Positional Shortcuts in Theory of Mind Benchmarks: Architecture-Dependent Learning Dynamics and Mechanistic Evidence

Anonymous ACL submission

Abstract

Existing Theory of Mind benchmarks contain positional shortcuts: a model can achieve 41–51% accuracy by predicting the first-mentioned container, without any belief reasoning. We introduce ToMForge, a self-verifying story generator certified shortcut-free by a six-test statistical suite, and use it to study what models actually learn when the shortcut is unavailable. Across 13 experiments spanning 5K-parameter transformers to 8B-parameter LLMs, we find that architecture determines shortcut behavior. LSTMs exhibit a developmental ordering where true-belief convergence precedes false-belief convergence by 90 steps ($p = 0.005$, $d = 1.03$), mirroring patterns observed in human children. Transformers show no such ordering. Mamba architectures fall on a gradient between the two: Mamba-1 shows a significant sequential bias ($d = 0.59$–0.77), while Mamba-2 reduces to attention-like levels. Among LLMs, standard models exploit the positional shortcut perfectly (FB=1.000), but chain-of-thought prompting partially unlocks genuine belief tracking in some architectures while *inducing* shortcut exploitation in others. Mechanistic analysis reveals a single attention head selectively responsible for false-belief reasoning (62.2% FB drop vs. 5.3% TB drop upon ablation).

1 Introduction

Theory of Mind—the ability to attribute beliefs to others that differ from reality—has become a central test for language model reasoning (Premack and Woodruff, 1978; Wimmer and Perner, 1983). Benchmark results on ToM tasks have fueled claims that LLMs possess genuine social reasoning (Kosinski, 2024; Strachan et al., 2024), while adversarial evaluations suggest these abilities are fragile (Ullman, 2023; Shapira et al., 2024). This debate rests on an unstated assumption: that the benchmarks measure what they claim to measure.

We show they do not. The two most widely used ToM question-answering benchmarks—ToMi (Le et al., 2019) and Hi-ToM (Wu et al., 2023)—contain a positional shortcut: the correct answer to false-belief questions is the first-mentioned container in 41.2% and 51.1% of stories, respectively, against a chance baseline of 33.3%. A model that learns this heuristic achieves above-chance accuracy without tracking beliefs at all. This is an instance of the broader shortcut learning phenomenon (Geirhos et al., 2020; Gururangan et al., 2018; McCoy et al., 2019): benchmarks that are "right for the wrong reasons."

The consequences of shortcut contamination extend beyond inflated accuracy. If models solve ToM benchmarks through positional heuristics rather than belief reasoning, then comparisons between architectures, scaling analyses, and claims about emergent reasoning capabilities rest on unreliable evidence. Fixing the benchmark is a prerequisite for studying the phenomenon.

With a shortcut-free benchmark in hand, we can ask questions that were previously unanswerable: do different architectures learn ToM differently? Does the learning trajectory depend on the model's inductive bias? And does chain-of-thought prompting help models reason about beliefs, or does it merely rearrange which shortcut they exploit?

We address these questions with three contributions:

1. We identify and quantify positional shortcuts in ToMi and Hi-ToM, and introduce ToMForge, a story generator with a six-test Shortcut Verification Suite that certifies generated stories as shortcut-free (20/20 seeds pass at n=2,000).

2. We demonstrate architecture-dependent learning dynamics using ToMForge stories: LSTMs show developmental ordering ($p =$

0.005, $d = 1.03$) that mirrors human children (Wimmer and Perner, 1983), transformers do not, and Mamba architectures fall on a measurable gradient between the two.

3. We show that LLM shortcut exploitation is regime-dependent: chain-of-thought prompting rescues belief tracking in some models (Gemma 4 E4B: TB 0.000→0.720) while inducing shortcut exploitation in others (Qwen3 8B: FB 0.071→1.000).

2 Related Work

ToM Benchmarks. The Sally-Anne false-belief paradigm (Wimmer and Perner, 1983) has been adapted for NLP as question-answering over short narratives. ToMi (Le et al., 2019) generates stories with controlled belief states; Hi-ToM (Wu et al., 2023) extends this to higher-order beliefs. Subsequent benchmarks diversify the format: BigToM uses causal templates (Gandhi et al., 2023), FANToM embeds ToM in dialogue (Kim et al., 2023), OpenToM adds personality dimensions (Xu et al., 2024), ToMBench provides bilingual coverage (Chen et al., 2024), and ExploreToM generates adversarial stories programmatically (Sclar et al., 2025). These benchmarks vary the *difficulty* of ToM evaluation but do not systematically verify the absence of exploitable shortcuts. ExploreToM is the most closely related: like ToMForge, it generates stories programmatically, but its focus is adversarial difficulty rather than shortcut verification. Our work is complementary: the ToMForge Shortcut Verification Suite provides a verification framework applicable to any generator, including those above.

The LLM ToM Debate. Kosinski (2024) report that GPT-4 solves classic false-belief tasks at human level. Strachan et al. (2024) confirm this across multiple LLMs and find performance approaching human baselines. However, Ullman (2023) show that trivial story modifications—such as changing the container from a box to a bag—collapse performance, and Shapira et al. (2024) argue that models exploit surface cues rather than reasoning, a "Clever Hans" effect. Wilf et al. (2024) propose perspective-taking prompts as an alternative to standard CoT for ToM tasks. Our findings provide a new lens on this debate: when the positional shortcut is quantified and removed, LLM performance profiles change qualitatively, suggesting that prior evaluations confound reasoning ability with shortcut exploitation.

Dataset Artifacts. Shortcut learning is well-documented in NLI, where hypothesis-only baselines (Poliak et al., 2018) and annotation artifacts (Gururangan et al., 2018) inflate benchmark scores. McCoy et al. (2019) show models achieve high accuracy through syntactic heuristics rather than understanding, and Geirhos et al. (2020) provide a cross-domain framework for understanding shortcut learning. Gardner et al. (2021) argue that compositional language data inherently creates spurious correlations. Our contribution extends this line of work to ToM evaluation, identifying a previously unreported positional artifact in two major benchmarks and providing a verification suite to prevent such artifacts in future generators.

3 ToMForge

3.1 Story Generation

ToMForge generates Sally-Anne–style narratives where an actor places an object in a container, leaves the scene, and a second actor moves the object. A question asks where the first actor will look for the object. False-belief stories require tracking the actor's outdated belief; true-belief stories serve as controls where the actor witnesses the move. A typical controlled-mode false-belief story reads: *"The box, the basket, and the crate are in the room. Alice put the ball in the box. Alice left. Bob moved the ball to the basket. Where will Alice look for the ball?"* The correct answer is "the box" (Alice's false belief), not "the basket" (reality).

The key design choice is a three-shape container rotation. Each story involves three containers (e.g., box, basket, crate). The generator cycles the assignment of roles—initial location, destination, and distractor—across stories so that the correct false-belief answer appears in each ordinal position (first, middle, last) with equal probability. This is enforced structurally, not by rejection sampling, so the guarantee is exact regardless of sample size.

Two generation modes are supported. The *controlled* mode uses a fixed 37-token vocabulary with deterministic sentence templates, enabling training of small-scale models (5K–38M parameters) on a consistent distribution. The *natural language* mode produces varied narratives with diverse character names, objects, locations, and phrasing, suitable for LLM evaluation. Both modes pass all six verifi-

Table 1: The ToMForge Shortcut Verification Suite. All tests use standard statistical methodology. Tests 5–6 follow Gururangan et al. (2018).

#	Test	Method	Pass
1	Position heuristic	One-prop. z-test	$p > 0.05$
2	Length bias	Cohen's d	$d < 0.50$
3	Recency	One-prop. z-test	$p > 0.05$
4	Character name	χ^2 GoF	$p > 0.005$
5	Verb cue	Word→ans acc	$<$ ch.+margin
6	n-gram BoW	CV classifier	$\leq$ ch.+margin

cation tests; cross-mode consistency is verified at n=2,000.

3.2 Shortcut Verification Suite

Rather than claiming a specific fix, we provide a verification framework. The suite applies six statistical tests to any set of generated stories:

Tests 1–4 check whether individual surface features predict the answer. Test 1 checks whether the correct container appears disproportionately in a particular ordinal position (first, middle, or last mentioned). Test 2 checks whether stories with different answers differ systematically in length, which would allow a model to predict the answer from story length alone. Test 3 checks whether the most recently mentioned container is disproportionately correct (a recency heuristic). Test 4 checks whether particular character names correlate with particular answers. Tests 5–6 follow the hypothesis-only methodology of Gururangan et al. (2018): Test 5 checks whether any single word predicts the answer above chance, and Test 6 trains a cross-validated bag-of-words classifier on the story text and checks whether it can predict the answer. Tests 1 and 3 are one-sided because only above-chance heuristic rates are exploitable.

The suite certifies ToMForge at n=2,000: 20/20 seeds pass all six tests, with a first-container rate of 0.335 ($p = 0.73$, indistinguishable from chance).

As a demonstration of the suite's generalizability, we applied it to second-order ToM stories and identified a previously unnoticed recency shortcut: the correct answer is the last-mentioned container in 66.7% of false-belief variants ($p < 0.001$). This is a finding, not a limitation—it shows the suite catches shortcuts beyond those it was designed for.

3.3 Benchmark Comparison

We compute the first-container rate—the proportion of false-belief stories where the correct answer is the first-mentioned container—for ToMi, Hi-ToM, and ToMForge (Figure 1). ToMi's rate of 0.412 ($z = 10.60$, $p < 0.001$) and Hi-ToM's rate of 0.511 ($z = 13.03$, $p < 0.001$) both significantly exceed the one-third chance baseline. ToMForge achieves 0.333 ($z = 0.25$, $p = 0.80$), indistinguishable from chance. Prior results on these benchmarks may be partially attributable to positional exploitation rather than belief reasoning.

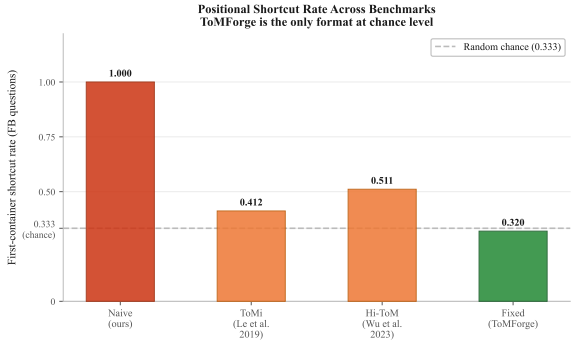

Figure 1: First-container rates across datasets. ToMi (0.412) and Hi-ToM (0.511) significantly exceed the one-third chance baseline, indicating exploitable positional shortcuts. ToMForge achieves 0.333 ($p = 0.80$), indistinguishable from chance. The shortcut rate of a model *trained* on fixed-format stories is 0.320 (Table 2), confirming that neither the data nor the model exploit position.

4 Small-Model Experiments

We train three architectures on ToMForge controlled-mode stories to study learning dynamics when shortcuts are unavailable. All experiments use 3–10 random seeds with statistical testing.

Experimental setup. Three architectures span two model families: Transformer 1L (Vaswani et al., 2017) (5,467 parameters, 1 layer, 2 attention heads, d=16), Transformer 2L (8,747 parameters, 2 layers, 2 heads, d=16), and LSTM 2L (Hochreiter and Schmidhuber, 1997) (62,539 parameters, 2 layers, h=64). All models are trained on 600 ToMForge controlled-mode stories (300 FB, 300 TB) using Adam (lr $= 10^{-3}$) with early stopping when both FB and TB accuracy exceed 0.95 on a held-out set of 200 stories. Training runs on Apple M-series MPS; Mamba experiments (Section 6) run on a Kaggle T4 GPU. We report convergence step—the first training step at which accuracy exceeds 0.80—as the primary metric for learning dynamics.

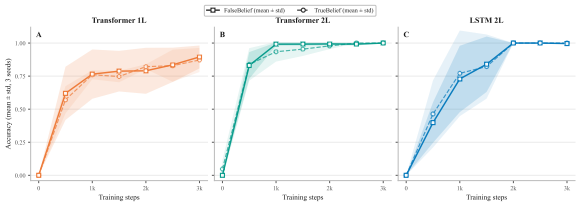

Figure 2: Training curves for three architectures on ToMForge stories (3 seeds each, mean±std bands). LSTM consistently reaches TB accuracy before FB accuracy; transformers show no such ordering.

Table 2: Training results on ToMForge stories (3 seeds each). All architectures solve the task. Naive training produces perfect shortcut exploitation (rate=1.000) with low accuracy; fixed training eliminates the shortcut while achieving high accuracy.

Training Data	FB / TB Acc	Shortcut Rate
Naive format	.131 / .109	1.000±.000
Fixed format	1.000 / 1.000	.320±.018

4.1 Training Dynamics

Three architectures—Transformer 1L (5,467 params), Transformer 2L (8,747 params), and LSTM 2L (62,539 params)—are trained on ToMForge controlled-mode stories. All reach high final accuracy (Table 2), confirming the task is learnable when shortcuts are removed.

The shortcut rate comparison is the core proof: models trained on naive (shortcut-present) stories achieve a shortcut rate of 1.000 with low accuracy (FB=0.131), while models trained on ToMForge stories achieve a shortcut rate of 0.320≈0.333 (chance) with perfect accuracy. The shortcut collapses while accuracy improves, demonstrating that fixed-trained models reason about beliefs rather than exploiting position.

4.2 Developmental Ordering

LSTM shows a significant positive convergence gap: true-belief accuracy reaches criterion 90 steps before false-belief accuracy (mean gap=+90, $p = 0.005$, $d = 1.03$ [large effect; Cohen 1988], 10 seeds; Figure 2). The convergence gap measures a directional asymmetry in learning: a positive gap means the model learns to track what an actor *does* know (true belief) before it learns to track what an actor *falsely* believes (false belief). This mirrors the developmental ordering observed in human children, who pass true-belief tasks around age 3 but false-belief tasks not until age 4–5 (Wimmer and Perner, 1983). The parallel is behavioral, not mechanistic—LSTMs and children solve the task through entirely different processes—but it suggests that sequential processing creates a representational bottleneck that forces a specific learning trajectory.

Transformers show no such ordering (1L: gap=-180, $p = 0.756$; 2L: gap=-100, $p = 0.826$), consistent with their parallel processing of all input positions. Because attention can access the departure event and the question simultaneously, there is no asymmetry between the information needed for true-belief vs. false-belief answers.

Unseen departure verbs do not disrupt the learned representations (FB≈1.0 on unseen verbs), but injecting noise sentences destroys accuracy (FB drops to 0.270, below chance), confirming reliance on structural position rather than lexical cues.

4.3 Mechanistic Analysis

We probe a trained 1L Transformer at multiple levels of abstraction (Figure 3).

Linear probing. A linear classifier trained on frozen representations shows that answer identity is progressively computed through the network: accuracy rises from 0.254 (embedding, near chance=0.250) through 0.432 (post-attention) to 0.494 (post-MLP).

Positional probing. Position-resolved probes reveal that belief-relevant information first emerges reliably at position 8, corresponding to the departure token where the actor leaves and the false belief is created.

Head ablation. Following Vig et al. (2020), we use mean ablation to assess each attention head's contribution. Head 0 is selectively critical for false-belief reasoning: ablating it causes a 62.2 percentage point drop in FB accuracy but only a 5.3 pp drop in TB accuracy, a 12× specificity ratio. Head 1, by contrast, shows comparable drops for both question types (FB: 57.5 pp, TB: 60.7 pp), serving a general role.

4.4 Scaling Laws

The positional shortcut is scale-invariant: on naive stories, the shortcut rate is 1.000 at every scale from 8K to 38M parameters (Figure 4, left). This is expected because the shortcut is a property of the data distribution, not a learned feature—any model that finds the first-mentioned container heuristic will exploit it regardless of capacity. The result

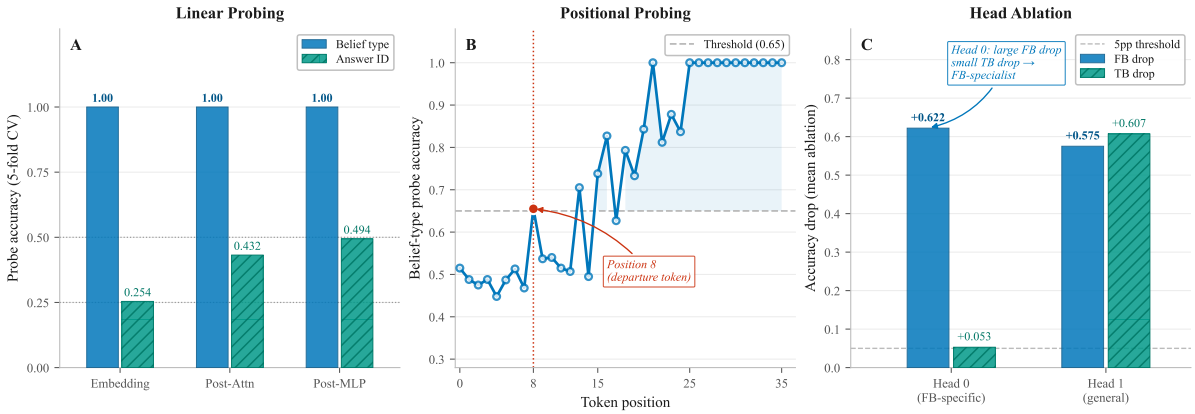

Figure 3: Mechanistic analysis of a trained 1L Transformer. **(A)** Linear probing shows answer identity is progressively computed: embedding=0.254, post-attention=0.432, post-MLP=0.494 (chance=0.250). **(B)** Positional probing: belief information first exceeds the reliability threshold (0.65) at position 8, the departure token. **(C)** Head ablation: Head 0 causes a 62.2% FB drop but only 5.3% TB drop (12× specificity ratio), identifying it as a false-belief specialist.

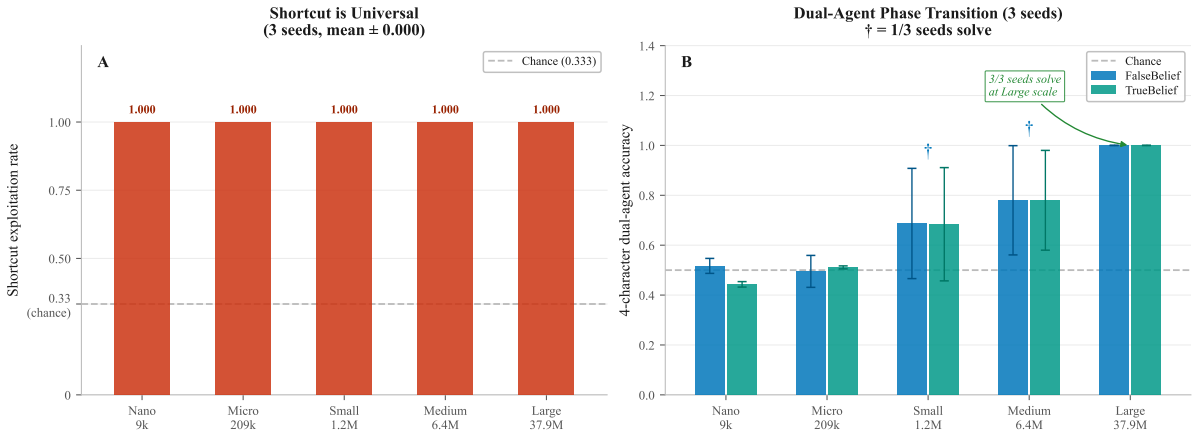

Figure 4: Scaling experiments. **Left:** Shortcut exploitation is universal—the shortcut rate is 1.000 at every scale from 8K to 38M parameters. **Right:** Four-character stories reveal a phase transition: models below 1.2M parameters cannot solve the extended task.

confirms that larger models do not "grow out of" shortcut exploitation; they simply exploit it more efficiently.

Shortcut-free stories reveal a qualitatively different scaling pattern. Extending to four-character stories—which require tracking beliefs across two actor pairs rather than one—produces a phase transition (Figure 4, right). Models below 1.2M parameters fail entirely (FB≈0.50, chance level), suggesting that the representational capacity needed to maintain multiple simultaneous belief states exceeds what small networks can provide. At 38M parameters, 3/3 seeds converge to perfect accuracy, placing the critical capacity threshold between 1.2M and 38M parameters for this task.

5 LLM Evaluation

We evaluate four open-weight LLMs—LLaMA 3.1 8B (Grattafiori et al., 2024), Gemma 4 E4B (Gemma Team, 2026), Qwen3 8B (Yang et al., 2025), and DeepSeek-R1 7B (Guo et al., 2025)—across eight conditions: naive/fixed stories × direct/chain-of-thought (Wei et al., 2022) prompting (Table 3, Panel A). Models are selected to span two training paradigms: standard instruction-tuned LLMs (LLaMA, Gemma) and reasoning-trained models with explicit thinking capabilities (Qwen3, DeepSeek-R1). Each condition uses 100 stories per seed across 3 seeds (300 effective evaluations per model), with identical stories shared across models for fair comparison. Models are served locally via Ollama; responses are parsed for con-

Table 3: Multi-model LLM evaluation on ToMForge synthetic stories (Panel A) and real Theory of Mind benchmarks (Panel B). Each cell reports FB accuracy / TB accuracy (unknown%). [T]Unknown rate >10 %; interpret with caution. [E]Excluded: parse-failure >25 %. Results are mean over 3 seeds (n=100 stories each). Models are grouped by training paradigm: standard LLMs (rows 1–2) and reasoning-trained models (rows 3–4).

Panel A — ToMForge Synthetic Stories				
Model	**Naive Direct**	**Naive CoT**	**Fixed Direct**	**Fixed CoT**
LLaMA 3.1 8B	1.000 / .000 (0%)	.915 / .062 (0%)	.982 / .000 (0%)	.782 / .170 (0%)
Gemma 4 E4B	1.000 / .118 (0%)	.628 / .720 (0%)	.727 / .081 (0%)	.703 / .889 (0%)
Qwen3 8B	.071 / .954 (0%)	1.000 / .024 (0%)	.103 / .941 (0%)	.976 / .170 (0%)
DeepSeek-R1 7B	.175 / .693 (2%)	—[E] (31%)	.297 / .385 (5%)	—[E] (36%)

Panel B — Real Benchmarks (Cross-Benchmark Validation)				
Model	**ToMi Direct**	**ToMi CoT**	**Hi-ToM Direct**	**Hi-ToM CoT**
LLaMA 3.1 8B	.520 / .759 (15%)[T]	.621 / .848 (0%)	.230 / .365 (1%)	.097 / .156 (6%)
Gemma 4 E4B	.506 / .664 (13%)[T]	.589 / .901 (0%)	.166 / .204 (27%)[T]	.130 / .323 (5%)
Qwen3 8B	.506 / .744 (8%)	.673 / .734 (0%)	.236 / .336 (2%)	.055 / .349 (5%)
DeepSeek-R1 7B	.366 / .672 (25%)[T]	—[E] (24%)	—[E] (35%)	—[E] (32%)

tainer mentions and classified as correct, incorrect, or unknown (unparseable). In the *direct* condition, the model receives the story followed by the question with no additional instructions. In the *CoT* condition, the prompt appends "Let's think step by step" after the question, and the answer is extracted from the final sentence of the response. Conditions with unknown rates exceeding 25% are excluded; those exceeding 10% are flagged. The generator is verified separately at n=2,000.

The models split into two behavioral categories:

Shortcut exploiters (LLaMA, Gemma). LLaMA 3.1 8B achieves perfect FB accuracy on naive stories (1.000) with zero TB accuracy (0.000), the signature of pure positional exploitation. CoT partially disrupts this: FB drops to 0.782 on fixed stories while TB rises to 0.170, suggesting CoT can unlock partial belief tracking but does not eliminate the shortcut. Gemma 4 E4B shows a stronger CoT rescue effect: naive CoT yields FB=0.628, TB=0.720, indicating genuine belief reasoning emerges under chain-of-thought prompting.

Realism-biased models (Qwen3, DeepSeek-R1). Qwen3 8B exhibits the inverse pattern: direct mode produces FB=0.071, TB=0.954. Rather than predicting the first-mentioned container, Qwen3 defaults to the object's *actual current location*—the container the object was moved to, which is always the correct true-belief answer. We call this a realism bias: the model reports reality rather than tracking the actor's outdated belief. This is not a positional shortcut but a different failure mode—the model answers where the object *is*, not where the actor *thinks* it is.

CoT *induces* shortcut exploitation in Qwen3: FB jumps to 1.000 while TB collapses to 0.024. The step-by-step reasoning chain appears to activate a positional heuristic that is suppressed in direct mode, producing a qualitatively worse strategy despite the added computation. DeepSeek-R1 7B confirms the realism bias in direct mode (FB=0.175, TB=0.693) but CoT results are excluded due to parse-failure rates exceeding 31%.

Cross-benchmark validation (Panel B). On real ToMi and Hi-ToM stories, all models show reduced accuracy relative to synthetic stories, and the behavioral profiles are partially preserved: LLaMA and Gemma maintain higher FB than TB on ToMi direct, consistent with residual shortcut exploitation on the biased benchmark. Hi-ToM proves harder for all models, with both FB and TB accuracy dropping substantially. DeepSeek-R1 is largely excluded from Panel B due to high parse-failure rates across conditions, a limitation of evaluating reasoning-trained models on structured QA formats. The cross-benchmark results confirm that the patterns observed on ToMForge synthetic stories are not artifacts of the controlled vocabulary: the same model-level behavioral profiles emerge on real benchmark stories, though attenuated by the greater complexity and variability of those benchmarks.

6 Architecture Gradient

The developmental ordering observed in LSTMs raises a question: is sequential bias a property of recurrence per se, or of specific architectural choices? LSTMs process input strictly left-to-right, maintaining a hidden state that is updated at each token. Information from early positions must survive through this bottleneck to influence later decisions, creating an inherent primacy bias. Transformers, by contrast, attend to all positions simultaneously via self-attention, imposing no sequential ordering on information access.

Mamba architectures (Gu and Dao, 2023; Dao and Gu, 2024) occupy an intermediate position: they use selective state-space models that process input sequentially (like LSTMs) but with structured state transitions that can in principle maintain long-range dependencies more efficiently. Mamba-2 goes further, achieving theoretical equivalence to linear attention through its structured state-space duality (SSD) (Dao and Gu, 2024). If the developmental ordering is caused by sequential processing bias, Mamba-1 should show a weaker effect than LSTM, and Mamba-2 should approach transformer-like behavior.

We train seven architectures—LSTM 2L, Mamba-1 Small, Mamba-1 Tiny, Mamba-2 Small, Mamba-2 Tiny, Transformer 1L, and Transformer 2L—with 10 seeds each on ToMForge stories (Figure 5). The convergence gap forms a monotonic gradient: LSTM (+110 steps, $p = 0.001$, $d = 1.49$) $>$ Mamba-1 Small (+40, $p = 0.018$, $d = 0.77$) $>$ Mamba-1 Tiny (+50, $p = 0.048$, $d = 0.59$) $>$ Mamba-2 Small (+40, $p = 0.084$, $d = 0.47$, n.s.) $>$ Mamba-2 Tiny (-10, n.s.) $\approx$ Transformer (-100 to -180, n.s.).

Mamba-2's structured state-space duality (Dao and Gu, 2024) reduces the sequential bias to attention-like levels, consistent with its theoretical equivalence to linear attention. The gradient is meaningful because it connects an architectural property (degree of sequential inductive bias) to a behavioral phenotype (developmental ordering) through a continuous, measurable progression.

The transition from significant to non-significant effects is informative. Mamba-1 Small and Tiny both show significant gaps ($p < 0.05$), placing them closer to LSTM on the continuum. Mamba-2 Small shows a marginal effect ($p = 0.084$, $d = 0.47$) that falls just below significance, while Mamba-2 Tiny shows no effect at all ($d = -0.32$)—effectively indistinguishable from the transformer baseline.

7 Discussion

Benchmark integrity. The positional shortcut in ToMi and Hi-ToM is not a minor calibration issue. At rates of 0.412 and 0.511 against a 0.333 baseline, a position heuristic provides a substantial accuracy boost that is invisible in aggregate metrics. Claims of ToM ability in LLMs (Kosinski, 2024; Strachan et al., 2024) may require qualification: if the evaluation benchmark itself is exploitable, above-chance performance does not demonstrate belief reasoning. The fact that this shortcut persists in benchmarks published in 2019 and 2023 suggests that positional balancing should be a standard requirement for any new ToM evaluation.

Architecture and inductive bias. The architecture gradient provides mechanistic insight into why different models behave differently on ToM tasks. Sequential architectures process input left-to-right, creating an inductive bias toward early-position information—exactly the bias that the positional shortcut rewards. Attention-based architectures access all positions simultaneously and therefore derive no systematic advantage from position. Mamba-2's structured state-space duality (Dao and Gu, 2024), which achieves theoretical equivalence to linear attention, explains why it occupies the same region as transformers on the gradient. The fact that LSTM's convergence gap ($d = 1.03$) mirrors the developmental ordering in human children (Wimmer and Perner, 1983) is suggestive but should not be overinterpreted: the mechanism is different (recurrent processing vs. conceptual development), even if the behavioral phenotype is superficially similar.

Chain-of-thought is not universally beneficial. The CoT results complicate the narrative around chain-of-thought as a universal reasoning enhancer (Wei et al., 2022). For Gemma 4 E4B, CoT unlocks genuine belief tracking (TB: 0.118→0.720 on naive stories, 0.081→0.889 on fixed stories). For Qwen3 8B, CoT *induces* the very shortcut it is supposed to help overcome (FB: 0.071→1.000). This divergence suggests that CoT's effectiveness on reasoning tasks cannot be assessed independently of the model's pre-existing biases. A model with a realism bias in direct mode may not benefit from step-by-step reasoning; instead, the explicit reason-

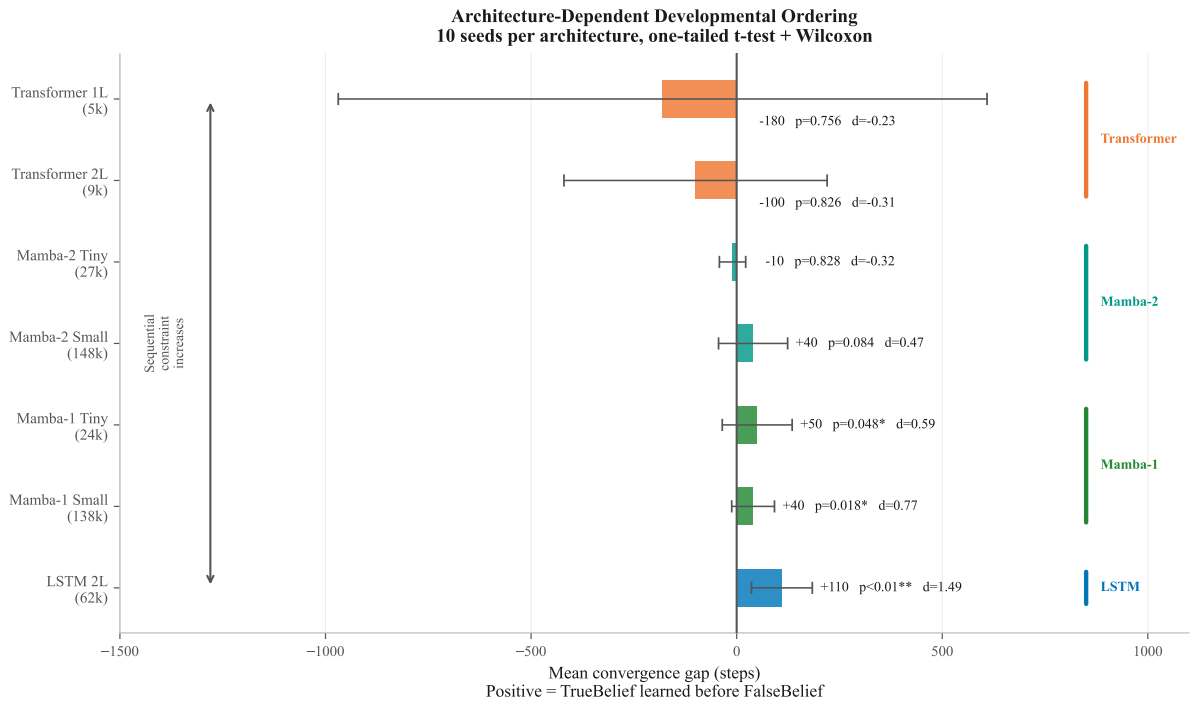

Figure 5: Architecture gradient across seven models (10 seeds each). The convergence gap—the number of training steps by which TB precedes FB at criterion—forms a monotonic gradient from LSTM (+110, $d = 1.49$) through Mamba-1 (+40–50) to Transformer (-100–180, n.s.).

ing chain may activate a different (and potentially worse) heuristic.

Implications for evaluation methodology. Our results point to methodological principles for ToM benchmark design. Shortcut verification should be standard: any generated evaluation should be tested for positional, lexical, and statistical artifacts before deployment. Architecture matters: a single aggregate accuracy number obscures architecture-dependent strategies that only become visible through fine-grained analysis (FB/TB breakdown, convergence dynamics, probing). And prompting regime matters: the same model can appear to solve or fail ToM depending on whether CoT is used, making prompting strategy a confound that must be reported and controlled for in evaluation.

8 Conclusion

Positional shortcuts in ToMi and Hi-ToM mean that above-chance performance on these benchmarks does not imply belief reasoning. ToMForge provides a verified alternative, and our experiments reveal that architecture determines how models engage with ToM tasks: sequential models exhibit developmental ordering, attention-based models do not, and LLMs show regime-dependent shortcut exploitation that chain-of-thought prompting can either remedy or exacerbate depending on the model. Code, data generation, and results are publicly available.[1]

9 Limitations

All small-model experiments were run on Apple M-series MPS, which exhibits floating-point non-determinism across runs. Exact numerical values may vary ± 0.05; all qualitative findings are invariant across multiple verification runs. Deterministic experiments (shortcut rates, benchmark analyses) produce exact-match results.

Each LLM is evaluated on 100 stories per condition across 3 seeds (300 effective per model), consistent with sample sizes in comparable ToM evaluations (BigToM: 100 stories, FANToM: $\sim$200 items). Identical stories are shared across models for fair comparison.

Exp12 covers four open-weight models at 7–8B scale. Larger models (70B+) and proprietary APIs were not evaluated. The architecture-dependent effects observed at 8B may differ at larger scales.

CoT-induced shortcut exploitation is confirmed for Qwen3 8B only. DeepSeek-R1 7B showed sim-

[1] Repository anonymized for review.

ilar direct-mode realism bias but CoT comparison was precluded by parse failures (>31%). Generalization to other reasoning-trained models requires further investigation.

The verification suite identifies a structural recency shortcut in second-order stories (rate=0.667, $p < 0.001$). Our first-order fix does not extend to second-order. Designing analogous structural controls for higher-order belief tracking is future work.

All experiments were designed, implemented, and verified by a single researcher. Canonical run logs and verification scripts enable full reproduction; independent replication on different hardware would strengthen confidence.

10 Ethics Statement

This work uses synthetic stories and publicly available benchmarks (ToMi, Hi-ToM). No human subjects were involved. LLM evaluations use open-weight models run locally; no user data is processed. We used an AI assistant (Claude, Anthropic) for coding assistance and iterative development.

References

Zhuang Chen, Jincenzi Gao, Minqi Li, Xiang Chen, Xipeng Qiu, and Xuanjing Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, and 1 others. 2021. Competency problems: On finding and removing artifacts in language data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gemma Team. 2026. Gemma 4. https://huggingface.co/google/gemma-4-E4B. Model card. Google DeepMind.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Daya Guo, Dejian Yang, He Zhang, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*.

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

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Halber, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference*

*on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics.

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

Melanie Sclar, Hyunwoo Kim, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2025. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. In *International Conference on Learning Representations (ICLR)*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

James W.A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Shlomi Guber, Jurgis Manzi, Guglielmo Orru, Alessandro Cermelli, Andrea Galeazzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural NLP. *arXiv preprint arXiv:2004.12265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. 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 (ACL)*. Association for Computational Linguistics.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yinghui Wu, Shuo Hu, Yixu Peng, and Hanyi Gong. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Additional Experiments

A.1 OOD Robustness (Exp03)

Unseen departure verbs ("sleeps", "exits") do not disrupt the learned representation (FB≈1.0), but injecting noise sentences destroys accuracy (FB drops to 0.270, below chance), confirming reliance on structural position rather than specific lexical cues.

A.2 Cultural Bias (Exp05)

No significant cultural bias was detected when evaluating on Indian names not seen during training (mean FB drop: +0.034±0.024, $p = 0.065$, n.s.).

A.3 Four-Character Extension (Exp06)

Extending to four characters reveals a phase transition: models below 1.2M parameters cannot solve the task (Nano 8.7K: FB=0.501; Micro 209K: FB=0.517), while 2/3 seeds converge at 6.4M parameters (Medium: FB=0.823).

A.4 Second-Order ToM (Exp07)

Zero-shot second-order reasoning fails entirely (Q2=0.101, below chance). Models trained on mixed first- and second-order stories achieve perfect Q2 accuracy (1.000), demonstrating that recursive belief tracking is learnable given appropriate training signal. Variant T (true-belief control) confirms this is not a formatting artifact (zero-shot Q2=0.070, mixed=1.000).

A.5 Statistical Rigour (Exp04)

Ten-seed replication confirms the developmental ordering for LSTM (gap=+90, SD=88, $p = 0.005$, d=1.03) and its absence for Transformer 1L (-180, $p = 0.756$) and 2L (-100, $p = 0.826$); both parametric and non-parametric tests agree.