StructToM-Bench: Locating Where Accuracy Is Lost in Theory-of-Mind Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Theory-of-mind (ToM) benchmarks usually report final-answer accuracy, which can conflate an incorrect answer with a rejected output or, for structured methods, a state that cannot answer the query. We introduce StructToM-Bench, a controlled benchmark of 1000 items with executable gold events, evaluated through direct answering and model-generated event graphs. In a preregistered, blinded two-choice audit, a majority of five human annotators favored the executable gold on each of the 31 items where it disagrees with the source label. For the graph path, strict accuracy separates into three observable stages: output compliance, structured-state availability, and query-result correctness. Across 22 model–deployment configurations of 1–15B parameters, evaluated over five passes against fixed denominators, at least one declared acceptance rule raises the direct score for 10 configurations. Two strict zeros rise to 0.577 and 0.610, while a punctuation-only check leaves 20 of 22 scores unchanged. Among strict graph-path failures, 77.1% are rejected outputs, 17.5% are accepted but unqueryable graphs, and 5.4% are queryable graphs whose computed answer differs from gold. An audit of our structural edge metric found that 30.2% of its gold edges belonged to a family the prompt instructed models to omit, so the metric penalized models for following the prompt. Because positional heuristics selected per operation on this corpus outperform every configuration within each operation, we interpret these findings as measurements of the evaluated pipelines rather than evidence of ToM competence. We recommend reporting strict accuracy alongside acceptance sensitivity, queryable coverage, and scorer and deployment provenance.

1 Introduction

Llama-3.2-1B scores 0.000 under this paper's strict direct-answer contract, which requires a `Final answer:` line exactly matching one of two options. Our internal answer-recovery rule also returns 0.000. Yet acceptance rules ported from the published OpenToM and FANToM evaluators (Xu et al., 2024; Kim et al., 2023), applied to the same retained responses, each return 0.577. The responses did not change; only the acceptance rule did.

Does that strict zero indicate failed belief tracking—following what each character has seen and therefore believes—or failure to satisfy the output contract? One false-belief item asks where Chloe will look for a radish after she leaves the room and Liam moves it from a bottle to a crate; the gold answer is the bottle, which Chloe last saw. Final-answer accuracy cannot distinguish these failure sources: the same loss can arise from an incorrect answer, an unusable response, a structured state omitting the queried entity, or a scorer rejecting recoverable content. This is a broader construct-validity concern (Bean et al., 2025); preferring either a strict or permissive score does not resolve it. Prior work reports strong performance on some standard false-belief tasks (Kosinski, 2024; Strachan et al., 2024), but also substantial ToM limitations (Sap et al., 2022), adversarial fragility (Shapira et al., 2024), and prompt-format sensitivity (Sclar et al., 2024). Some datasets also admit shortcuts that bypass the intended construct (Le et al., 2019)—including, as we show, the corpus used here. What remains unclear is where measured losses enter the pipeline, and which conclusions survive changes to its acceptance rules, elicitation path and serving conditions.

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.

Research question. We ask where along the evaluation pipeline strict failures become observable, and how far reported scores depend on declared scoring choices rather than on the model. For a graph response, we ask three successive questions: was its output accepted, could its events answer the question, and did the computed answer match gold? The corresponding rates are accepted responses over planned calls, queryable graphs over accepted graphs, and correct answers over queryable graphs. Strict accuracy therefore factorizes as:

$$\Pr(\text{correct}) = \underbrace{\Pr(\text{accepted})}_{\text{compliance}} \cdot \underbrace{\Pr(\text{queryable} \mid \text{accepted})}_{\text{structured-state availability}} \cdot \underbrace{\Pr(\text{gold} \mid \text{queryable})}_{\text{query-result correctness}},$$

nested by construction, since queryability is defined only for accepted graphs (Figure 1). The factorization is an accounting identity, observational rather than cognitive; the findings are how failures distribute across it. The direct path exposes acceptance and final correctness but no separate state or query result, so a wrong direct answer cannot be localized across these measurements.

The instrument. StructToM-Bench pairs 1000 controlled story–question items, drawn from 875 distinct narratives, with gold events and a deterministic query engine. The direct path asks for an answer; the event-graph path asks for typed story events (such as entering, leaving, or moving an object) and a stated answer in one object. For its primary score, the benchmark ignores the stated answer and computes one from the emitted events. The model is still solving the question while producing the graph, so we measure structured-state availability *under joint event-and-answer elicitation*. The emitted graph is an observable output, not a readout of latent state, and the two prompts differ in several respects: their contrast compares evaluation pipelines rather than isolating a causal variable. We evaluate 22 model–deployment configurations of 1–15B parameters; frontier endpoints enter only as host-pinned, same-lineage scope checks.

Contributions. **(i) A measurement framework.** A controlled benchmark with executable gold whose structured path separates output compliance, structured-state availability and query-result correctness while retaining strict accuracy as the primary score. In a preregistered, blinded two-choice audit, human annotators resolved all 31 disagreements between the executable gold and the source labels in favor of the executable gold. **(ii) An empirical audit of where losses enter.** The best of the declared evaluation rules changes measured scores and rankings on both paths, while a punctuation-only check barely moves them. Most strict graph-path failures occur before the graph can answer a query, and an audit found that our structural scorer penalized models for following the prompt (Sections 4 and 6). **(iii) Robustness and scope.** The two pipelines differ in opposite directions across operations, with four contrasts surviving Holm correction, but same-corpus shortcut baselines beat every system within each operation. Two independently constructed datasets, a preregistered prompt test and host-pinning comparisons further bound these conclusions: they describe the evaluated pipelines, not ToM competence.

Section 3 defines the instrument and measurements; Section 4 compares scores under declared rules; Sections 5 and 6 locate the losses; and Section 7 tests their scope.

2 Related Work

Theory-of-mind benchmarks and construct validity. Story-based false-belief evaluation has a long history in NLP (Nematzadeh et al., 2018; Le et al., 2019). Later benchmarks add longer narratives (Xu et al., 2024), information-asymmetric conversation (Kim et al., 2023), and recursive beliefs (Wu et al., 2023). Others test broader social cognition and interaction (Chen et al., 2024; Bortoletto et al., 2025; Tong et al., 2026), common ground and role-play (Soubki et al., 2024; Shinoda et al., 2025), or changing mental states (Xiao et al., 2025). Surveys document the diversity of resulting claims (Chen et al., 2025). Other work asks how far performance is explained by language statistics rather than belief tracking (Trott et al., 2026); across balanced variants and 17 open-weight models, scale helps but not monotonically (Kouwenhoven et al., 2026), which is our regime. StructToM-Bench trades naturalistic breadth for event-level control, and does not address whether literal belief prediction captures functional adaptation to another agent (Riemer et al., 2025). Construct-validity critiques (Bean et al., 2025; Riemer et al., 2025) ask what these benchmark scores measure; we address a complementary question: where in the evaluation pipeline a measured loss enters.

Controlled diagnostics and structured elicitation. BigToM demonstrates scalable causal templates (Gandhi et al., 2023); CheckList and contrast sets motivate behavioral slices and minimally changed examples (Ribeiro et al., 2020; Gardner et al., 2020). The same control creates a validity risk—template-derived labels can coexist with construction-aligned shortcuts, and fewer than half of the ToM studies surveyed by Soubki & Rambow (2025) validate against machine-exploitable patterns—so we report template provenance and same-corpus positional shortcut bounds by operation, and treat held-out-template validation as remaining work. SymbolicToM tracks higher-order belief states with an explicit graphical representation during decoding (Sclar et al., 2023); SimToM, PercepToM and TimeToM expose perspective or temporal state (Wilf et al., 2024; Jung et al., 2024; Hou et al., 2024); DEL-ToM and EnigmaToM use verifiable or neuro-symbolic representations (Wu et al., 2025; Xu et al., 2025); and PDDL-Mind translates narratives into PDDL actions and verifies state transitions against a domain, improving ToM answering through explicit state tracking (Zhu et al., 2026). We do not propose the graph path as a better reasoning method; we measure the acceptance, queryability and query-result correctness of the state a model itself emits, which leaves missing or unusable state visible.

Evaluation as measurement, and metric artifacts. Semantically equivalent prompt formats produce large accuracy spreads (Sclar et al., 2024), and prompt-wrapper choice alone shifts accuracy and parseability (Mehta, 2026). Structured-output compliance varies across models (Liu et al., 2024), and leaderboard audits show answer selection reordering rankings by as many as eight positions (Alzahrani et al., 2024). Schaeffer et al. (2023) show that apparent capability jumps can be artifacts of a discontinuous metric. Two lines of work are closest to ours. Galeone et al. (2026) show that 7–9B models can produce correct answers that invalid JSON makes unusable, and repair this by prompt optimization; Yu et al. (2025) show that the answer-extraction step of the evaluation chain changes measured accuracy, and train an extractor to improve it. Our contribution is specific to event-graph ToM evaluation: fixed-denominator accounting of where each strict failure becomes observable, coupled to audits of declared evaluation rules and of our own scorer, with serving provenance treated as part of the measured system.

3 The Instrument and Its Three Measurements

3.1 Corpus and executable gold

Items. StructToM-Bench contains 1000 story–question items from 875 distinct narratives in FixedToMi, the disambiguated ToMi variant (Le et al., 2019) distributed with SimToM (Wilf et al., 2024). Stories describe characters entering and leaving rooms and objects being placed or moved, as in the introduction's Chloe example. Questions ask for an object's earlier location (memory), its final location (reality), one character's belief (first-order), or one character's belief about another's (second-order). Each belief type has four variants—ToM or no-ToM in generator versions 0 and 1—giving ten operations of 100 items each. In the source labels, a ToM variant's answer differs from the object's final location, while a no-ToM variant's answer matches it. Source stories are tagged first- or second-order false belief for 665 items and true belief for 335.

Executable gold. Each item stores time-ordered gold events (such as enter, leave, put, and move) and a fixed query, such as Chloe's belief about the radish. Query targets are specified in the item, never inferred from model output. Replaying the events with a deterministic updater tracks object locations and what characters believe from what they witnessed; querying the resulting state yields the gold answer. Gold events are converted from structured source data, not independently annotated. The same updater processes a model's events, so graph-path failures can be located at output rejection, lack of a queryable state, or a queryable state whose answer differs from gold.

Source-label disagreements. Executable gold agrees with the source answer on 969 of 1000 items. All 31 disagreements occur in no-ToM operations of stories tagged first-order false belief; the ToM/no-ToM and 0/1 names are inherited generator labels, not guaranteed descriptions of the adjudicated answers. An internal LLM adjudicator agreed with executable gold on all 31 disagreements. In a preregistered, blinded two-choice audit, a majority of five qualifying human annotators chose executable gold on every disagreement (29 unanimously), confirming the registered prediction. Appendix L gives the construction, adjudication procedure and provenance.

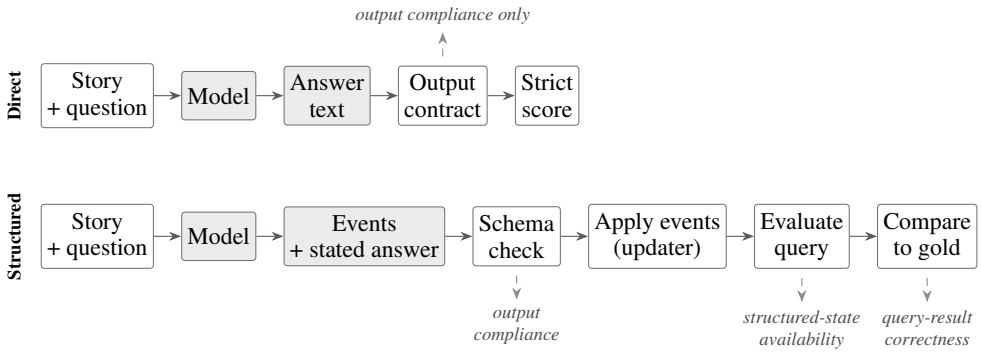


Figure 1: **The structured path exposes additional observable stages.** Shaded boxes mark model-side generation and its outputs; unshaded boxes mark benchmark inputs, validation and computation. On the structured path the benchmark validates the schema, applies the emitted events with its deterministic updater, evaluates the declared query on the resulting state, and compares the result to gold; dashed labels mark where each measured stage is decided. The stated answer in the graph response is requested but ignored by the graph score. Section 3.2 defines the terms.

3.2 Protocol and measurements

Each primary configuration evaluates all 1000 items under each method over five fresh passes at temperature zero; because hosted inference is not bit-reproducible even then, we report the passes and their spread rather than a single run. This gives 5000 *planned observations* per configuration–method cell, 110 000 per method across the cohort, and 220 000 across both methods. Missing calls, rejected outputs, unqueryable graphs and incorrect results are all losses against that fixed denominator. *Strict accuracy*—satisfying the method-specific output contract *and* yielding the gold answer—is the primary end-to-end metric, and no diagnostic is ever substituted for it.

A direct response is *accepted* iff it contains the required `Final answer:` marker and the following text exactly matches one option. A graph response is accepted iff it parses as a JSON object whose `events` payload is a schema-valid event array; the prompt also requests a stated `final_answer`, but acceptance does not require it and the graph score ignores it. An accepted graph is *queryable* iff its events produce a state on which the declared query evaluates. Each failure to reach queryability has a recorded reason; Appendix F separates failures before and after payload acceptance. On the graph path, *output compliance* is measured as payload acceptance (accepted over planned), *structured-state availability* as conditional queryability (queryable over accepted), and *query-result correctness* as correct answers over queryable graphs. *Queryable coverage* (queryable over planned) is a separate unconditional rate outside the factorization. Every conditional structural score is reported beside an unconditional denominator.

Declared, deterministic evaluation rules diagnose rejected responses without replacing strict accuracy; the punctuation rule changes almost no scores (Section 4). Structured-state availability is measured under joint event-and-answer elicitation: in a story-only ablation, only one of four cells met the registered equivalence criterion, so we do not assume it is independent of the prompt (Appendix H). Structural Belief Graph Alignment (SBGA) measures typed-edge precision and recall of emitted graphs against gold; it is a separate diagnostic, not a factorization term.

3.3 Diagnostic controls

Shortcut bound. First- and last-mentioned-container heuristics require no model and score near chance corpus-wide. Selecting the better of the two *separately for each operation on this evaluation corpus* gives per-operation accuracy of 0.860–1.000 and **exceeds all 22 systems on every operation**. Because that choice uses the corpus labels, this is a descriptive same-corpus bound, not a held-out baseline; it is nonetheless enough to show that the operations admit construction-aligned shortcuts. We therefore interpret operation-conditioned contrasts as pipeline behavior rather than ToM competence. The shortcut bound applies only to answer accuracy: the heuristics select options but emit no event graphs, so they do not bound diagnostics of whether a model–prompt system sat-

isfies the graph contract and produces queryable state. Those diagnostics likewise do not establish belief tracking (Appendix E).

Response policy. Abstention is one loss a strict score cannot separate from inability, so we probe it as a control rather than a fourth measurement. Only 2 of 22 configurations abstain spontaneously. Gemma-4-12B refuses 1995 of 2000 second-order direct items, yet returns a scorable graph for 1683 of those same items, so abstention here is a property of the elicitation path rather than of the item (Appendix M).

4 The 22-Configuration Landscape

All results in Sections 4–7 are computed on the current 22-configuration cohort: 1000 items, five fresh passes, 44 reportable model–method cells. Two superseded Gemma/OpenRouter rows are retained for audit and excluded from every aggregate.

Table 1 reports the cohort. It does not collapse the two methods into one score: direct and event-graph elicitation are different pipelines whose ordering changes across systems and operations. Standard deviations are across the five passes and are small throughout, so the movements discussed here are not pass-to-pass noise; deployment is a separate source of variation (Section 7).

Two properties matter for what follows. The two methods disagree sharply—Gemma-4-E2B scores 0.626 direct and 0.003 graph, Ministral-3B 0.152 and 0.609—and across all 22 configurations the direct and graph orderings correlate only at Spearman $\rho = 0.416$, so the elicitation path is part of the measured system (Appendix N). And at least one declared acceptance rule raises ten direct scores, two of them from a reported strict zero.

What each declared rule changes. Table 7 (Appendix D) aggregates every declared rule. On the direct path, salvage raises 9 of 22 scores and changes 17 competition ranks ($\tau_b = 0.7115$); the two borrowed rules raise 10 each and change 19 and 18 ranks ($\tau_b = 0.6508$). The OpenToM-derived rule also *lowers* one score (Ministral-14B, $0.7128 \rightarrow .7112$), so a declared rule is not simply a more lenient scorer. On the graph path, salvage alone raises 21 of 22 scores (10 ranks, $\tau_b = 0.8417$). The event-time fallback raises only three, but sharply: it reconstructs missing event times from array order, which lifts Gemma-4-E2B's graph score from 0.003 to 0.654, a larger intervention than ordinary parsing tolerance. A check anchors these comparisons: terminal-punctuation tolerance should move almost nothing, and it does not—20 of 22 systems are exactly flat and no rank changes. Only two move at all, one of them by 0.0002. That near-null result argues against a generic permissiveness explanation for the larger movements. This rule was adopted after an audit of the scorer found the artifact, so its narrowness is a measured property of applying it uniformly to all 22 configurations rather than a prediction about which ones it would spare.

The declared rule registry. Every declared rule is applied uniformly to every eligible configuration, never enabled selectively (Appendix G), and two of the four direct rules are not ours: they port the matching logic of the published OpenToM and FANToM evaluators. The largest recoveries come from those borrowed rules—for Llama-3.2-1B our internal salvage rule recovers nothing, while each borrowed rule recovers 0.577—so the best declared rule column does not reduce to an in-house parser tuned on our own failures.

5 Direct and Structured Pipelines Differ Across Operations

Table 2 reports all ten balanced operations, using $20{,}000$ cluster-bootstrap draws over stories. We resample stories rather than items because 125 item pairs ask different questions of the same narrative. **Four large contrasts survive a Holm correction across the ten operations** (adjusted $p \leq 0.001$, the resolution floor at this draw count, under all eight declared seeds): memory and reality favor direct answering by 0.168 and 0.119, and both second-order no-ToM strata favor the graph path by 0.206 and 0.249. No other contrast survives. Second-order 1 ToM excludes zero pointwise but not after adjustment (Holm $p = 0.21$–0.22 across seeds), and second-order 0 ToM is seed-sensitive even pointwise; we treat both as inconclusive and interpret neither direction. The result is heterogeneity: a single cohort-wide average would hide large, adjustment-robust differences of opposite sign, and

Table 1: **The best declared evaluation rule raises the measured score for 10 of 22 configurations on the direct path and all 22 on the graph path.** Evaluation rules are the declared direct acceptance rules and the graph-path salvage and event-time fallback, which reconstructs missing event times rather than only relaxing parsing. Rows are ordered by direct strict accuracy, the primary metric; rule values are diagnostics and never replace it. "Best declared rule" is the maximum over that path's declared rules, subscripted with the rule that achieved it only where it exceeds strict (ties take the first rule in declared order): direct SV salvage, PN punctuation, OT OpenToM-derived, FT FANToM-derived; graph GS salvage, TF event-time fallback. Per-rule values are in Table 8. "Strict payload acc." is event payloads accepted under the strict contract over planned graph observations. ^{P} marks a host-pinned replacement, * a dedicated Hugging Face endpoint. Pass-level standard deviations are ≤ 0.0139; Appendix B gives host and provenance per row.

	Direct		Event graph		
System	strict	best declared rule	strict	best declared rule	strict payload acc.
Phi-4	0.7390	0.7390	0.8470	0.8676 GS	0.9348
Ministral-8B	0.7374	0.7524 SV	0.8066	0.8130 GS	0.9694
Nemotron-Nano-9B*	0.7302	0.7302	0.6298	0.6474 GS	0.8926
Gemma-3n-E4B	0.7268	0.7268	0.4948	0.5550 GS	0.5542
Ministral-14B	0.7128	0.7128	0.8280	0.8534 GS	0.9274
Gemma-3-4B*	0.7126	0.7126	0.5712	0.8230 GS	0.6552
Gemma-4-E4B*	0.6994	0.6994	0.1102	0.5878 TF	0.1696
Qwen3-14B*	0.6910	0.6910	0.7238	0.7430 GS	0.9298
Qwen3-8B	0.6398	0.6398	0.7130	0.7340 GS	0.8874
Gemma-4-E2B*	0.6262	0.6262	0.0030	0.6538 TF	0.0072
Granite-4.1-8B*	0.6194	0.6194	0.6166	0.7166 GS	0.7366
Granite-H-Micro	0.6112	0.6156 SV	0.0040	0.3724 TF	0.0196
SmolLM3-3B*	0.5720	0.5720	0.1402	0.1894 GS	0.5090
Gemma-3-12B	0.5460	0.7004 SV	0.7946	0.8204 GS	0.9362
Gemma-4-12B*	0.4362	0.4362	0.8958	0.8990 GS	0.9908
Llama-3.2-3B^{P}	0.4016	0.5406 OT	0.3644	0.5070 GS	0.6520
Ministral-3B	0.1518	0.6548 FT	0.6088	0.7166 GS	0.6618
Llama-3.1-8B^{P}	0.1038	0.6000 OT	0.6586	0.7826 GS	0.7420
Gemma-3-1B*	0.0632	0.5538 FT	0.0000	0.0010 GS	0.0000
Qwen2.5-7B*	0.0114	0.6798 SV	0.3770	0.6444 GS	0.4832
Llama-3.2-1B*	0.0000	0.5772 OT	0.1500	0.3610 GS	0.2120
Phi-4-mini*	0.0000	0.6102 SV	0.2962	0.6694 GS	0.3652

configurations disagree even where the mean contrast is large (on memory, 6 of 22 still score higher on the graph path). Appendix J gives the per-seed and draw-count audit.

We do not claim to explain this pattern, but the graph-path factorization localizes it. Payload acceptance is similar across operations (0.568–0.622), so every operation carries a large compliance loss, and it is lower on each ToM stratum than on each no-ToM stratum. The differences become larger after acceptance. Accepted reality graphs are always queryable and return gold on 0.990 of queries; memory graphs are almost always queryable (0.976) but correct less often (0.929); no-ToM strata are queryable on 0.893–0.902 and correct on 0.972–0.989; the ToM strata are queryable on only 0.712–0.756 and correct on 0.882–0.894. These conditional rates pool configurations by how often each reaches the stage, so they are descriptive and not comparable with the direct column, which also varies by operation; no single stage determines a contrast's sign. Format restrictions can themselves degrade reasoning (Tam et al., 2024), but our prompts differ in more than format, so isolating causes requires factorial ablations.

6 Most Graph-Path Failures Occur Before a Query Can Be Evaluated

The factorization, term by term. Section 3.2 defines strict graph accuracy as the product of acceptance, structured-state availability and query-result correctness; every term is recoverable from the retained per-example rows, and the identity closes exactly on all 22 configurations. Of $110\,000$ planned graph observations, $53\,168$ succeed strictly and $56\,832$ fail. Of those failures, $43\,820$ (77.1%) are rejected payloads and 9951 (17.5%) are accepted payloads that yield no queryable state, so $53\,771$ of $56\,832$ failures (94.6%) occur before a query returns a result. Only 3061 (5.4%) are queryable graphs whose deterministic result differs from gold; once a graph is queryable it usually

Table 2: **Direct and structured pipelines differ by operation, in both directions.** Values are strict accuracy, means across the 22 configurations; intervals are pointwise 95% cluster bootstraps that resample story texts while keeping this cohort fixed (20,000 draws). Differences are graph minus direct and compare pipelines, not an isolated causal variable; "Wins" counts configurations whose graph score is strictly larger. **Bold**: survives Holm adjustment across the ten operations under all eight seeds; ‡: excludes zero pointwise only; †: seed-unstable pointwise verdict. "Shortcut" is the better of two positional heuristics (F first-, L last-mentioned container), selected per operation *on this same corpus*: a same-corpus bound, not a held-out baseline, which **exceeds every measured system on every operation**. Operation names are inherited source-generator labels (Section 3); Appendix J gives the bootstrap audit.

Operation	Direct	Graph	Graph − direct	95% CI	Wins	Shortcut
Memory	0.7320	0.5639	−0.168	**[-0.1962, -0.1409]**	6/22	1.000 F
Reality	0.7307	0.6112	−0.119	**[-0.1392, -0.1005]**	6/22	1.000 L
First-order 0, no-ToM	0.5656	0.5369	−0.029	[-0.0671, +0.0109]	11/22	0.920 L
First-order 0, ToM	0.3685	0.4043	+0.036	[-0.0105, +0.0819]	12/22	1.000 F
First-order 1, no-ToM	0.5758	0.5486	−0.027	[-0.0669, +0.0140]	12/22	0.950 L
First-order 1, ToM	0.3892	0.3791	−0.010	[-0.0527, +0.0313]	11/22	0.980 F
Second-order 0, no-ToM	0.3164	0.5221	+0.206	**[+0.1637, +0.2470]**	16/22	0.860 L
Second-order 0, ToM	0.4105	0.3677	−0.043	[-0.0865, +0.0001] †	10/22	1.000 F
Second-order 1, no-ToM	0.2943	0.5429	+0.249	**[+0.2146, +0.2825]**	16/22	0.930 L
Second-order 1, ToM	0.4043	0.3567	−0.048	[-0.0928, -0.0028] ‡	10/22	1.000 F

returns gold (53 168 of 56 229). Acceptance is also the most dispersed term, 3.1 times as dispersed as query-result correctness (Appendix F).

Payload acceptance at the extremes. Table 1's strict payload-acceptance column shows the range: Gemma-4-E2B accepts only 36 of 5000 payloads and Gemma-3-1B none, while Gemma-4-12B pairs payload acceptance 0.991 with the cohort's best graph score, 0.896.

Structural alignment and a prompt–metric mismatch. SBGA scores the 66 180 accepted payloads against gold edges. On the corrected three-family score, precision exceeds recall in 20 of the 21 scorable configurations (Table 11), a descriptive pattern of *omitted* rather than fabricated state. That correction repairs the clearest apparatus-side failure in this study, one we introduced ourselves. Gold graphs encode containment as a `container_in` edge, but the structured-generation prompt *instructs models not to emit them*, and the scorer counted those edges in its gold denominator anyway. Models were penalized for following the benchmark's own output instruction, so the resulting unanimity is compliance rather than convergence: across the current cohort's 22 graph cells the strict SBGA view carries 132,360 gold `container_in` edges, **zero** are matched, and **0 of 21** scorable configurations emit even one.

That zero is not evidence that models cannot represent containment; it is near-perfect adherence to an instruction, scored as failure. Because `container_in` is 132,360 of the 438,309 pooled strict-view gold edges (30.2%) and matches none, pooled four-family recall cannot exceed 0.698; observed pooled recall is 0.627. At a different aggregation level, excluding the family moves the median configuration's recall from 0.650 to 0.931. This three-family score is a corrective rescore adopted after the prompt–metric mismatch was identified, not a preregistered primary result, and the four-family figures are retained for comparison (Table 10). The alternative repair—inferring container scenes from the physical edges models do emit—is identified but not applied. Both repairs preserve the intended specification; retaining the mismatched denominator would instead score adherence to the prompt as structural failure.

Rejected payloads are therefore the dominant observed failure stage, followed by accepted graphs with no queryable state. This is a statement about the structured path, not a claim that these stages limit every method, operation or system. The prompt asks models only for physical events (enter, leave, stay, put, move), and the benchmark's updater infers who witnessed each move from presence events. An omitted enter or leave event can therefore change a derived belief while leaving the graph queryable, so the stages locate *where* a loss becomes observable, not its cause. Requesting perception output explicitly is an untested design proposal: it would make the reported witness state

explicit and separately scoreable when a model supplies it, but would not guarantee that it is supplied or correct.

7 Robustness and Scope

Cross-dataset transfer. The acceptance gap is not confined to our templates. OpenToM and FANToM are independently constructed; we stage both as two-option forced choice so chance remains 0.50. For eight systems, three predicted clean and five predicted to carry a specific artifact before the runs, we test whether each system's direct-path gap status transfers. On OpenToM, **6 of 8 classifications transfer**; both misses are systems that gained an external gap, and none lost one. On FANToM, **4 of 8 transfer**: every system with an internal gap retains it, while all four without one acquire one. FANToM is therefore a boundary condition rather than a replication—its options average 35.3 words against OpenToM's 1.8, so verbatim output becomes a transcription task for every system. External calls were not host-pinned (Appendix C).

Prospective prompt test. A preregistered, frozen intervention (van Miltenburg et al., 2021) tested whether rule recovery reflects contract failure. The v3 prompts change only the worked examples, demonstrating the output format instead of describing it. For each of nine treatment cells, the preregistration predicted that strict accuracy would rise to within ± 0.10 of the recovery its declared rule had already measured; five controls with no measured gain under the rule targeted by their arm were to stay within ± 0.05. On a matched 300-item subset at three passes, six treatments landed in band, one overshot and two missed. Three of five controls stayed in band against the registered full-corpus baselines, and four of five against baselines recomputed on the same 300 items after the run; Qwen3-8B, outside the band under both, fell by 0.081. **The registered joint criterion failed either way**, so the test is partial support, not confirmation of the mechanism. Llama-3.2-1B never emits the required marker in 900 calls while often stating a candidate answer in prose, so compliance loss is real but heterogeneously repairable. Matched-subset sensitivity and the unpinned routed cells are detailed in Appendix I.

Deployment sensitivity. Ten model–method cells were served by more than one upstream host; six of them, three models on both paths, were remeasured with the host pinned. Strict accuracy moved by up to 0.215, or 0.167 excluding one host that returned 3207 empty responses, in no consistent direction, and per-pass variance fell in five of six. Serving configuration therefore changes measured accuracy and pass variance, so deployment provenance is part of the evaluation unit, and superseded routed rows are excluded from every current aggregate (Appendix B).

Frontier endpoints. Same-lineage checks suggest the output-contract gap shrinks to near zero at three host-pinned frontier endpoints, but these comparisons also change training data and serving and do not isolate scale (Appendix K).

8 Limitations

Benchmark validity. StructToM-Bench is template-derived and measures controlled story-based belief tracking rather than naturalistic or functional ToM, and within individual operations positional shortcuts selected on the same corpus outperform every evaluated system; held-out-template validation remains to be done. Its gold events are deterministic conversions from structured source data, not an independent annotation set; blinded, author-recruited human adjudication covers only the 31 disagreements, as a two-way choice. Repeated stories and questions limit lexical diversity, and no contamination search has been run.

Deployment and causal identification. The cohort is purposive and mixes routed and dedicated serving, so family, scale, training and deployment are not separable causal factors and the ranking does not isolate model weights; host pinning repairs known cells but recovers neither historical model revisions nor runtime precision. Public weights for the three Ministral configurations were unavailable at audit time, so they are excluded from every self-hosted comparison. The primary comparison uses one direct prompt and one graph prompt; the prompt ablations show that results are prompt-sensitive but do not establish generality across prompt designs. The two paths also

differ in entity inventory, option presentation, output contract and computation, so their operation-conditioned contrast is descriptive until factorial ablations isolate those components.

Statistical and interpretive scope. The story-bootstrap intervals condition on the fixed 22-configuration cohort; they do not quantify variation across models or deployments. The prospective prompt test is partial support: its registered joint criterion failed, it used 300 items and three passes, most cells were routed, and one in-band treatment (Gemma-4-E4B) is only 0.0009 inside the band under the matched sensitivity check. The best declared rule depends on a finite registry of declared rules, and Granite-H-Micro's overshoot shows that a rule score can understate what a prompt change recovers. Declared rules require independent semantic validation, and response-policy results are partly coupled to corpus construction and reported as controls. We claim no scaling law, architecture effect, universal advantage for structured elicitation, latent capability recovered by any declared rule, naturalistic ToM, contamination-free data, or identification of the three observable stages as independent causal mechanisms.

9 Discussion and Conclusion

A single theory-of-mind accuracy obscures where an evaluation succeeds or fails, because it is measured through a stack of model, prompt, output contract, scorer and deployment rather than model weights alone. StructToM-Bench separates three observable stages of that stack, and the separation changes what the scores mean: declared evaluation rules alter measured scores and rankings; most structured-path failures occur before a query can be evaluated; and a prompt–metric mismatch in our own scorer manufactured an apparent structural failure. The strict score still answers which configurations satisfy the complete contract; the diagnostics say where a loss entered, and none of this licenses replacing strict accuracy with the most favorable diagnostic. For future evaluations, we recommend:

1. Report the planned denominator, unconditional strict accuracy and, for structured outputs, acceptance, queryability given acceptance and correctness given queryability.
2. Identify the evaluated model–prompt–scorer–deployment system.
3. Disclose acceptance sensitivities against a near-null control specified in advance; our punctuation rule was a later diagnostic, not such a control.
4. Select shortcut rules on a development split and evaluate them on an untouched test split, or call the result a same-corpus bound.
5. Exclude superseded rows from current aggregates.

Reproducibility statement

The anonymized supplementary archive (83 MB) contains the benchmark items, executable gold events and adjudicated labels; the exact prompt templates and run configurations; the scoring, analysis and table-generation code with a pinned dependency lock; the response records behind the reported results, with each model's full output, for the 22-configuration cohort, the host-pinning comparisons, the story-only ablation and the frontier endpoints; the derived artifacts, supersession and host metadata, and preregistrations; and the paper source. One command regenerates the primary offline derived artifacts from the response records. In a clean extraction with dependencies installed from the pinned lockfile, 18 derived results matched the bundled artifacts, and one further sensitivity check passed. The analyses required no provider calls, and every generated table passed its consistency check. The external-dataset arms, the permission-menu response-policy probes and the prospective prompt test retained only per-cell aggregates; these are included as recorded but cannot be regenerated from responses.

AI use statement

Generative AI tools were used to propose and critique paper framings, provide feedback on research methodology, support offline result interpretation and evidence reconciliation, draft and edit manuscript text, identify and summarize literature, and assist with analysis code. Human authors reviewed the AI-assisted work and checked paper-facing empirical claims against retained artifacts. The authors take responsibility for the final text, claims, code, and artifacts.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-0590. URL https://papers.neurips.cc/paper_files/paper/2025/hash/1967e0fc3aa6cbbace562f5cb8e3954e-Abstract-Datasets_and_Benchmarks_Track.html.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. Tom-ssi: Evaluating theory of mind in situated social interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32264–32289, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1642. URL https://aclanthology.org/2025.emnlp-main.1642/.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 31539–31558. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1522. URL https://aclanthology.org/2025.acl-long.1522/.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 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)*, pp. 15959–15983, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Cosimo Galeone, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. When correct isn't usable: Improving structured output reliability in small language models. *arXiv preprint arXiv:2605.02363*, 2026. URL https://arxiv.org/abs/2605.02363.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 13518–13529, 2023. doi: 10.52202/075280-0595. URL https://arxiv.org/abs/2306.15448.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11532–11547. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.685. URL https://aclanthology.org/2024.findings-acl.685/.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19794–19809, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1105. URL https://aclanthology.org/2024.emnlp-main.1105/.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 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*, pp. 14397–14413, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121. URL https://www.pnas.org/doi/10.1073/pnas.2405460121.

Tom Kouwenhoven, Michiel T. van der Meer, and Max J. van Duijn. Traces of social competence in large language models. In *Proceedings of the 30th Conference on Computational Natural Language Learning (CoNLL)*, pp. 742–759. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.conll-main.45/.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598/.

Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. Are llms good at structured outputs? a benchmark for evaluating structured output capabilities in llms. *Information Processing & Management*, 61(5):103809, 2024. doi: 10.1016/j.ipm.2024.103809. URL https://www.sciencedirect.com/science/article/pii/S0306457324001687.

Deep Pankajbhai Mehta. Format sensitivity index: Token-controlled prompt wrapper robustness and schema compliance in LLM benchmarking. *arXiv preprint arXiv:2607.09665*, 2026. URL https://arxiv.org/abs/2607.09665.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2392–2400, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1261. URL https://aclanthology.org/D18-1261/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82091–82130. PMLR, 2025. URL https://proceedings.mlr.press/v267/riemer25a.html.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.248. URL https://aclanthology.org/2022.emnlp-main.248/.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2304.15004.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13960–13980. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.acl-long.780/.

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 *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11324.

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. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.138. URL https://aclanthology.org/2024.eacl-long.138/.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1520–1528, 2025. doi: 10.1609/aaai.v39i2.32143. URL https://ojs.aaai.org/index.php/AAAI/article/view/32143.

Adil Soubki and Owen Rambow. Machine theory of mind needs machine validation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18495–18505, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-acl.951/.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14815–14823, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.880. URL https://aclanthology.org/2024.findings-acl.880/.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z. URL https://www.nature.com/articles/s41562-024-01882-z.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Haibo Tong, Zeyang Yue, Feifei Zhao, Erliang Lin, Lu Jia, Ruolin Chen, Yinqian Sun, Qian Zhang, and Yi Zeng. Cogtom: A comprehensive theory of mind benchmark inspired by human cognition for large language models. In *Proceedings of the 64th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pp. 31383–31425. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1448. URL https://aclanthology.org/2026.acl-long.1448/.

Sean Trott, Samuel M. Taylor, Cameron Robert Jones, James A. Michaelov, and Pamela D. Rivière. Language statistics and false belief reasoning: Evidence from 41 open-weight LMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39816–39833. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1849. URL https://aclanthology.org/2026.acl-long.1849/.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 613–623. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.51. URL https://aclanthology.org/2021.naacl-main.51/.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8292–8308, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.451. URL https://aclanthology.org/2024.acl-long.451/.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 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*, pp. 10691–10706, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717/.

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. Del-tom: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11383–11397, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.573. URL https://aclanthology.org/2025.emnlp-main.573/.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. Towards dynamic theory of mind: Evaluating llm adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24036–24057, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1171. URL https://aclanthology.org/2025.acl-long.1171/.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.466. URL https://aclanthology.org/2024.acl-long.466/.

Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. Enigmatom: Improve llms' theory-of-mind reasoning capabilities with neural knowledge base of entity states. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13598–13622, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.699. URL https://aclanthology.org/2025.findings-acl.699/.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9602d22a8c791f23f8e4d1398e3fb5be-Abstract-Conference.html.

Wang Bill Zhu, Qiutong Tony Yi, Robin Jia, and Jesse Thomason. PDDL-mind: Large language models are capable on belief reasoning with reliable state tracking. *arXiv preprint arXiv:2604.17819*, 2026. URL https://arxiv.org/abs/2604.17819.

A Evidence Status

Table 3: Paper-facing evidence status after offline reconciliation.

Evidence block	Status	Permitted use
Primary 22-system core and lenses (App. D, Tables 7–8)	Current	Main strict, compliance, availability, and operation results
Prospective prompt intervention (App. I, Table 14)	Secondary	Partial support; routed cells not pinned; controls reported under both baselines
Response-policy probes (App. M, Table 17)	Secondary	Model-side control; not repeated on pinned hosts
Provider before/after comparison	Historical audit	Deployment sensitivity; never current ranking
Six pinned frontier cells (App. K, Table 16)	Secondary	Same-lineage scope evidence; `--strict-hosts` passed
DeepSeek-V4-Pro rung	Excluded	Host-unattributable, and no same-lineage small rung; not used for any scale claim
External arms (App. C, Tables 5–6)	Provisional	Internal side recomputed; external calls not host-pinned
Synthetic construct validity	Current	Offline accounting over checked-in corpora; boundary for benchmark claims

B Model and Deployment Metadata

Host-pinning comparisons in full. The sharpest of the ten multi-host cells is a graph cell mixing a working host with 3207 empty Cloudflare responses, whose strict score moved $0.1490 \rightarrow 0.3644$ on pinning. Across the six before/after comparisons absolute strict change reaches 0.215 with that failure and 0.167 without, while the best disclosed direct evaluator moves at most 0.04. Per-pass standard deviation falls in five of the six cells and by up to $9.5\times$ (Llama-3.1-8B direct, $0.0123 \rightarrow 0.0013$); in the sixth it rises by 0.0005. This is consistent with host mixing contributing substantially to run-to-run noise, independently of whether pinning raises or lowers a given score. The primary synthesis and the declared acceptance lenses regenerate byte-for-byte from stored artifacts without provider calls; that claim does not extend to arms not repeated under the current serving policy. The machine-readable comparison records the pinned host for each of these six cells in `artifacts/analysis/host-comparison/host_comparison.json`; the frontier host and available precision declarations are recorded separately in `artifacts/analysis/offline-evidence-reconciliation/evidence_manifest.json`.

Table 4: The 22 primary configurations. Parameter counts are read directly from each repository's `safetensors` index on Hugging Face—the weights actually loaded—rather than from a model card or a model name, and are given in billions. Three systems use per-layer embeddings, for which the publisher-reported *effective* count is shown in parentheses beside the total. Two systems are hybrid or mixture-of-experts architectures (`GraniteMoeHybrid`, `NemotronH`); the remainder are dense. *Provenance* is *pinned* for host-pinned replacements, *dedicated* for a dedicated Hugging Face endpoint, and *routed* otherwise. Every configuration was evaluated at five passes.

System	Params (B)	Host	Provenance	Run date
Gemma-3-12B	12.19	OpenRouter	routed	2026-07-28
Gemma-3-1B	1.00	HF endpoint	dedicated	2026-07-29
Gemma-3-4B	4.30	HF endpoint	dedicated	2026-07-29
Gemma-3n-E4B	7.85 (4B eff.)	OpenRouter	routed	2026-07-28
Gemma-4-12B	11.96	HF endpoint	dedicated	2026-07-29
Gemma-4-E2B	5.12 (2.3B eff.)	HF endpoint	dedicated	2026-07-29
Gemma-4-E4B	8.00 (4.5B eff.)	HF endpoint	dedicated	2026-07-29
Granite-4.1-8B	8.79	HF endpoint	dedicated	2026-09-14
Granite-H-Micro	3.19	OpenRouter	routed	2026-07-28
Llama-3.1-8B	8.03	OpenRouter	pinned	2026-08-25
Llama-3.2-1B	1.24	HF endpoint	dedicated	2026-09-14
Llama-3.2-3B	3.21	OpenRouter	pinned	2026-08-25
Ministral-14B	13.95	OpenRouter	routed	2026-07-28
Ministral-3B	3.85	OpenRouter	routed	2026-07-27
Ministral-8B	8.92	OpenRouter	routed	2026-07-28
Nemotron-Nano-9B	8.89	HF endpoint	dedicated	2026-07-29
Phi-4	14.66	OpenRouter	routed	2026-07-27
Phi-4-mini	3.84	HF endpoint	dedicated	2026-07-29
Qwen2.5-7B	7.62	HF endpoint	dedicated	2026-09-14
Qwen3-14B	14.77	HF endpoint	dedicated	2026-09-14
Qwen3-8B	8.19	OpenRouter	routed	2026-07-27
SmolLM3-3B	3.08	HF endpoint	dedicated	2026-07-29

C External Arms in Full

Testing independently constructed stimuli asks whether the measured acceptance gap is confined to StructToM-Bench's templates. Persistence supports cross-stimulus robustness of the gap; it does not make the gap an intrinsic system property, because system-stimulus and serving interactions remain. OpenToM and FANToM are independently constructed and not descended from FixedToMi (Xu et al., 2024; Kim et al., 2023); both were staged into a native two-option forced choice with every gold answer inside its option set, so chance remains 0.50 on both sides and scoring is identical to the internal protocol—strict primary, and the same maximum over all four declared acceptance rules on each side. Two of the eight internal baselines were superseded by host-pinned replacements after these arms ran; the comparison below is computed against the pinned values, and every verdict is unchanged. The external calls themselves were not host-pinned, so for those same two systems the comparison crosses a serving boundary that only a re-run can close.

How to read Tables 5 and 6. *Roster category* is the artifact family predicted for each system *before* the external runs—clean, bare-answer, punctuation, or t-field. It is a four-way prediction, not a gap flag. The *outcome* column tests binary direct-path gap status only: a classification *transfers* when $(\text{internal gap} > \varepsilon) = (\text{external gap} > \varepsilon)$ with $\varepsilon = 0.02$, so a system predicted to carry an artifact can still *differ* if its internal gap sits below ε. This is a predeclared descriptive classification, not causal transfer: it records whether the acceptance gap reappears on independently constructed stimuli, not why. "Gap" is the best declared lens minus strict, maximised over salvage, OpenToM-derived, FANToM-derived and punctuation. Internal columns are current reconciled cohort values on StructToM-Bench; [s] marks the two systems whose internal row is a host-pinned replacement, and the other six are routed or dedicated deployments as recorded in Appendix B. Both external arms ran 300 examples at three passes and were *not* host-pinned, so the two marked comparisons put a host-pinned internal value against an unpinned external call.

Systems were sorted into gap classes *before* the external runs. On OpenToM, **six of eight pre-declared classifications transfer**, and both misses fail conservatively: one develops only a 0.025 gap against a 0.02 threshold, and one *gains* an external gap rather than losing an internal one. No system lost a gap it had. A newly acquired external gap indicates a system-stimulus interaction rather than invariance, and we report it as such; a system *losing* a gap it had internally would have been the disconfirming outcome, and none did.

Table 5: **OpenToM: 6 of 8 pre-declared classifications transfer.** A classification *transfers* when internal and external direct-path gap status agree at $\varepsilon = 0.02$; *differs* records disagreement, which a system predicted to carry an artifact can show whenever its internal gap sits below ε. The reading conventions, the four roster categories, the gap definition and the provenance of the ^{s} marker are given once for both tables above.

		Internal (current cohort)			External (unpinned)			
System	Predeclared roster category	strict	best	gap	strict	best	gap	Outcome
Phi-4	clean, no gap	0.7390	0.7390	0.0000	0.8089	0.8122	0.0033	transfers
Ministral-14B	clean, no gap	0.7128	0.7128	0.0000	0.7511	0.7756	0.0245	*differs*
Qwen3-8B	clean, no gap	0.6398	0.6398	0.0000	0.6756	0.6889	0.0133	transfers
Granite-H-Micro	*t*-field artifact	0.6112	0.6156	0.0044	0.5667	0.6511	0.0844	*differs*
Llama-3.2-3B^{s}	punctuation artifact	0.4016	0.5406	0.1390	0.1044	0.6156	0.5112	transfers
Ministral-3B	bare-answer artifact	0.1518	0.6548	0.5030	0.1656	0.6289	0.4633	transfers
Llama-3.1-8B^{s}	punctuation artifact	0.1038	0.6000	0.4962	0.2511	0.6300	0.3789	transfers
Qwen2.5-7B	bare-answer artifact	0.0114	0.6798	0.6684	0.0100	0.6889	0.6789	transfers

C.1 FANToM: WHERE THE CONTRACT ITSELF BECOMES THE BURDEN

FANToM makes a boundary condition visible that OpenToM cannot. Under the same scoring code, **four of eight** classifications transfer—but the count is the least informative part of the result. All four systems with an internal gap above 0.02 keep their gaps (0.326, 0.387, 0.284, 0.272). The four "failures" are systems that *gained* a gap: three previously clean systems acquire small ones (0.030–0.077) and one acquires a large one (0.408, at strict 0.000).

The answer-length evidence supports a transcription-burden interpretation: FANToM's options average 35.3 words (median 29, max 189) against OpenToM's 1.8 (median 1), and the strict contract requires reproducing one verbatim. Two systems score strict 0.000 while scoring 0.32–0.39 under the lenses. FANToM is therefore consistent with the general claim—a verbatim-match contract can understate measured ability to the point of reporting zero—but provides little evidence for the system-specific claim, since the contract imposes a substantial burden across systems. The consequence generalizes: identical scoring code does not guarantee a score represents the same burden across datasets. For the same reason we report no FANToM graph arm as structured-state availability evidence; its dialogue format does not match the physical-event schema.

Table 6: **FANToM: 4 of 8 pre-declared classifications transfer.** A classification *transfers* when internal and external direct-path gap status agree at $\varepsilon = 0.02$; *differs* records disagreement, which a system predicted to carry an artifact can show whenever its internal gap sits below ε. The reading conventions, the four roster categories, the gap definition and the provenance of the ^{s} marker are given once for both tables above.

		Internal (current cohort)			External (unpinned)			
System	Predeclared roster category	strict	best	gap	strict	best	gap	Outcome
Phi-4	clean, no gap	0.7390	0.7390	0.0000	0.4744	0.5156	0.0412	*differs*
Ministral-14B	clean, no gap	0.7128	0.7128	0.0000	0.4800	0.5567	0.0767	*differs*
Qwen3-8B	clean, no gap	0.6398	0.6398	0.0000	0.4244	0.4544	0.0300	*differs*
Granite-H-Micro	*t*-field artifact	0.6112	0.6156	0.0044	0.0000	0.4078	0.4078	*differs*
Llama-3.2-3B^{s}	punctuation artifact	0.4016	0.5406	0.1390	0.0800	0.3522	0.2722	transfers
Ministral-3B	bare-answer artifact	0.1518	0.6548	0.5030	0.0000	0.3867	0.3867	transfers
Llama-3.1-8B^{s}	punctuation artifact	0.1038	0.6000	0.4962	0.1244	0.4089	0.2845	transfers
Qwen2.5-7B	bare-answer artifact	0.0114	0.6798	0.6684	0.0978	0.4233	0.3255	transfers

D ACCEPTANCE DIAGNOSTICS IN FULL

Table 7: **Declared acceptance diagnostics on the current cohort.** Generated from the evidence artifacts. *Raised*, *lowered* and *flat* count configurations whose score under the rule is above, below or equal to strict accuracy (unrounded, tolerance 5×10^{-5}). *Ranks changed* counts configurations whose competition rank differs from the strict ordering; τ_b compares the two orderings. The *best declared rule* row takes each configuration's maximum over that path's rules. The punctuation rule was adopted after a scorer audit and is reported as a check.

Path	Rule	Raised	Lowered	Flat	Ranks changed	τ_b
Direct	salvage	9/22	0/22	13/22	17/22	0.7115
Direct	OpenToM-derived	10/22	1/22	11/22	19/22	0.6508
Direct	FANToM-derived	10/22	0/22	12/22	18/22	0.6508
Direct	terminal punctuation *(check)*	2/22	0/22	20/22	0/22	1.0000
Direct	**best declared rule**	10/22	0/22	12/22	18/22	0.6508
Graph	salvage	21/22	0/22	1/22	10/22	0.8417
Graph	event-time fallback	3/22	0/22	19/22	13/22	0.8009
Graph	**best declared rule**	22/22	0/22	0/22	16/22	0.7202

Table 8: **Every declared lens on every eligible configuration.** Strict joint accuracy is the ranking metric; all other columns are diagnostics and never determine ordering. Direct lenses (salvage, OpenToM-derived, FANToM-derived, terminal punctuation) apply only to `direct_qa` cells and graph lenses only to `event_graph_with_answer` cells; — marks an ineligible cell, which is distinct from a lens that applied and moved nothing. Superseded routed rows are excluded.

	Direct path					Graph path		
Configuration	strict	salvage	OpenToM	FANToM	punct.	strict	salvage	t-fall.
Phi-4	0.7390	0.7390	0.7390	0.7390	0.7390	0.8470	0.8676	0.8470
Ministral-8B	0.7374	0.7524	0.7524	0.7524	0.7374	0.8066	0.8130	0.8066
Nemotron-Nano-9B	0.7302	0.7302	0.7302	0.7302	0.7302	0.6298	0.6474	0.6298
Gemma-3n-E4B	0.7268	0.7268	0.7268	0.7268	0.7268	0.4948	0.5550	0.4948
Ministral-14B	0.7128	0.7128	0.7112	0.7128	0.7128	0.8280	0.8534	0.8280
Gemma-3-4B	0.7126	0.7126	0.7126	0.7126	0.7126	0.5712	0.8230	0.5712
Gemma-4-E4B	0.6994	0.6994	0.6994	0.6994	0.6994	0.1102	0.1132	0.5878
Qwen3-14B	0.6910	0.6910	0.6910	0.6910	0.6910	0.7238	0.7430	0.7238
Qwen3-8B	0.6398	0.6398	0.6398	0.6398	0.6398	0.7130	0.7340	0.7130
Gemma-4-E2B	0.6262	0.6262	0.6262	0.6262	0.6262	0.0030	0.0030	0.6538
Granite-4.1-8B	0.6194	0.6194	0.6194	0.6194	0.6194	0.6166	0.7166	0.6166
Granite-H-Micro	0.6112	0.6156	0.6156	0.6156	0.6112	0.0040	0.0050	0.3724
SmolLM3-3B	0.5720	0.5720	0.5720	0.5720	0.5720	0.1402	0.1894	0.1402
Gemma-3-12B	0.5460	0.7004	0.7004	0.7004	0.5460	0.7946	0.8204	0.7946
Gemma-4-12B	0.4362	0.4362	0.4362	0.4362	0.4362	0.8958	0.8990	0.8958
Llama-3.2-3B	0.4016	0.5380	0.5406	0.5406	0.4016	0.3644	0.5070	0.3644
Ministral-3B	0.1518	0.6450	0.5414	0.6548	0.1520	0.6088	0.7166	0.6088
Llama-3.1-8B	0.1038	0.2894	0.6000	0.5888	0.1470	0.6586	0.7826	0.6586
Gemma-3-1B	0.0632	0.4844	0.5524	0.5538	0.0632	0.0000	0.0010	0.0000
Qwen2.5-7B	0.0114	0.6798	0.6798	0.6798	0.0114	0.3770	0.6444	0.3770
Llama-3.2-1B	0.0000	0.0000	0.5772	0.5772	0.0000	0.1500	0.3610	0.1500
Phi-4-mini	0.0000	0.6102	0.6102	0.6102	0.0000	0.2962	0.6694	0.2962

E SHORTCUT BASELINES (CORPUS LEVEL)

Table 9: **Label-free shortcut baselines score near chance corpus-wide.** Each baseline applies one fixed rule to all 1000 items without using labels; accuracy is against the executable gold (chance 0.50).

Label-free baseline	Corpus accuracy
Option A (first listed)	0.496
Option B (second listed)	0.504
First-mentioned container	0.529
Last-mentioned container	0.468

The four baselines in Table 9 are corpus-level: each applies one fixed rule to all 1000 items without using labels, and all score near chance. The per-operation bound in Section 3.3 and Table 2 is different: for each operation separately, it keeps whichever of the first- and last-mentioned-container rules scores higher on this same corpus. That selection uses the labels, so it is a same-corpus bound, not a held-out baseline.

F CONTRACT, QUERYABILITY, AND SBGA

Per-operation terms. The pooled acceptance, queryability and correctness rates Section 5 quotes are recorded per operation in `artifacts/analysis/graph-factorization/graph_factorization.json`. Each operation has 11 000 planned observations, and its pooled strict rate reproduces Table 2's graph column.

Dispersion across the three terms. One configuration (Gemma-3-1B) has no accepted graph payloads, so its downstream conditional rates are undefined. Dispersion across the remaining 21 configurations points the same way, and more sharply once the terms are separated. Acceptance spans 0.007 to 0.991, descriptive CV = 0.493; queryability given acceptance spans 0.308 to 0.940, CV = 0.193; and the gold rate given a queryable state spans 0.333 to 0.996 with a median of 0.947 and CV = 0.157. Acceptance is 3.1 times as dispersed as query-result correctness. We read these as relative dispersion and not as a decomposition of variance, which a coefficient of variation does not license; they are descriptive over every evaluated configuration rather than estimates for a model population, and are computed from retained rows with superseded cells excluded.

CONTRACT, QUERYABILITY, AND SBGA IN FULL

Strict contract. A direct response is *accepted* iff it carries the required `Final answer:` marker and the text after it matches one option of the two-option set exactly. A graph response is accepted iff it parses as a JSON object whose `events` payload is a schema-valid event array. The prompt also requests a stated `final_answer`; a missing one fails only its separate answer parse, which is reported as a diagnostic, and neither acceptance nor the graph score depends on it. Anything else—missing marker, unparseable JSON, schema violation, missing call—is a loss against the fixed planned denominator.

Queryability. An accepted graph is *queryable* iff applying its events to the deterministic updater yields a well-formed final state on which the example's declared query can be evaluated. Non-queryable observations each receive a recorded skip reason: missing predictions and event-parse or event-schema failures prevent payload acceptance, while graph-construction and unsupported-state failures arise after acceptance. The scorer also defines an *unscored label* status for items outside the gold audit; it does not occur in the current cohort, where all 1000 labels are accepted. *Queryable coverage* is queryable observations over planned observations—equivalently, the fraction that are neither skipped nor missing. *Payload acceptance*, the quantity Table 1 reports and the denominator SBGA scores against, counts accepted event payloads over planned observations and is no smaller than queryable coverage. Every conditional structural score in this paper is reported beside its unconditional denominator.

SBGA. Structural Belief Graph Alignment scores exact typed-edge set matching. The gold graph is a set of typed edges over four families: `agent_in`, `object_in`, `container_in`, and `believes_object_in`. Writing G_τ and P_τ for the gold and predicted edge sets of family τ, Structural Belief Graph Alignment scores exact set matching per family,

$$\mathrm{prec}_\tau = \frac{|G_\tau \cap P_\tau|}{|P_\tau|}, \qquad \mathrm{rec}_\tau = \frac{|G_\tau \cap P_\tau|}{|G_\tau|},$$

with F_1 their harmonic mean. Headline figures are computed over three families; `container_in` is excluded, and the exclusion is a measurement decision we document rather than a difficulty judgement.

Why `container_in` is excluded. Section 6 reports the measured consequence; this is the cause of the measurement defect. Gold carries `container_in` 2,000 times per 1,000 examples, and **no system in the corpus emits a single one**. It encodes a static placement ("the envelope is in the garden"), not an action, and every system treats it as scene-setting rather than an event. Without a container-to-scene placement the updater defaults the scene to UNKNOWN, so `container_in(drawer, unknown)` can never match `container_in(drawer, den)`; the container name is recovered correctly and only the scene is unresolved. The measured consequence: with `container_in` accounting for 30.2% of pooled gold edges and matching none, pooled four-family recall cannot exceed 0.698; observed pooled recall is 0.627 (Table 10). Scoring a family the prompt never requests measures the prompt, not the system. The principled repair—inferring container scenes from the physical edges models do emit—is not applied here; this exclusion is the honest interim.

Table 10: **Edge-family alignment, strict view.** The four pooled columns and the two median columns are *different estimands* and must not be read as one aggregation: pooled recall is matched/gold over scored records and is exact, while the medians are macro medians across configuration-level five-pass means. Pooled four-family recall is 0.6270 against a macro median of 0.6500, and the total row leaves the medians blank because medians do not sum. Pooled *precision* does not exist here: the artifact records gold and matched counts but no predicted-edge counts. `container_in` precision is *undefined* rather than zero—no system predicted one such edge, so its precision denominator is zero—while its recall is defined at 0.0000 because every cell has a nonzero gold denominator (per-cell counts range from 72 to 9,908).

	Pooled over scored records				Macro median over cells	
Edge family	gold	share	matched	recall	precision	recall
`believes_object_in`	138,113	31.5%	120,070	0.8694	0.9660	0.9158
`container_in`	132,360	30.2%	0	0.0000	—	0.0000
`agent_in`	101,656	23.2%	90,780	0.8930	0.9511	0.9196
`object_in`	66,180	15.1%	63,977	0.9667	0.9966	0.9966
All four diagnostic families	438,309	100.0%	274,827	0.6270	—	—

Table 11 turns the same scored records around, reporting overall precision and recall per configuration. The ordering makes the omission pattern legible: the four systems with the largest precision–recall gaps are also the four with the lowest structural recall, so they emit a small amount of state accurately rather than a large amount carelessly.

Table 11: **Precision exceeds recall in 20 of 21 configurations.** Overall strict-view precision and recall per configuration over the three headline families, with `container_in` excluded; values are five-pass means and the difference column makes the omission pattern inspectable. The sole exception, Llama-3.1-8B, reverses by 0.0005—on the same scale as pass-to-pass variation in both quantities (pass-level standard deviations 0.00045 for precision and 0.00067 for recall). Differences are computed from unrounded means. Across the cohort, the predominant descriptive pattern is lower recall than precision, consistent with omission being more common than fabrication.

Configuration	Precision	Recall	$P - R$
SmolLM3-3B	0.8684	0.3609	$+0.5075$
Llama-3.2-3B	0.9157	0.5827	$+0.3330$
Gemma-4-E2B	0.8381	0.6333	$+0.2048$
Granite-H-Micro	0.8147	0.6944	$+0.1203$
Gemma-4-E4B	0.9842	0.9063	$+0.0779$
Ministral-14B	0.9879	0.9359	$+0.0519$
Qwen3-14B	0.9665	0.9260	$+0.0405$
Gemma-3-4B	0.9752	0.9400	$+0.0352$
Llama-3.2-1B	0.7473	0.7132	$+0.0341$
Ministral-8B	0.9617	0.9305	$+0.0312$
Phi-4-mini	0.9707	0.9398	$+0.0310$
Qwen3-8B	0.9819	0.9550	$+0.0269$
Qwen2.5-7B	0.9313	0.9092	$+0.0221$
Nemotron-Nano-9B	0.8690	0.8498	$+0.0192$
Ministral-3B	0.9483	0.9298	$+0.0184$
Granite-4.1-8B	0.9634	0.9468	$+0.0166$
Phi-4	0.9981	0.9833	$+0.0148$
Gemma-4-12B	0.9966	0.9832	$+0.0134$
Gemma-3-12B	0.9742	0.9630	$+0.0111$
Gemma-3n-E4B	0.9803	0.9698	$+0.0104$
Llama-3.1-8B	0.9731	0.9736	-0.0005

G Declared Lens Registry

Table 12: Declared acceptance lenses. All are deterministic, fail-closed, and applied to retained responses without new provider calls. The two borrowed rules adapt the matching logic of OpenToM's `check_answer_for_fg_location` and FANToM's `evaluate_list_q`; in our implementation both read the answer from the line after a `Final answer:` or `Answer:` marker when one is present and from the whole response otherwise, and both leave ties unscored. Strict scoring remains the primary metric.

Lens	Path	Rule	Source
salvage (SV)	direct	A response with no `Final answer:` marker gets one attempt at a single-line answer of at most six words; rows that already parsed strictly are untouched	internal
punctuation (PN)	direct	Strip trailing punctuation before exact match	internal
OpenToM (OT)	direct	Argmax lexical overlap against the option set; unscored on a tie	borrowed, OpenToM harness
FANToM (FT)	direct	Dual-sided substring containment; unscored when zero or more than one option matches	borrowed, FANToM harness
`t` fallback	graph	Fill each event's `t` from array position, only when no event carries an integer `t`	internal
graph salvage	graph	Drop schema-invalid event rows only when they cannot affect the query path; retain the valid physical events	internal

Borrowed-rule provenance. The OpenToM rule adapts `compute_lexical_overlap` and `check_answer_for_fg_location` from `seacowx/OpenToM`, commit `3f22b662`. The FANToM rule adapts `evaluate_list_q` from `skywalker023/fantom`, commit `1cae6fa`. Full commit URLs and our marker, option and tie policies are documented in the bundled `external_lenses.py`; neither rule reproduces the source benchmark's overall metric.

H The Story-Only Prompt Ablation

The structured prompt asks for two things: an event array *and* a directly stated answer. Whether the emitted state depends on that second request was preregistered before any story-only call, with three dated amendments recording every change to scope, the authoritative measurement copy, and one infrastructure fault.

Design. A second prompt was derived from the deployed one by changing three lines: the instruction to answer, the `final_answer` field in the output schema, and the closing reminder. Nothing else differs. Group A—*primary*—re-ran **both** arms contemporaneously on dedicated endpoints we deployed ourselves, one endpoint per cell, $K{=}5$ over all 1000 items, so the only difference between arms is the prompt. Group B ran the story-only arm for the remaining eight OpenRouter cells against their frozen with-answer baselines; because those arms are 48 days apart on another host, prompt effect and serving drift are confounded there by construction, and Group B is reported as breadth evidence that cannot overturn Group A.

Decision rule, fixed before data collection. Paired differences (story-only − with-answer) on five measurements, with 95% cluster-bootstrap intervals resampled over `story_id` (2000 resamples). Equivalence required the interval to lie inside $\pm.05$ on at least four of five measurements for at least three of four Group A cells. No p-values.

Result: the registered equivalence criterion was not met. One of four cells satisfied the criterion.

Table 13: Group A paired differences, story-only minus with-answer, with 95% cluster-bootstrap intervals over `story_id`. A quantity is *inside* when its interval lies wholly within $\pm.05$; **bold** marks an interval that crosses the band. The registered rule has **five criteria**, not six: SBGA precision and recall together form one criterion, counted inside only when *both* are. Equivalence required at least four of five criteria for at least three of four cells.

	acceptance	query|acc	queryable cov.	SBGA prec	SBGA rec	gold|qry
Qwen2.5-7B — 5 of 5 criteria inside						
Δ	$+0.010$	-0.011	$+0.003$	-0.019	$+0.003$	$+0.016$
95% CI	$[-0.020, +0.039]$	$[-0.035, +0.012]$	$[-0.024, +0.030]$	$[-0.034, -0.004]$	$[-0.023, +0.030]$	$[-0.010, +0.043]$
Granite-4.1-8B — 3 of 5 (acceptance, queryable cov. outside)						
Δ	-0.030	$+0.001$	-0.027	$+0.009$	-0.021	$+0.016$
95% CI	$[\mathbf{-0.058}, \mathbf{-0.005}]$	$[-0.012, +0.014]$	$[\mathbf{-0.054}, \mathbf{-0.002}]$	$[+0.003, +0.017]$	$[-0.047, +0.003]$	$[-0.002, +0.036]$
Qwen3-14B — 3 of 5 (queryable cov., SBGA outside)						
Δ	$+0.019$	$+0.018$	$+0.033$	$+0.017$	$+0.041$	$+0.021$
95% CI	$[+0.004, +0.035]$	$[+0.003, +0.033]$	$[\mathbf{+0.016}, \mathbf{+0.052}]$	$[+0.010, +0.024]$	$[\mathbf{+0.027}, \mathbf{+0.056}]$	$[+0.007, +0.035]$
Llama-3.2-1B — 0 of 5						
Δ	-0.053	-0.020	-0.049	-0.020	-0.043	-0.110
95% CI	$[\mathbf{-0.081}, \mathbf{-0.026}]$	$[\mathbf{-0.080}, \mathbf{+0.040}]$	$[\mathbf{-0.075}, \mathbf{-0.023}]$	$[\mathbf{-0.062}, \mathbf{+0.021}]$	$[\mathbf{-0.065}, \mathbf{-0.022}]$	$[\mathbf{-0.190}, \mathbf{-0.029}]$

Only Qwen2.5-7B met the criterion. Granite-4.1-8B and Qwen3-14B fail in the interval tails only—every point estimate lies inside the band—while Llama-3.2-1B fails on all five. Two distinctions matter for reading this table, and conflating them overstates the result. *Point-estimate direction* is uniform within a cell: all six Qwen3-14B differences are positive, all six Llama-3.2-1B differences negative. *Statistically supported direction* is not symmetric: all six Qwen3-14B intervals exclude zero, whereas only four of six do for Llama-3.2-1B—queryability given acceptance and SBGA precision both span it. The observed effects are therefore system-specific and bidirectional, which is weaker than claiming every direction is established for every quantity.

Group B places four of eight cells inside the band, and its two largest excursions are informative about *deployment* rather than the prompt: Llama-3.2-3B emitted schema-valid *empty* event arrays

on 35.0% of its accepted payloads in the frozen with-answer baseline and none in the story-only arm recorded on the same host, and Qwen2.5-7B emitted them on 20.7% of accepted payloads on OpenRouter and none on our own endpoint, where host, runtime, precision and date all changed together. An empty array satisfies the schema, so it is scored as an accepted payload while carrying no state; Appendix B records the affected cells.

Clustering sensitivity. The registered analysis resamples by `story_id`, which is unique per example, so it treats all 1000 items as independent clusters. The corpus in fact contains 875 distinct narratives: 125 pairs of items share a story and differ only in the question asked. We therefore recomputed every interval clustering on the story text instead. Empirically, mean interval width rose by 0.0024 and 23 of the 24 quantities widened, one narrowing by 0.0011; **no per-cell count and no verdict changed**—Qwen2.5-7B remains 5 of 5, Granite-4.1-8B and Qwen3-14B 3 of 5, Llama-3.2-1B 0 of 5, and the registered criterion remains unmet at 1 of 4. The registered `story_id` result stands as the primary analysis because it is what was fixed in advance; the sensitivity is reported so the conclusion is not resting on that choice.

Consequence. The preregistered response to a failed equivalence criterion is terminological, and it is applied throughout this paper: the middle factorization term is *structured-state availability under joint event-and-answer elicitation*. The cohort is not re-run—the with-answer prompt produced all existing evidence, including arms that cannot be cheaply rebuilt—and no result elsewhere in the paper is restated as a prompt-independent property of any model.

I Prospective Prompt Test in Full

The registration (`artifacts/analysis/paper-gap-11/PREREGISTRATION.md`) was committed before any call. It fixed nine treatment cells, each with a measured compliance artifact, and five controls with no measured gain under the rule targeted by their arm; two graph controls do gain under graph salvage, which their arm did not test. The v3 prompts differ from v1 only in their worked examples, which demonstrate the required output literally instead of in prose shorthand; task framing, field definitions and scope rules are byte-identical. Each treatment's predicted strict score is the recovery its declared rule had measured on the full 1000-item v1 run. The run used a stratified 300-item subset at three passes, so predictions and outcomes are measured on different item sets. As a sensitivity check, `rule_sensitivity_300.json` recomputes each treatment's registered rule on those 300 items from the same v1 run; each source run first reproduces its registered baseline and prediction on the full corpus. The count stays at six of nine in band, one overshot and two missed, but Gemma-4-E4B's margin shrinks from 0.009 to 0.0009. The frozen registered verdicts stand; the check measures only the effect of this particular 300-item subset, and Gemma-4-E4B's near-boundary result shows that its in-band verdict is fragile. The registered falsifier was controls *rising* by more than 0.05, which would indicate a generally better prompt rather than a contract repair. Against the registered baselines Qwen3-14B rose by 0.053; against the matched baselines, a correction made after the run and before interpretation, it rose by 0.047. Qwen3-8B fell under both. Three treatment cells ran on dedicated endpoints and the remaining eleven cells were routed and not host-pinned.

Table 14: **Prospective prompt test, all 14 registered cells.** Generated from `final_14cell_comparison.json`. The v3 prompts change only the worked examples, demonstrating the output format instead of describing it. Each treatment's *predicted* strict score is the recovery its declared rule had measured on the full 1000-item v1 run; outcomes are three passes on a matched 300-item subset. A treatment is *in band* within ± 0.10 of its prediction. As a sensitivity check, *Rule@300* recomputes the same rule on the 300 items the run used, from the same v1 run; it does not replace the registered verdict. The five controls have no measured gain under the rule targeted by their arm and must stay within ± 0.05; their change is shown against the registered full-corpus baseline and against a baseline recomputed on the same 300 items after the run. 6 of 9 treatments are in band (6 of 9 against Rule@300); 3 of 5 controls stay in band against the registered baseline and 4 of 5 against the matched one.

Treatments

System	Serving	Predicted	v3	v3 − pred.	Verdict	Rule@300	v3 − R@300
Ministral-3B	routed	0.6500	0.7056	+0.0556	in band	0.6647	+0.0409
Gemma-4-E4B	dedicated	0.5900	0.6811	+0.0911	in band	0.5820	+0.0991
Qwen2.5-7B	routed	0.6700	0.6756	+0.0056	in band	0.6613	+0.0143
Phi-4-mini	dedicated	0.6100	0.6700	+0.0600	in band	0.6327	+0.0373
Gemma-4-E2B	dedicated	0.6500	0.6667	+0.0167	in band	0.6593	+0.0074
Llama-3.1-8B	routed	0.6400	0.6022	−0.0378	in band	0.6053	−0.0031
Granite-H-Micro	routed	0.3700	0.5100	+0.1400	overshot	0.3700	+0.1400
Llama-3.2-3B	routed	0.5700	0.4011	−0.1689	missed	0.5813	−0.1802
Llama-3.2-1B	routed	0.5700	0.0000	−0.5700	missed	0.5740	−0.5740

Controls

System	Serving	v3	Δ registered	Δ matched	In band (registered / matched)
Qwen3-14B	routed	0.7933	+0.0529	+0.0466	no / yes
Phi-4	routed	0.7333	−0.0057	+0.0140	yes / yes
Ministral-14B	routed	0.7100	−0.0028	+0.0047	yes / yes
Gemma-3n-E4B	routed	0.4744	−0.0204	−0.0089	yes / yes
Qwen3-8B	routed	0.5722	−0.0676	−0.0811	no / no

J Operation-Effect Bootstrap Audit

Stage pairing and denominators. Group A cells contribute `direct_qa` from their cohort stage and `event_graph_with_answer` from their ablation stage, under the declared pairing; no `counterfactual` rows enter this analysis. Denominators are fixed at 5000 observations per cell and method—every example appears exactly five times—with provider errors counted as strict failures rather than dropped. The cluster bootstrap resamples the 875 distinct story texts, not the 1000 story–question pairs.

The operation effects use pointwise, unadjusted 95% cluster-bootstrap intervals with 20,000 draws. The larger count is not what makes the stable verdicts stable; it is what makes an unstable one visible. At 1,000 and 2,000 draws every operation returns the same verdict under all eight seeds, and second-order 0 ToM excludes zero under 8 of 8; at 5,000 it falls to 6 of 8 and at 20,000 to 5 of 8. The instability is therefore already visible at 5,000 draws. The interval is the reason: at the reported seed its upper bound is -0.0019 at both 1,000 and 2,000 draws, $+0.0006$ at 5,000 and $+0.0001$ at 20,000, with the width rising from 0.0826 to 0.0866 and settling by roughly 5,000. A 1,000-draw resample estimates its tail quantiles from the 26th and 976th order statistics, which is too coarse to place an endpoint this close to zero, so the interval comes out too narrow and the operation appears secure under every seed. We therefore read the low-draw agreement as an artifact of the estimator rather than evidence of a stable effect. This stability analysis diagnoses numerical sensitivity; it is not a multiplicity correction and provides no simultaneous-coverage guarantee.

Table 15: **Seed stability of the operation-effect verdicts.** Each mark is one bootstrap seed, in the order 1, 7, 42, 101, 999, 2024, 20260916, 31337: $+$ and $-$ mark an interval excluding zero above or below it, $\cdot$ an interval spanning zero. Verdicts are decided on *unrounded* endpoints while the numeric columns come from the stored endpoints, rounded to four decimals, so the two can differ in their last digit. *Nearest-bound distance* is the distance from zero to the nearer bound, ranged over the eight seeds: an exclusion margin where the verdict excludes zero, and how far zero sits from the nearer edge where it does not. Four effects clear zero by more than 0.10; second-order 1 ToM also excludes zero under all eight seeds, but its nearest-bound distance is only 0.0025–0.0034. The pre-substitution columns are a *superseded diagnostic* cohort, not a result: with estimator, seeds and draw count held fixed, they show the two borderline operations' verdict counts are also sensitive to cohort composition, while the other eight retain theirs.

	Current cohort			Pre-substitution (superseded)		
Operation	Seeds	n	Nearest-bound distance	Seeds	n	Nearest-bound distance
Second-order 1 no-ToM	++++++++	8/8	0.2134–0.2146	++++++++	8/8	0.2115–0.2127
Second-order 0 no-ToM	++++++++	8/8	0.1630–0.1638	++++++++	8/8	0.1572–0.1582
Memory	−−−−−−−−	8/8	0.1406–0.1413	−−−−−−−−	8/8	0.1566–0.1574
Reality	−−−−−−−−	8/8	0.1003–0.1010	−−−−−−−−	8/8	0.1204–0.1209
Second-order 1 ToM	−−−−−−−−	8/8	0.0025–0.0034	·····$\cdot$	0/8	0.0000–0.0010
Second-order 0 ToM	−·−·−−·−	5/8	0.0000–0.0008	−−−−−−−−	8/8	0.0004–0.0021
First-order 1 ToM	·····$\cdot$	0/8	0.0308–0.0322	·····$\cdot$	0/8	0.0333–0.0346
First-order 1 no-ToM	·····$\cdot$	0/8	0.0133–0.0140	·····$\cdot$	0/8	0.0115–0.0124
First-order 0 no-ToM	·····$\cdot$	0/8	0.0104–0.0114	·····$\cdot$	0/8	0.0169–0.0181
First-order 0 ToM	·····$\cdot$	0/8	0.0101–0.0115	·····$\cdot$	0/8	0.0098–0.0111

An implementation audit also found that bootstrap streams had initially been keyed by an operation's position in a list. Re-keying each stream by a stable operation-name identifier left the point estimate for second-order 0 ToM unchanged but moved its upper interval bound across zero at the fourth decimal place. The operation excludes zero under 5 of 8 seeds but not under the reported seed, so we classify it as inconclusive. The five pointwise effects retain their verdict under all eight seeds and under the stable-key implementation.

Multiplicity. Seed stability is a Monte Carlo check, not a multiplicity correction, so we also adjust across the ten operations. From the same $20{,}000$ draws and name-keyed streams we compute two-sided percentile-bootstrap p-values and apply Holm's step-down procedure. With $\hat{\Delta}_b^*$ the graph-minus-direct difference in draw b of $B = 20{,}000$, and $p_{(1)} \le \cdots \le p_{(m)}$ the ordered values over the $m = 10$ operations,

$$k_{\le 0} = \#\{b : \hat{\Delta}_b^* \le 0\}, \qquad k_{\ge 0} = \#\{b : \hat{\Delta}_b^* \ge 0\},$$
$$p = \min\Big\{1, \frac{2(1 + \min(k_{\le 0}, k_{\ge 0}))}{B + 1}\Big\},$$
$$p_{(i)}^{\text{Holm}} = \max_{j \le i} \min\big\{1, (m - j + 1)p_{(j)}\big\}.$$

The smallest attainable raw value is $2/(B+1) \approx 1.0 \times 10^{-4}$. Memory, reality and both second-order no-ToM contrasts reach the smallest attainable adjusted value, $p_{\text{Holm}} \le 0.001$, under all eight seeds; no draw of the $20{,}000$ crosses zero for any of them. Second-order 1 ToM, whose pointwise interval excludes zero with its nearer endpoint 0.0025–0.0034 from it, has raw p between 0.034 and 0.038 and p_{Holm} between 0.20 and 0.23 across seeds; second-order 0 ToM has p_{Holm} between 0.23 and 0.26, and every first-order contrast exceeds 0.50. Bonferroni-adjusted 99.5% intervals agree: at the reported seed they are $[-0.2090, -0.1295]$, $[-0.1474, -0.0930]$, $[+0.1465, +0.2628]$ and $[+0.1978, +0.2961]$ for the four surviving contrasts, and $[-0.1117, +0.0178]$ for second-order 1 ToM. Section 5 therefore interprets only the four adjusted contrasts. Every value here is frozen in `artifacts/analysis/operation-effects/operation_effects_multiplicity.json`, regenerated by `scripts/build_operation_effects_artifact.py` with `--multiplicity`, which recomputes the Holm adjustment from the stored raw p-values before writing.

K SCALING LADDER IN FULL

The small end of each lineage is where compliance failures concentrate. On the current cohort the strict-to-maximum declared recovery gap at the smallest rung is 0.668 for Qwen (Qwen2.5-7B, salvage), 0.577 for Llama (Llama-3.2-1B, OpenToM-derived), 0.503 for Mistral (Ministral-3B, FANToM-derived), and 0.000 for Nemotron, which carries no artifact. Three of the four have positive gaps; two use borrowed rules and Qwen uses internal salvage.

This axis asks whether those failures persist at a larger endpoint *within the same lineage*. It is a same-lineage comparison, not a scaling law: task performance, training data and serving conditions are not separable across rungs, and prior work shows input-format sensitivity need not vanish with scale (Sclar et al., 2024).

Three lineages with host-pinned frontier endpoints. Table 16 partitions the measured strict-score movement from the smallest rung to the frontier rung into the movement under declared recovery and the change in the acceptance gap. This is an algebraic comparison, not a causal attribution. All three frontier cells were remeasured with the upstream host pinned and fallbacks disabled, and each passed `verify_run_integrity --strict-hosts` on 10,000 raw responses with a single recorded host.

Table 16: **Most apparent strict-score improvement within these lineages is associated with the disappearance of the output-contract gap rather than with increased recoverable score.** *Strict* Δ is the strict-to-strict movement a conventional scaling plot would show; *Recovery* Δ is the movement under maximum declared recovery; *Contract* is the difference between them, and *Share* its fraction of the strict movement. Share exceeds 100% when the recovered score falls while the strict score rises: for Mistral, Recovery Δ is -0.009, so the contract component ($0.494 + 0.009 = 0.503$) exceeds the strict movement. All three frontier endpoints are host-pinned and gated: Llama-4-Maverick and Qwen3.5-397B-A17B to `DeepInfra`, whose recorded quantization is fp8, and Mistral-Large-2512 to its first-party host, whose numerical precision is undisclosed. Serving precision is therefore not held constant across rungs, and the small rungs (bf16 or dedicated) and the frontier rungs differ in serving conditions as well as scale.

Lineage	Rungs	Strict Δ	Recovery Δ	Contract	Share
Llama	3.1-8B $\rightarrow$ 4-Maverick	$+0.681$	$+0.185$	0.496	72.8%
Qwen	2.5-7B $\rightarrow$ 3.5-397B	$+0.746$	$+0.078$	0.668	89.6%
Mistral	Ministral-3B $\rightarrow$ Large-2512	$+0.494$	-0.009	0.503	101.8%

In every lineage the majority of the apparent gain is associated with closure of the output-contract gap rather than movement in recoverable score: 73%, 90%, and 102%. The Mistral case is the limiting one. From Ministral-3B to Mistral-Large-2512, strict accuracy rises 0.494 while the declared recovered score *falls* 0.009, so the contract component slightly exceeds the entire strict movement. A conventional scaling curve drawn on strict accuracy would report that lineage as substantially improving; measured under a declared recovery rule it does not improve at all.

At the frontier rung the compliance gap is 0.0004 for Maverick and exactly 0.000 for both Qwen3.5-397B and Mistral-Large-2512, where every declared rule returns the strict value. We do not observe the compact-model compliance artifact at these three endpoints.

This is analogous to the metric-induced capability shifts studied by Schaeffer et al. (2023), but here the discontinuity is localized empirically to the output contract rather than inferred from score curves.

The same dissociation inside the cohort. Re-measuring Qwen2.5-7B on an endpoint we deploy ourselves moves its direct strict score from 0.091 to 0.011, which read alone would suggest a collapse in task performance. It is not one. Strictly parsed responses fall from 553 to 57 of 5000 while *every* parsed response on the new endpoint is correct; the unparsed majority are bare option strings such as `container`, answering the question without the required `Final answer:` marker. Maximum declared recovery is 0.680 on the new endpoint against 0.670 on the old. The answer

content is unchanged and the output contract compliance is not, so the movement is a deployment-sensitive compliance effect measured on the direct path, not evidence about belief tracking. Host, runtime, precision and date changed together here, so the cause is the deployment stack rather than any one of its components.

What this does not show. Three further frontier cells are pinned and gated from an earlier round. For DeepSeek-V4-Flash, pinning moved strict $+0.065$ while maximum declared recovery moved -0.009, and the gap collapsed from 0.074 to 0.000. The gap therefore appears in the routed measurement but not in the pinned remeasurement; this comparison does not isolate which serving-stack component caused the difference. For Qwen3.6-35B-A3B, pinning lifted graph parse rate only to 0.707, far below every other frontier endpoint. One cell does not identify whether architecture or active parameter count explains this difference.

Compliance scaling and task-performance scaling are distinct. The Mistral lineage also separates them directly. Its 675B rung scores *below* its own 8B rung on both measures: -0.092 strict and -0.107 under declared recovery, so recoverable performance is non-monotonic within a single lineage under a fixed protocol. We therefore distinguish two findings. Output-contract artifacts largely disappear at the frontier endpoints of every lineage we measured. The recoverable-score trajectory is not monotonic: a larger endpoint can remain flat or perform worse than a smaller same-lineage model.

We claim neither a general scaling law nor an architecture effect, and we do not identify the recoverable-score trajectory with reasoning capability; the shortcut baselines of Section 3.3 and the pending validation in Section 8 both bound that reading. Two of the six host-pinned frontier cells are attributable to a host but not to a numerical precision, for unrelated reasons. Mistral-Large-2512 is served only first-party with undeclared quantization, which is a property of the provider. Nemotron-3-Super-120B-A12B retained no endpoint metadata at all, which is a gap in our own record rather than the provider's and is repairable by re-fetching it.

L Instrument Construction Detail

Question types and record structure. Memory questions request an earlier object location, reality questions the final world location, first-order questions one agent's belief, and second-order questions one agent's attribution of another's belief. The eight factorial first- and second-order strata keep their generator-defined names and are not reinterpreted after model results are seen. Each record carries typed agents, objects, containers and scenes; a time-ordered event sequence that updates the physical world and the beliefs of observing agents; and query metadata whose targets are fixed in the record and never inferred from model output.

Corpus accounting. The 1000 story–question pairs comprise 875 unique normalized story texts and 870 unique normalized question texts; 125 pairs share a narrative and differ only in the question asked. These are accounting figures, not a claim of linguistic diversity: the data remain template-controlled, and no training-contamination search has been performed. The operation-effect bootstraps resample the 875 story texts rather than the 1000 pairs. The preregistered story-only analysis instead resamples the item-level `story_id` it registered, and reports a story-text sensitivity check (Appendix H). Three of the 22 configurations use per-layer embeddings, for which Appendix B reports total and effective parameter counts separately; the stated 1.00B–14.77B range refers to total parameters.

Gold-label audit provenance. Project records date the audit behind the 969-of-1000 concordance to 2026-05-30 and describe a high-reasoning OpenRouter rerun. The code defaults identify `deepseek/deepseek-v4-flash`, prompt `gold-label-audit` v1, reasoning effort *high*, and temperature zero; raw requests and per-example provider metadata were not retained, so this configuration is reconstructed rather than recorded and cannot be independently verified from the released responses. This LLM adjudication is distinct from the blinded human protocol registered below, whose result is reported separately and never pooled with it.

A preregistered test of that concordance. Treating the above as a hypothesis rather than a result, we registered a prediction before collecting any independent judgements: *blinded adjudicators, shown story, question and two candidate answers without provenance, will select the graph-derived label on at least 26 of the 31 disagreements*. The adjudication instrument contains 91 items—the 31 disagreements plus 60 agreement controls, six per operation, with presentation order randomised—and adjudicators are admitted to the primary analysis only if they answer at least 51 of the 60 controls correctly, since controls have a known answer on which source, graph and LLM already agree. Outcomes are declared in advance as confirmed ($\geq$ 26), partial (16–25) or disconfirmed ($\leq$ 15), with UNSURE counted in the denominator. We report the result whichever way it falls, and if it is disconfirmed we revisit the affected labels and recompute every analysis that depends on the gold set. The registration, the frozen item set and its hashes are released with the benchmark under `artifacts/analysis/label-audit-2026-09-09/`, which also carries the scoring implementation and its amendment history.

Result: confirmed. Seven human adjudicators were run, and recruitment closed at the registered five qualifying sets, whose control accuracy ranged from 52 to 60 of 60. Of the two sets not counted, one failed the gate ($46/60$). The other passed it ($58/60$) but was not exported from the annotation form, and its provenance could not be confirmed. A resubmission of that set made after feedback on control items was also set aside. The item-level majority resolves **all** 31 **disagreements to the graph-derived label**, with none to the source label, no ties and no UNSURE responses. The registered outcome is therefore *confirmed*. Twenty-nine items are unanimous and two are 4–1; 153 of 155 individual votes choose the graph label. Three adjudicators choose it on all 31 items and two on 30. Adding the two uncounted sets leaves the majority at 31 of 31. Each adjudicator chooses option 1 on 14–16 of the 31 items, and the source label is option 1 on 16, so the result does not reflect a position preference. The operational labels are unchanged. The test is narrower than an independent annotation. Adjudicators chose between two given answers, were recruited by the authors rather than from a crowdsourcing pool, and judged only the 31 disagreements and 60 controls, which establish task competence rather than validate the 969 agreement items. The registration, the 91-item set, the answer key and the scorer were committed on 2026-09-13, before any response was collected, and the released item file has the same SHA-256 as that commit (`cf4d6ae0...`); the adjudicator responses and the scoring output were committed with the result on 2026-09-22. Raw responses, pseudonymised as A1–A7, per-adjudicator scores and the scoring output are released in the same directory.

One trace, end to end. A single item makes the stages concrete. Suppose a model emits the gold-consistent graph below; the trace is illustrative rather than a retained response. The story is *Chloe entered the bedroom. Liam entered the bedroom. The radish is in the bottle. The bottle is in the bedroom. Chloe exited the bedroom. Liam moved the radish to the crate. The crate is in the bedroom.*, asked as *Where will Chloe look for the radish?* over the options {`bottle`, `crate`}. On the direct path the model returns text and the only observable is whether it is accepted and equals the gold option. On the structured path:

Stage	What is observed
Model emits	`enter(Chloe,bedroom,t=1)`, `enter(Liam,bedroom,t=2)`, `put(radish,bottle,t=3)`, `leave(Chloe,bedroom,t=5)`, `move(actor=Liam,object=radish,to=crate,t=6)`, alongside a stated `final_answer` the score ignores. Container-to-scene statements are omitted, as the prompt directs
Compliance	The JSON object parses and its `events` payload is schema-valid, so the graph is *accepted*
Structured-state availability	Applying the events to the benchmark's updater yields a state on which `first_order_belief_location(Chloe, radish)` evaluates, so the graph is *queryable* and counts toward queryable coverage
Query-result correctness	Chloe left before the move, so her belief is not updated and the deterministic query returns `bottle`, which equals the gold answer

A failure at any stage is visible as that stage: a malformed array is a compliance loss, a graph missing the `leave` event yields `crate` and is a query-result loss, and a graph naming no agent at all is

unqueryable and is an availability loss. That second case is worth stating plainly: the underlying defect is an omitted event, yet it surfaces in the conditional query-result term, because the graph remains queryable. The factorization localizes where a loss becomes *observable* in the pipeline; it does not identify the causal origin of the underlying output defect. The direct path returns `bottle` or `crate` and admits none of these distinctions.

M Response Policy Probes in Full

This section gives the complete four-probe account summarised in Section 3.3: probe sizes, per-probe rates, scoring conventions and the partition argument.

Abstention is the one behavior in this cohort that a strict score cannot distinguish from inability: a refused item and a wrong item are both losses against the planned denominator. We treat it as a control rather than a fourth measurement, and the reason is empirical. Taken together, the probes show that hedging depends on the response menu and is higher on second-order than first-order items in the probed systems. They do not establish a latent disposition, item difficulty, or suppressed competence recoverable by rescoring. All figures below are measured on second-order items unless stated otherwise.

Spontaneous hedging is concentrated, not cohort-wide. Under the standard direct contract, only 2 of the 22 configurations abstain on second-order items at all: Gemma-4-12B on 1995 of 2000 responses (0.9975, five passes) and Granite-4.1-8B on 313 (0.1565). The remaining 20 hedge on exactly zero responses. A behavior this skewed is easy to mistake for a property of two models.

Prompting makes hedging widespread. When the response menu explicitly offers "cannot be determined" as an accepted answer, **all** 10 **of** 10 **probed configurations take it** on second-order items—including the nine that never volunteer a refusal under the standard prompt (Table 17).[1] Thus, spontaneous hedging by two systems does not imply that the other probed systems lack the same response option; the elicitation contract determines whether it is expressed.

Hedging is higher on second-order items. In the permission arm, 10 **of** 10 **configurations hedge more on second-order items than on first-order items**—a unanimous ordering we did not predict. The gap is large and system-specific (Table 17). In this probe, hedging is systematically higher on the recursive items rather than uniformly distributed.

[1] Ten configurations were evaluated under both permission arms: nine routed through OpenRouter and Gemma-4-12B on its dedicated endpoint. Each cell is 200 sampled items $\times$ 3 passes; outcomes are scored under bare-answer salvage, because the question is behavioral rather than one of format compliance, with the strict marker rate reported alongside.

Table 17: **Response policy under standard and permission prompts.** Panel A: the two nonzero rows of the 22-configuration cohort, 2,000 second-order responses over 5 passes each; the other 20 are exactly zero. Panel B: the separate 10-configuration permission cohort, a strict subset (12 were not probed), matched by exact model identifier, with rates averaged over 3 passes of 200 planned examples per prefix and unparsed responses counted as losses. Granite-4.1-8B and Granite-H-Micro are distinct configurations; the former was not probed. All ten hedge more on second-order items, spanning 0.4617–1.0000 against 0.0400–0.7183 on first-order items.

Panel A: spontaneous hedging, all 22 configurations

Configuration	Hedges / responses	Rate
Gemma-4-12B	1,995 / 2,000	0.9975
Granite-4.1-8B	313 / 2,000	0.1565
Other 20 configurations	0 / 2,000 each	0.0000

Panel B: permission cohort, 10 configurations

	Volunteered	Permitted hedge rate	
Configuration	hedge rate	first-order	second-order
Gemma-4-12B	0.9975	0.6650	1.0000
Qwen3-14B	0.0000	0.5533	1.0000
Qwen2.5-7B	0.0000	0.7183	0.9950
Llama-3.1-8B	0.0000	0.5100	0.7983
Gemma-3-12B	0.0000	0.0400	0.7667
Ministral-14B	0.0000	0.2117	0.7617
Ministral-8B	0.0000	0.5067	0.6633
Llama-3.2-3B	0.0000	0.3400	0.6467
Gemma-3n-E4B	0.0000	0.5067	0.5767
Granite-H-Micro	0.0000	0.1067	0.4617

For Gemma-4-12B, forced answers are numerically close to chance. When abstention is forbidden, Gemma-4-12B emits zero hedges and scores 0.5050 ± 0.0141 on second-order items against a two-option chance rate of 0.500, and 0.5400 ± 0.0041 on first-order items. This differs from a rejected output contract that hides recoverable answer content: Llama-3.2-1B moves from 0.000 to 0.577 under a declared rule. Both cases lose points under strict scoring, but the probes provide different diagnostic evidence.

But the refusal belongs to the elicitation path, not the item. The same items tell a different story under structured elicitation. Of the 1995 second-order items Gemma-4-12B refuses in the direct path, 1683 yield a scorable graph under the structured contract (0.8436), against a direct answer rate of 0.0025 on that same set; the configuration's cohort-wide graph-path strict accuracy is 0.8958, the best in the cohort, while its direct-path strict accuracy of 0.4362 ranks 15th of 22. Whether this system "refuses theory-of-mind questions" is therefore not a fact about the system or about the items. It is a fact about which of the two paths was used to ask.

Why this stays a control. Because the response categories partition the denominator—correct, wrong, hedge and unparsed sum to one—a negative association between hedging and correctness is close to arithmetic, and we do not report one. The informative quantity is accuracy *among items the configuration chose to answer*, whose cohort median is 0.505 against chance 0.500. Answer-location preference is reported on the same footing: the original location is correct on false-belief items by construction, so location preference is partly coupled to corpus construction. Both remain diagnostics, not independent mechanisms.

N MEASUREMENT ANGLES IN FULL

Table 1 supports several summaries of the same 22 configurations. The direct and graph strict orderings correlate at Spearman $\rho = 0.416$. Three measurement angles have three different leaders; their median best-to-worst rank spread is 6, and 5 configurations appear in at least one angle's top three. These are descriptive comparisons of evaluation pipelines, not corrected capability estimates. The shortcut baselines in Section 3.3 further bound a ToM interpretation.

Table 18: **Three summaries of the current cohort have three different leaders.** Recovery is the best declared *direct-path score*.

Angle	Leader	Value
Direct strict	Phi-4	0.7390
Graph strict	Gemma-4-12B	0.8958
Direct declared recovery	Ministral-8B	0.7524

Table 19: Ranks under the three angles (1 is best; the two tied direct strict zeros receive average ranks). Spread is worst minus best rank, with cohort median 6.

System	Direct	Graph	Recovery	Spread
Gemma-4-12B	15	1	22	21
Gemma-4-E4B	7	19	8	12
Gemma-4-E2B	10	21	13	11
Llama-3.1-8B	18	8	17	10
Qwen2.5-7B	20	14	10	10
Gemma-3-12B	14	5	7	9
Gemma-3n-E4B	4	13	4	9
Granite-H-Micro	12	20	15	8
Gemma-3-4B	6	12	6	6
Llama-3.2-3B	16	15	21	6
Ministral-3B	17	11	11	6
Nemotron-Nano-9B	3	9	3	6
SmolLM3-3B	13	18	19	6
Phi-4-mini	21.5	16	16	5.5
Qwen3-8B	9	7	12	5
Llama-3.2-1B	21.5	17	18	4.5
Granite-4.1-8B	11	10	14	4
Gemma-3-1B	19	22	20	3
Ministral-8B	2	4	1	3
Qwen3-14B	8	6	9	3
Ministral-14B	5	3	5	2
Phi-4	1	2	2	1