ACCESS, WORDING, AND KNOWLEDGE ATTRIBUTION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Separating social reasoning from reading calls for a control condition, and one device is a non-mental question over the same story, read so that the difference isolates the social part. ToMi uses this device, pairing its belief questions with memory and reality controls, and so does the benchmark we study. We report a replicated positive result and a boundary on it. On nine instruction-tuned models from four families, recalling the two locations the question depends on raises balanced accuracy by 9.0 to 22.8 points *under an identical explicit instruction naming the comparison to perform*. With both additions present, five models reach 97.9 to 99.7; the four that begin at or near chance reach only 59.6 to 77.6, even though the same four score 95.4 to 99.4 when the witnesses are named and answering reduces to checking list membership. The nine models do not read as a continuous slope: a model chosen because it fell in the empirical gap lands on the crossing, though one model there cannot fix where the transition lies. Stating the procedure is not the bottleneck and can be actively harmful, costing up to 22.6 points when the facts are already supplied. We reach this after auditing our own control, which named the actor of the event while the knowledge question required recovering it; with designation matched we no longer detect the growth of the gap with group size, although our intervals are too wide to demonstrate an absence of effect. A contrast the confound cannot touch—an explicit perceptual block with the question held word-for-word identical—shows no omniscience bias in any of fourteen models. We audited two published benchmarks for the same asymmetry and found it in neither. Prespecifications, reproduction gates and internal controls are released, including an internal control that failed on seven models of nine.

1 INTRODUCTION

To ask whether a language model reasons about what others know, one device is to pair a mental-state question with a non-mental control over the same story. The control is meant to absorb what the two share—reading, entity tracking, narrative following—so that the remainder isolates social reasoning.

We examined two published benchmarks. ToMi (Le et al., 2019) uses this device; BigToM (Gandhi et al., 2023) instead keeps the question fixed and perturbs the story. We therefore make *no* claim about how common either strategy is: the design we audit is the one our own benchmark uses, and which ToMi shares.

This paper makes one positive claim about that remainder and one cautionary claim about how easily the pairing fails.

The positive claim is behavioural, and it comes with a boundary. On a task whose epistemic rule is stated in the prompt, recalling the two locations the question depends on raises accuracy by 9 to 23 points on *every* model we tested, with an explicit instruction naming the comparison present in both conditions, so the benefit is not obtained merely by telling the model what to do—indeed that instruction can cost a model twenty points. But only five of the nine go on to reach ceiling; the four that begin at or near chance improve substantially and still fail, and the transition between the groups looks like a threshold. We do not claim to identify which processing stage benefits.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The cautionary claim concerns how we got there. Our control question named the actor of the event; our knowledge question did not. Answering the knowledge question therefore required a step the control never required. Before the audit our headline was that the gap widens with the number of agents, which reads as a limit on how many mental states a model can track. With designation matched we no longer detect that growth.

Our contributions are: a controlled generator that **states its epistemic rule in the prompt** (§3); an audit of our own control and a matched-designation experiment on nine models (§5); a 2×2 separating *supplying the premises* from *stating the procedure*, replicated on nine models from four families (§6); a threshold separating list lookup from a single inference (§7); and a perceptual block probe across fourteen models plus a negative audit of two published benchmarks (§8, §9).

2 Related work

Benchmarks derived from the Sally–Anne false-belief task (Baron-Cohen et al., 1985) evaluate belief attribution by question answering. ToMi (Le et al., 2019) is itself a methodological correction: its authors show that existing benchmark tasks can be solved without theory of mind (Premack & Woodruff, 1978) because of dataset biases, and add *memory* and *reality* control questions. BigToM (Gandhi et al., 2023) generates evaluations procedurally and ships control conditions in which a random event replaces the target event. Our contribution is in the same spirit but concerns a different property of the pairing: whether the non-mental control matches the target question in what it makes the model *retrieve*. We audit both benchmarks for that property in §9.

What these benchmarks measure is contested. Kosinski (2023) reports that a recent model solves a majority of false-belief tasks and raises the possibility that theory of mind emerged as a by-product of language ability. Ullman (2023) answers that small variations preserving the principles of the task reverse such results, and Shapira et al. (2023) find that models struggle with adversarial examples, which they read as reliance on shallow heuristics. Our results bear on both sides. The crossover of §7 is an instance of the heuristic pattern: two models score near ceiling where answering reduces to checking a list, and fall by more than twenty points where the same information requires one inference. Our internal control is an instance of the fragility Ullman (2023) describes, in a form that does not even alter the task: a sentence that is true and wholly redundant costs up to 12 points. That such a sentence is not neutral is itself documented, from two directions. Shi et al. (2023) show that adding *irrelevant* context sharply degrades reasoning, and Sclar et al. (2024) show that *meaning-preserving* changes to prompt formatting move accuracy by as much as 76 points, and argue that a single format therefore does not characterise a model. Our case sits between the two: the added sentence is neither irrelevant nor a formatting choice but a true restatement, which makes its cost harder to dismiss either as distraction or as surface form.

A second line reports that deliberation does not reliably help on mental-state tasks. Gong et al. (2026) find that reasoning models do not consistently outperform non-reasoning ones and sometimes do worse, and describe a *slow thinking collapse* in which extended reasoning degrades performance. We preregistered the corresponding prediction for our task and did not reproduce it (§4.1).

Our positive result is adjacent to work on how models use the information in their context. Liu et al. (2024) show that performance depends strongly on *where* relevant information sits, being highest at the beginning or end of the context and degrading in the middle. Our recall manipulation moves the relevant facts next to the question and changes their availability, proximity, repetition and relevance at once; we are explicit throughout that our design does not separate these, and we set out the control that would in §10.

3 The controlled benchmark

Each item states the perceptual rule explicitly: before the event nobody knows what it reveals; everyone stays in their room; rooms are separate; everyone in the event room perceives and remembers; and a person learns the fact if and only if they perform the event or see or hear it. The story then places N named agents in rooms and describes one event performed by one actor, drawn from five event types in equal proportion. Because the rule is given, a correct answer is derivable from the text, which distinguishes the task from benchmarks leaving the epistemic rule implicit.

```
Rules for this story:  Before the event, nobody knows the
information that the event will reveal.  Everyone stays in their
stated room throughout the event.  Rooms are separate:  nobody
can see or hear events in another room.  Unless explicitly blocked
below, everyone in the event room watches all visible actions and
listens to all speech, understands the information, and remembers
it.  A person learns the fact if and only if they perform the event
or see or hear the information revealing it.  Nobody receives any
other information.
There are 4 people:  Tom, Bob, Kate, and Zoe.  Tom is in the office.
Bob is in the kitchen.  Kate is in the office.  Zoe is in the
office.
Kate moved the blue key from the table into the drawer in full view.
Does Tom know where the blue key is?
```

Table 1: A **rule/open/knowledge** item at $N{=}4$, shown in full and copied from the released data. The spatial control replaces the final line with `Is Tom in the same room as Kate?`, which names the actor that the knowledge question leaves to be recovered. The label is `True`.

	balanced accuracy			"knows" rate	
model	rule	named	spatial	open	blocked
	Mistral-7B	-12.2	$[-13.6, -10.8]$		
	Llama-3.1-8B	$+0.9$	$[-0.2, +2.0]$		
Llama-3.2-3B	52.3	79.5	59.5	10.2	0.0
Qwen2.5-3B	53.5	97.8	63.3	13.0	0.1
gemma-2-2b	54.0	95.4	51.5	42.0	0.5
Mistral-7B-v0.3	54.8	99.4	67.2	11.5	0.0
Mistral-Nemo-12B	55.1	98.4	76.1	7.0	0.0
Llama-3.1-8B	55.6	97.1	85.9	7.3	0.0
Qwen2.5-7B	58.7	95.4	69.4	83.0	0.0
Mistral-Small-22B	65.0	99.7	91.0	26.0	0.0
gemma-2-9b	67.2	88.8	90.3	77.1	0.0
gemma-2-27b	72.5	93.0	97.0	76.0	0.0
Qwen2.5-14B	76.6	90.0	94.6	58.9	0.0
Qwen2.5-32B	88.5	98.7	97.7	51.1	0.0
Qwen2.5-72B	90.5	98.5	97.4	47.0	0.0
Llama-3.1-70B	96.0	99.9	99.8	53.4	0.0

Table 2: Fourteen models, `enumerated` layout. The last two columns give the rate of "knows" answers under **open** access and under an explicit perceptual block, where every such answer is an error (§8).

Table 1 shows a complete item. Three cue conditions share the underlying scenario. In **rule** the witness set must be derived. In **named** an added sentence lists the witnesses; this adds text, raising median story length from 864 to 940 characters, so **named** is not length-matched. In the **spatial** control the question asks about co-location.

Access is **open** or **blocked**, the latter giving the queried person an opaque blindfold and ear defenders. Group size takes eleven values from 2 to 20 with 100 true and 100 false items each, and two layouts render the same scenario. Labels are computed by construction. Models answer in one pass, scored by the logits of single-token `True`/`False` continuations. We report balanced accuracy; all intervals are scenario-level paired bootstraps with 2,000 resamples, resampling `pair_id` so that every observation of a scenario moves together.

4 THE APPARENT DEFICIT AND ITS SCALING

Table 2 gives the fourteen models. The gap between **rule** and the **spatial** control widens with N, which read at face value is a scaling result. Two facts temper it before any audit. Six models sit between 52 and 59 points, so their apparent N effect is a floor artefact. Among the rest the effect

is concentrated: ranked by how much the gap widens between $N \in \{2, 4\}$ and $N \in \{12, 20\}$, it is -12.5 for Qwen2.5-72B, -8.8 for Qwen2.5-14B and gemma-2-27b, -6.2 for Qwen2.5-32B, and at most -3.0 for every other above-chance model, including -1.2 for the strongest, Llama-3.1-70B. The two models we later put through the matched-designation protocol are therefore drawn from the top of that ranking rather than favourably selected.

4.1 DELIBERATION

Chain-of-thought prompting (Wei et al., 2022) on Qwen2.5-14B, with a prespecified validation gate that the direct arm reproduced to within 0.1 points, raises **rule** from 78.1 to 91.4. The gap survives: at N=20 the spatial control reaches 99.5 and knowledge attribution 86.5. Following Gong et al. (2026) we preregistered that the benefit would shrink with N; the opposite occurred, -8.0 points with a paired interval of $[-13.2, -2.8]$. Generated length did not grow with N: its median fell from 168 tokens at N=2 to 148 at N=20.

5 THE CONTROL WAS CONFOUNDED

The two questions are, verbatim:

```
Is {query} in the same room as {actor}?
Does {query} know {fact}?
```

The spatial control hands over the actor; the knowledge question requires recovering it from the event sentence. The contrast therefore mixes applying the epistemic rule with identifying the agent who acted. A retrieval-based account is *compatible* with every result in the previous section—the gap, its growth with N as the candidate set grows, the benefit of naming witnesses, the benefit of deliberation—which is what motivates the repair; compatibility is not demonstration, and the experiments below are what test it.

One fact makes the repair interpretable. Across all $2{,}200$ scenarios the event room is the actor's room, the stories are byte-identical between the two tasks, and the spatial label equals the knowledge label in $2{,}200$ cases out of $2{,}200$, so every cell carries one label per scenario.

5.1 CONSTRUCTION

We derive cells by rewriting *only* the `question` field of existing items, leaving the generator untouched so that no random path moves; every story is asserted byte-identical to its source. With `DESC` an actor description fixed per event type (for `object_move`, `the person who moved the blue key`), the cells are:

`MAP`	`The person who {desc} is {actor}.`
`RECALL`	`{actor} is in the {room`$_a$`}. {query} is in the {room`$_q$`}.`
`INSTR`	`To answer, compare the room {query} is in with the room {actor} is in.`

Order is `MAP` + [`RECALL`] + [`INSTR`] + question, so that `INSTR` always sits immediately before the question in both cells carrying it and `RECALL` always immediately after `MAP`; no single order can match both factors' positions at once. Two cells replay the original questions verbatim as a reproduction gate. **All eighteen replayed anchors, across nine models, reproduce the original run to within** 0.1 **points** over $2{,}200$ paired scenarios each.

5.2 RESULTS

Replacing the actor's name by a description in the *spatial* question costs 19.4 points on gemma-2-27b, on a question containing no epistemic inference. Table 3 summarises nine models.

Naming the actor is heterogeneous and reverses. The gain is $+5.0$ on gemma-2-27b and -3.0 on Qwen2.5-32B, both intervals excluding zero in opposite directions. We report the heterogeneity and claim no general benefit.

model	G	naming	residual	internal ctrl.
below the threshold: knowledge anchor under 60				
gemma-2-2b	-2.5	-1.2	$+1.8$	-0.5
Mistral-7B	12.4	-3.4	-3.5	-12.2
Llama-3.1-8B	30.3	$+2.7$	-28.4	$+0.9$
Qwen2.5-7B	10.7	-4.6	-8.7	-6.6
Mistral-Small-22B	26.1	$+1.5$	-16.5	-8.0
gemma-2-9b	23.0	$+1.1$	-16.0	-6.0
gemma-2-27b	24.5	$+5.0$	-14.0	-5.5
Qwen2.5-14B	18.0	$+1.4$	-12.3	-4.3
Qwen2.5-32B	9.2	-3.0	-11.0	-1.3

Table 3: Matched designation, nine models ordered by starting accuracy. G is the original gap; *naming* is the gain from naming the actor in the knowledge question; *residual* is the gap with designation matched; *internal ctrl.* is the cost of adding `MAP` to a spatial question that already names the actor, which we prespecified should be zero and which excludes zero on seven of the nine. Contrasts for models below the threshold are bounded by the floor. Intervals are in Appendix D.

An internal control failed on seven models of nine. Adding `MAP` to a spatial question already naming the actor costs between 1.3 and 12.2 points on seven models, every interval excluding zero; only Llama-3.1-8B at $+0.9$ and gemma-2-2b at -0.5 are indistinguishable from zero, and the latter is at chance throughout. A true but wholly redundant sentence is therefore not behaviourally neutral, which limits the naming estimate: it measures the net effect of a textual intervention. `MAP` is identical in both cells of the residual contrast, making it paired under a common prefix; we do not claim its cost cancels, which would assume an untested absence of prefix-by-task interaction.

Growth with N is not detected. Among the four models above chance, the residual gap at $N \in \{12, 20\}$ minus $N \in \{2, 4\}$ is $+6.5$, $+3.8$, $+0.8$ and -3.8, none significant under Holm correction across five primary contrasts, and the point estimates do not share a sign. The two floor models are excluded here by the rule stated in §6. We take this as a limitation on the original interpretation rather than a demonstrated absence. A null can carry an argument given tight enough intervals or a met equivalence margin; ours are wide, and in our generator N covaries with room count (two to four), text length and distractor count.

6 Supplying the premises versus stating the procedure

Recall changes availability, proximity, repetition and relevance at once, and in particular makes salient *that* a comparison is required. We therefore cross the two additions rather than compare one against the other, since comparing instruction-without-facts against facts-without-instruction would move two factors simultaneously. Writing k for the cell naming the actor only:

$$Q_1 = (k{+}\texttt{INSTR}) - k \qquad \text{procedure alone}$$
$$Q_2 = (k{+}\texttt{RECALL}{+}\texttt{INSTR}) - (k{+}\texttt{INSTR}) \qquad \text{facts, instruction constant}$$
$$Q_3 = Q_2 - \big[(k{+}\texttt{RECALL}) - k\big] \qquad \text{interaction}$$

Holm correction (Holm, 1979) is applied across Q_1, Q_2, Q_3; everything else in this section is exploratory.

Main result, and its boundary. With the instruction held constant, supplying the facts is worth between $+9.0$ and $+22.8$ on all nine models, every interval well clear of zero. The effect replicates across four families and requires no prior competence: the largest gain belongs to Mistral-7B, which is at chance without it.

The *ceiling* does not follow. With both additions five models reach 97.9 to 99.7, while the four beginning at or near chance reach only 59.6 to 77.6; taking instead the better of recall alone and both additions, which we do not do in any contrast, would move those four only to 66.1–82.2. Supplying the premises improves them substantially without making them succeed. We do not identify which processing stage benefits.

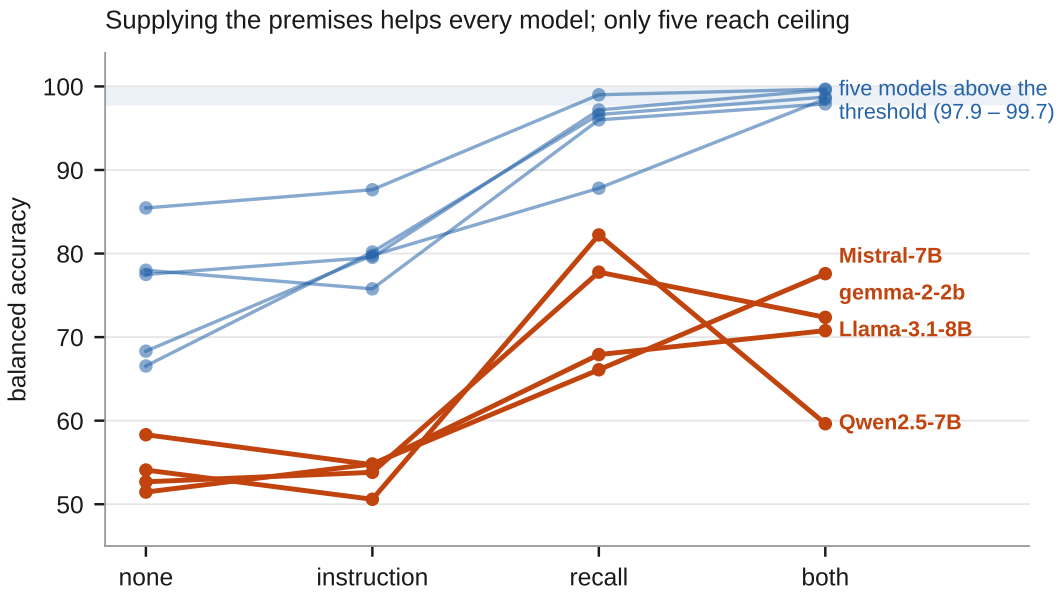


Figure 1: The four conditions on nine models. "none" already names the actor; "instruction" states which rooms to compare without supplying them; "recall" supplies them. The four models that begin at or near chance gain as much as the others and still end far below them, and two of them are *worse* with both additions than with recall alone.

model	Q_1 instruction	Q_2 recall	Q_3 interaction
below the threshold			
gemma-2-2b	+1.1	+18.5	− 6.5
Mistral-7B	+3.4	+22.8	+ 8.1
Llama-3.1-8B	−3.6	+16.0	+ 6.5
Qwen2.5-7B	−3.5	+ 9.0	−19.1
Mistral-Small-22B	+13.6	+18.5	−11.5
gemma-2-9b	+11.5	+18.7	− 0.8
gemma-2-27b	+2.0	+20.0	+ 0.4
Qwen2.5-14B	−2.2	+22.1	+ 4.1
Qwen2.5-32B	+2.2	+12.0	− 1.5

Table 4: Contrasts for Figure 1, nine models ordered by starting accuracy. Q_2 excludes zero on every model. All intervals are in Appendix D.

The instruction can do real harm. With the facts already supplied, adding the instruction is worth +11.5 on Mistral-7B and +10.7 on gemma-2-9b, but −5.4 on gemma-2-2b and −22.6 on Qwen2.5-7B, every interval excluding zero. On Qwen2.5-7B recall alone is worth +23.5 from the anchor and the instruction takes 22.6 of that back. Two models are therefore *better* with recall alone than with both additions, so the best condition is not the same for every model. We report both conditions separately rather than taking the better of the two, which would be a post-hoc selection; the prespecified Q_2 remains the reference measure.

A floor rule, fixed in advance. Balanced accuracy near 50 carries no signal, so before running these two models we declared that for any model whose knowledge anchor falls below 60 we would report the cells and Q_2, treat P_1, P_2, Q_1 and Q_3 as floor-bounded, and treat P_3 as *uninterpretable*: an N-slope on a gap with one term at chance measures nothing. We hold to that here, where P_3 is +16.0 and +13.2 with intervals excluding zero on the two floor models and we do not interpret it.

The instruction is heterogeneous. Q_1 ranges from −2.2 on Qwen2.5-14B to +11.5 on gemma-2-9b; the effects on the remaining models are small but not absent. On Qwen2.5-14B the instruction

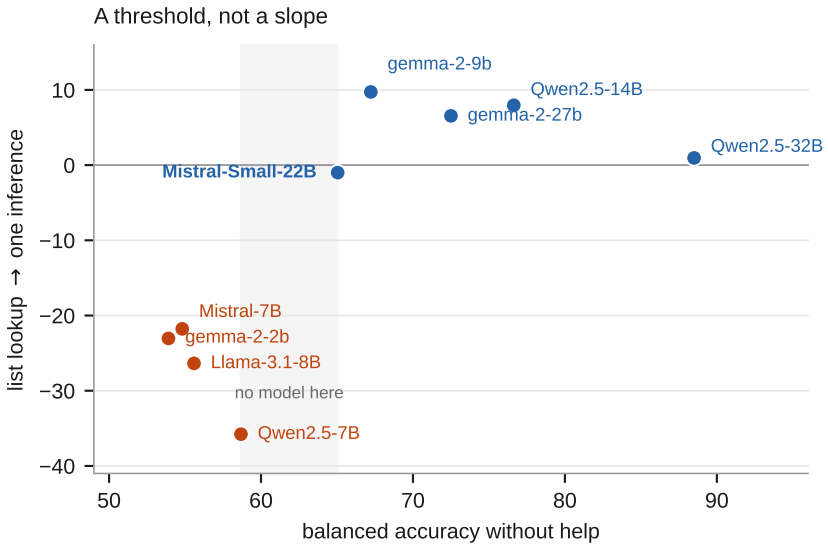


Figure 2: The difference between the two ways of supplying the answer's ingredients, plotted against starting accuracy. Positive means the model does better with the rooms than with the witness list. The shaded band is the region where we had no model, which is why Mistral-Small-22B was run.

hurts without recall and helps with it, an interaction of $+4.1$; the null interaction on gemma-2-27b does not generalise.

What this does not show. It does not show the models already know the procedure: the instruction may be redundant, insufficiently effective, or imperfectly followed, and these are not told apart. It does not show the recall benefit bypasses any change in how the procedure is executed, only that it survives an explicit constant instruction. The ceiling figures are ceilings, not evidence of general competence.

An unplanned regularity. Among the four models above chance, the cost of the redundant `MAP` sentence falls monotonically with accuracy: -6.0, -5.5, -4.3, -1.3 against 67.2, 72.5, 76.6 and 88.5. The original gap G does *not* follow monotonically (23.0, 24.5, 18.0, 9.2), so we claim monotonicity only for the internal control, and only within that group: the two floor models break it in both directions, Mistral-7B at -12.2 and Llama-3.1-8B at $+0.9$ despite near-identical original accuracies of 54.8 and 55.6. We report this as descriptive and exploratory. It suggests that part of what we measured as an epistemic deficit may be a general sensitivity to phrasing that diminishes with competence.

7 A THRESHOLD BETWEEN LIST LOOKUP AND A SINGLE INFERENCE

The four models below the threshold are not uniformly weak. In the **named** condition of §3, where an added sentence lists the witnesses, they score 95.4, 99.4, 97.1 and 95.4 (Table 2). There, answering reduces to checking whether the queried person appears in a list. Our recall condition instead supplies the actor's name and both rooms and leaves one inference: *same room, therefore knows*.

Figure 2 plots that difference against each model's accuracy without help. Five models are at or above zero, from $+1.0$ to $+9.7$: they do *better* when handed the rooms than when handed the list. Four fall by 21.8 to 35.8 points. Every interval excludes zero.

The transition is a threshold, and a model was run to test it. After six models we had a gap in starting accuracy between 58.7 and 67.2 with nothing in it, and could not tell a threshold from a slope. Mistral-Small-22B was chosen for that reason, at 65.0, before its result was known. It lands on the crossing at -1.0, the only model near zero. Across 6.3 points of starting accuracy the difference moves from -35.8 to -1.0, while within each group it is comparatively flat, -21.8 to

-35.8 below and $+1.0$ to $+9.7$ above. With one model in the gap we can neither locate the threshold nor measure how sharp it is; what we can say is that the nine points do not read as a continuous slope.

It is not a response-bias artefact. The four models below the threshold answer "knows" at very different rates without help: 7.3% for Llama-3.1-8B, 11.5% for Mistral-7B, 42.0% for gemma-2-2b and 83.0% for Qwen2.5-7B. Three quite different response profiles give the same pattern, so a conservative bias does not explain it.

The comparison is confounded with length. The two conditions differ by two sentences as well as by the inference, so we cannot attribute the reversal to the inference alone. That matters because the cost of a redundant sentence, measured by our internal control, ranges from -12.2 on Mistral-7B to $+0.9$ on Llama-3.1-8B. Part of Mistral's drop may be a burden of added text. That explanation does not cover Llama-3.1-8B, whose internal control is indistinguishable from zero and which still falls 26.4 points. A length- and position-matched control is the experiment this calls for, and we have not run it.

8 A CONTRAST THE CONFOUND CANNOT TOUCH

In the **blocked** condition the queried person wears an opaque blindfold and ear defenders and the question is word-for-word identical to the open condition, so the actor-retrieval load is the same on both sides. All fourteen models answer "knows" at essentially zero under the block ($\leq 0.5\%$), against 7.0% to 83.0% in the open condition (Table 2). This is not a response-bias artefact: gemma-2-9b answers "knows" 77.1% of the time when access is open and 0.0% when blocked, and Qwen2.5-7B goes from 83.0% to 0.0%. We find no omniscience bias when the block is stated. Because the blocked cells carry a single label this is a bias probe rather than an accuracy measure, which is why we read it as a rate and check it against the open-condition rate.

9 DOES THE ASYMMETRY OCCUR ELSEWHERE?

We looked for a published benchmark contrasting a mental-state question against a control that supplies an entity the target question makes the model retrieve, taking the published contrast rather than the item as the unit, and did not find one. **We examined two benchmarks**, which is a small sample; we draw no conclusion about prevalence from it, in either direction.

In BigToM (Gandhi et al., 2023) the belief question names the agent whose belief is queried, and the released generator produces control conditions that substitute a random event while keeping the question format, so condition and control are designation-matched by construction. In ToMi (Le et al., 2019) every question names every entity it refers to, in every item, since each question type has a single template; there is consequently no designation intervention to make. ToMi's belief question does require more retrieval than its controls, but that is retrieval *depth*, and a mental-state task requiring more retrieval than its control is an expected property rather than a confound.

We therefore report the confound as a property of our own design. The claim that published reading controls fail to match access difficulty has, at present, no published example behind it.

10 LIMITATIONS

Nine models from four families, answering in one pass on a single controlled task whose rule is given; nothing here speaks to open-ended or naturalistic settings.

The mechanism is not identified. Recall changes availability, proximity, repetition and relevance together. The natural next step is a length- and position-matched recall of *irrelevant* facts: if it hurts while relevant recall helps, content matters at matched length and position; if it also helps, the contribution is non-specific, which would not by itself identify position or salience, and distance would have to be manipulated separately (Liu et al., 2024).

Group size covaries with room count, text length and distractor count, so our null on growth with N does not isolate an effect of agent number in either direction. A cross-model analysis relating the

apparent deficit to a serial-position effect over the fourteen models gives a correlation of -0.04; an earlier version on three models suggested otherwise and did not survive extension.

Four of the nine sit at or near chance on the original task, so several contrasts are floor-bounded for them and one, P_3, we declared uninterpretable in advance rather than report a spuriously significant slope. The threshold of §7 rests on a single model inside the gap, which fixes neither its location nor its sharpness; a second model between 59 and 65 would.

We did not test a model trained to deliberate. That experiment was prespecified, launched and frozen; its data are released unanalysed, and its direct arm is not reported on its own because a reasoning-trained model with deliberation disabled would raise the question without answering it.

11 Conclusion

On a task whose epistemic rule is stated, recalling the premises beside the question helps all nine models we tested, across four families, even when the comparison to perform is spelled out—and spelling that procedure out is worth anything from $+11.5$ to -22.6 depending on the model, so it is not what the models were missing. Only five then answer almost perfectly. The other four improve and still fail, yet score between 95.4 and 99.4 in a condition where answering reduces to checking a list, and the gap between the two groups looks like a threshold rather than a slope. What looked like a limit on attributing knowledge across a group is, on this task, better described as a difficulty in using facts that are already present in the context—and, for the weakest models, as an inability to take even one step from facts that have been handed over.

We reached that description by auditing our own non-mental control, which named an entity the mental-state question required recovering. The lesson is not that such controls are generally broken—we checked two published benchmarks and found them matched—but that a control absorbs only the difficulty it actually shares, and that this is worth checking question by question.

Use of Large Language Models

Per the ICLR 2027 policy, we disclose how large language models were used in this work. This section is not counted towards the page limit.

An LLM-based coding assistant was used throughout, under author direction, for: **research execution**—writing the stimulus-derivation, inference and analysis scripts, launching the cluster jobs, and computing the bootstraps reported here; **drafting**—producing first drafts of the manuscript text, which the authors then revised; **figures**—writing the plotting code, which reads the released prediction files directly so that no figure value is transcribed by hand; and **retrieval**—locating candidate related work, each item of which was then checked against its source record before being cited.

The experimental design decisions, the prespecifications, the equivalence margin and the floor rule were set by the authors before the corresponding data existed. Several claims drafted with assistance were subsequently weakened or retracted by the authors on review, including a serial-position analysis that did not survive extension from three models to fourteen, and an earlier reading that attributed a reversal to an inference step which the design cannot isolate; both are reported in the paper as such.

We have reviewed all AI-assisted work: every number in the tables was recomputed from the released predictions by a verification script, every reference was checked against a bibliographic record, and the reproduction gates were run before any result was interpreted. We take responsibility for the final content of this work.

References

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

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*, 2023.

Nanxu Gong, Haotian Li, Sixun Dong, Jianxun Lian, Yanjie Fu, and Xing Xie. To think or not to think, that is the question for large reasoning models in theory of mind tasks. *arXiv preprint arXiv:2602.10625*, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *arXiv preprint arXiv:2302.02083*, 2023.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

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

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*, 2023.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

A Effects broken down by group size

Figure 3 gives the knowledge–spatial gap at each of the eleven group sizes, before and after matching designation. Two features are visible that the aggregate hides. First, neither gap grows smoothly: both jump between $N{=}2$ and $N{=}4$ and then plateau, so the aggregate "growth with N" is largely a step between the two smallest sizes. Second, matching designation shifts the whole curve upward rather than changing its shape.

B The deliberation protocol and its gate

The chain-of-thought experiment was prespecified before the run. Because it replaces one-pass logit scoring with short greedy generation, we required the direct arm to reproduce the original logit-scored accuracies before interpreting the deliberation arm. It did, within 0.1 points on all three cells: 78.1 against 78.1 for **rule**/knowledge, 90.1 against 90.0 for **named**/knowledge, and 94.3 against 94.3 for **rule**/spatial.

Answers were parsed from the generation with a declared fallback order. No fallback was needed: across all three cells and both arms, 0.0% of responses were unparsable and 0.0% were truncated at the 512-token budget, the longest observed generation being 390 tokens. Median generated length fell with group size, from 168 tokens at $N{=}2$ to 148 at $N{=}20$, which is the observation that bears on the *slow thinking collapse* prediction of Gong et al. (2026); we report it as descriptive, since a correct answer can be short.

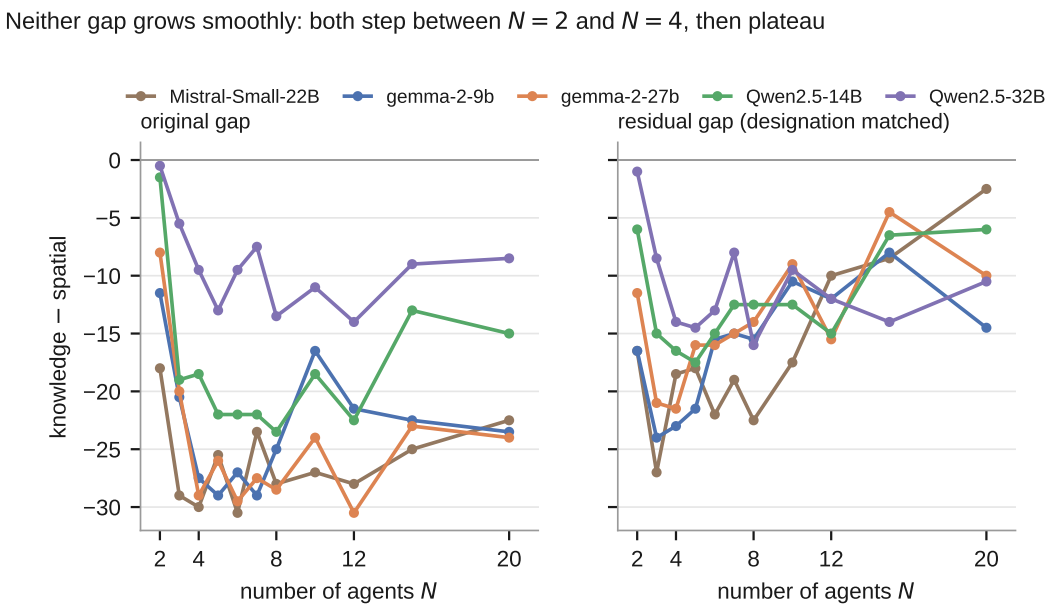


Figure 3: The knowledge–spatial gap at each of the eleven group sizes, 200 items per point, for the four models above chance. The summary contrast P_3 compares the mean of $N \in \{12, 20\}$ against $N \in \{2, 4\}$.

C GENERATION, MATCHING AND SCORING CONTROLS

Stimuli. The matched-designation cells are derived by rewriting only the `question` field of existing items. The generator is not modified, because adding a branch would move its random path and silently change the stories. The derivation asserts, for every one of the 2,200 scenarios, that the story is byte-identical to its source, that the label is identical across all cells, that the actor is in the event room, that the two anchor questions match their original text exactly, and that no actor description contains any room name used in the dataset. Files are hashed and revalidated by re-deriving in memory and comparing bytes.

Matching. Every contrast is computed on the same 2,200 scenarios, joined on `pair_id`. Bootstraps resample scenarios, not items, so that all observations of a scenario move together.

Scoring. Answers are scored as `max_logit_over_single_token_variants` over single-token `True` and `False` continuations, in `bfloat16` with batch size 16. The same runner, template detection and token-id selection are used for the new cells as for the original run, which is what makes the anchor gate meaningful; the model revision recorded in each run manifest is checked against the local snapshot before launch.

Reproduction gate. Two cells replay the original questions verbatim. Across four models this yields eight replayed anchors, all reproducing the original accuracies at exactly $+0.00$ over 2,200 paired scenarios. The analysis halts if the gate fails.

D FULL PER-MODEL CELLS AND INTERVALS

Exploratory cells and analyses. The knowledge cell with both additions is a ceiling check rather than a contrast, since supplying both rooms and the epistemic rule makes the answer nearly derivable. The relation between model accuracy and the cost of the redundant sentence (§6) is descriptive over four points. The cross-model serial-position analysis reported in §10 covers all fourteen models of Table 2.

	spatial			knowledge				
model	anch.	desc.	+MAP	anch.	MAP	REC	INS	both
gemma-2-2b	51.5	51.0	50.9	53.9	52.7	77.8	53.8	72.4
Mistral-7B	67.2	58.8	55.0	54.8	51.5	66.1	54.8	77.6
Llama-3.1-8B	85.9	76.5	86.7	55.6	58.3	67.9	54.7	70.8
Qwen2.5-7B	69.4	60.2	62.8	58.7	54.1	82.2	50.6	59.6
Mistral-Small-22B	91.1	79.7	83.1	65.0	66.5	96.6	80.2	98.7
gemma-2-9b	90.3	68.9	84.3	67.2	68.3	87.8	79.8	98.5
gemma-2-27b	97.0	77.6	91.5	72.5	77.5	97.2	79.5	99.6
Qwen2.5-14B	94.6	81.8	90.3	76.6	78.0	96.0	75.8	97.9
Qwen2.5-32B	97.7	90.8	96.5	88.5	85.5	99.0	87.6	99.7

Table 5: All eight cells for the nine models, balanced accuracy over 2,200 scenarios, ordered by starting accuracy. The four above the rule are below the threshold of §7.

model	P_1 naming	P_2 residual	Q_1 instr.	Q_2 recall	Q_3 inter.	internal
gemma-2-2b	$-1.2\,[-3.3, +0.9]$	$+1.8\,[-0.2, +3.9]$	$+1.1\,[-1.0, +3.1]$	$+18.5\,[+17.0, +20.3]$	$-6.5\,[-9.2, -3.8]$	$-0.5\,[-1.1, +0.0]$
Mistral-7B	$-3.4\,[-4.5, -2.3]$	$-3.5\,[-4.8, -2.4]$	$+3.4\,[+1.4, +5.4]$	$+22.8\,[+20.7, +24.7]$	$+8.1\,[+5.8, +10.5]$	$-12.2\,[-13.6, -10.8]$
Llama-3.1-8B	$+2.7\,[+1.5, +4.0]$	$-28.4\,[-30.2, -26.6]$	$-3.6\,[-4.9, -2.2]$	$+16.0\,[+14.6, +17.4]$	$+6.5\,[+4.6, +8.2]$	$+0.9\,[-0.2, +2.0]$
Qwen2.5-7B	$-4.6\,[-6.0, -3.2]$	$-8.7\,[-11.2, -6.3]$	$-3.5\,[-5.0, -1.9]$	$+9.0\,[+7.8, +10.3]$	$-19.1\,[-21.2, -17.1]$	$-6.6\,[-8.4, -4.9]$
Mistral-Small-22B	$+1.5\,[+0.1, +2.8]$	$-16.5\,[-18.2, -14.8]$	$+13.6\,[+11.9, +15.2]$	$+18.5\,[+17.0, +20.1]$	$-11.5\,[-13.3, -9.7]$	$-8.0\,[-9.3, -6.8]$
gemma-2-9b	$+1.1\,[-0.2, +2.3]$	$-16.0\,[-17.6, -14.2]$	$+11.5\,[+9.6, +13.2]$	$+18.7\,[+17.0, +20.4]$	$-0.8\,[-2.9, +1.3]$	$-6.0\,[-7.2, -4.8]$
gemma-2-27b	$+5.0\,[+3.5, +6.6]$	$-14.0\,[-15.7, -12.2]$	$+2.0\,[+0.8, +3.3]$	$+20.0\,[+18.4, +21.7]$	$+0.4\,[-1.1, +1.8]$	$-5.5\,[-6.6, -4.6]$
Qwen2.5-14B	$+1.4\,[+0.2, +2.5]$	$-12.3\,[-14.1, -10.4]$	$-2.2\,[-3.9, -0.6]$	$+22.1\,[+20.4, +23.9]$	$+4.1\,[+2.4, +6.0]$	$-4.3\,[-5.2, -3.5]$
Qwen2.5-32B	$-3.0\,[-4.2, -1.9]$	$-11.0\,[-12.4, -9.5]$	$+2.2\,[+0.6, +3.8]$	$+12.0\,[+10.8, +13.3]$	$-1.5\,[-3.0, +0.2]$	$-1.3\,[-1.8, -0.8]$