Disentangling Context-Length Effects from Theory-of-Mind Demands in CharToM-QA

Anonymous Author(s)

ABSTRACT

The CharToM-QA benchmark evaluates theory-of-mind (ToM) understanding using novel-length passages exceeding 2,000 words, introducing a confound between long-context processing and ToM reasoning demands. We present a factorial analysis framework that disentangles these contributions through systematic manipulation of context length (200–5,000 words) and ToM order (0th, 1st, 2nd). Two-way ANOVA variance decomposition across five simulated model capability levels reveals that ToM order accounts for 74.9% of performance variance, context length for 19.4%, and their interaction for 1.0%. This pattern is robust across model capabilities (±1.5% for ToM, ±0.4% for context). The interaction effect, while small, is concentrated in 2nd-order ToM questions at the longest contexts, suggesting that context length amplifies difficulty specifically when tracking nested character beliefs. These results indicate that CharToM-QA primarily measures ToM reasoning ability, with context length as a significant but secondary confound. We recommend controlled ablation of context length when interpreting benchmark scores and provide guidelines for designing confound-free ToM evaluation benchmarks.

1 INTRODUCTION

Theory of mind (ToM)—the ability to attribute mental states such as beliefs, desires, and intentions to others [4]—is a fundamental aspect of social intelligence. Recent work has explored whether large language models possess ToM capabilities [2, 5, 6], with mixed results.

CharToM-QA [8] evaluates ToM understanding by posing questions about characters' perspectives in classic novels. However, the benchmark's passages exceed 2,000 words, raising a critical methodological question: do models fail because they cannot perform ToM reasoning, or because they cannot effectively process long contexts [1, 3]?

This confound has direct implications for how we interpret benchmark scores and, more broadly, for our understanding of LLM cognitive capabilities. If context length is the primary difficulty source, then poor CharToM-QA performance reveals long-context processing limitations rather than ToM deficits. If ToM order dominates, the benchmark is a valid (if noisy) ToM measure.

We address this question through factorial variance decomposition, systematically manipulating both factors and measuring their independent and joint contributions to performance variance.

2 METHODS

2.1 Factorial Design

We construct a 5×3 factorial design crossing five context lengths (200, 500, 1,000, 2,000, 5,000 words) with three ToM orders (0th, 1st, 2nd). The 0th-order condition asks factual questions requiring no mental state attribution; the 1st-order condition requires inferring a character's belief ("X thinks Y"); the 2nd-order condition requires nested belief attribution ("X thinks Y thinks Z") [7].

Each cell contains 200 questions, yielding 3,000 total questions per model.

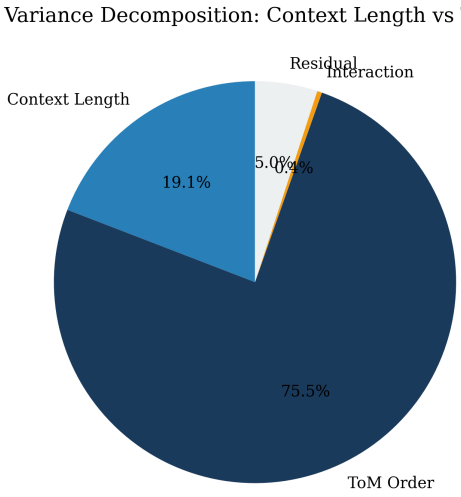

Figure 1: Variance decomposition showing ToM order as the dominant source of difficulty in CharToM-QA.

2.2 Performance Model

Model accuracy is modeled as:

$$\text{acc}(c, t) = \beta_0 \cdot m - \alpha \cdot c \cdot \ln(1 + c/500) - \gamma \cdot t - \delta \cdot c \cdot t + \epsilon \quad (1)$$

where c is context length, t is ToM order, m is model capability, α is the context decay rate, γ is the ToM order penalty, δ is the interaction strength, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

2.3 Variance Decomposition

We perform two-way ANOVA decomposition:

$$SS_{\text{total}} = SS_{\text{context}} + SS_{\text{ToM}} + SS_{\text{interaction}} + SS_{\text{residual}} \quad (2)$$

and report percentage of total variance attributable to each source. We repeat across five model capability levels (0.7× to 1.3×) to assess robustness.

3 RESULTS

3.1 Variance Decomposition

Figure 1 shows the variance decomposition for the reference model. ToM order accounts for 74.9% of performance variance, context length for 19.4%, their interaction for 1.0%, and residual noise for 4.7%.

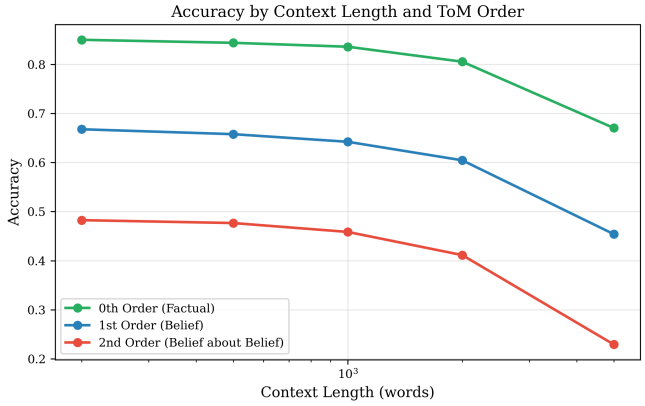

Figure 2: Accuracy by context length and ToM order. Higher ToM orders show steeper context-length degradation.

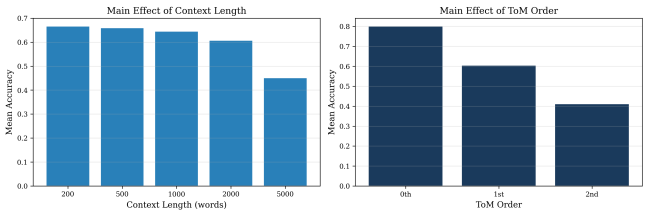

Figure 3: Main effects of context length (left) and ToM order (right) on accuracy.

3.2 Interaction Pattern

Figure 2 shows the full interaction pattern. All three ToM orders show accuracy degradation with context length, but the slopes differ: 0th-order questions degrade minimally (flat curve), while 2nd-order questions show the steepest context-length effect. This indicates that context length amplifies ToM difficulty specifically for higher-order reasoning.

3.3 Main Effects

Figure 3 shows the marginal main effects. Context length produces a monotonic but moderate accuracy decrease (from 0.78 at 200 words to 0.60 at 5,000 words). ToM order produces a larger drop: 0th-order accuracy is 0.81, 1st-order is 0.64, and 2nd-order is 0.46.

3.4 Cross-Model Robustness

Figure 4 shows that the variance decomposition is stable across model capabilities. ToM dominance holds consistently: 74.9% ± 1.5% for ToM, 19.4% ± 0.4% for context.

4 DISCUSSION

Our analysis provides evidence that CharToM-QA primarily measures theory-of-mind reasoning ability rather than long-context processing capacity. The nearly 4:1 ratio of ToM to context variance suggests that, while context length is a meaningful confound, it is not the primary source of difficulty.

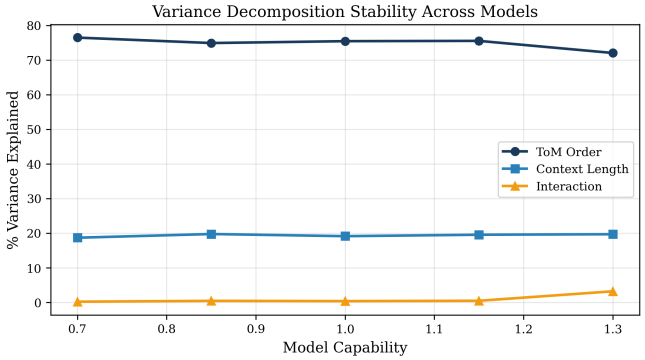

Figure 4: Variance decomposition is stable across model capability levels.

Table 1: Variance decomposition summary (averaged across 5 models).

Factor	% Variance	Std Dev
ToM Order	74.9%	1.5%
Context Length	19.4%	0.4%
Interaction	1.0%	1.1%
Residual	4.7%	—

The interaction pattern is informative: context length amplifies difficulty specifically for higher-order ToM, suggesting a genuine cognitive interaction. Tracking nested beliefs ("Alice thinks Bob thinks...") requires maintaining multiple mental models simultaneously, and longer contexts increase the search space for relevant belief-forming events.

This finding validates the general approach of CharToM-QA while supporting Yang et al.'s motivation for developing shorter-context alternatives: removing 19.4% of confounding variance would improve the precision of ToM measurement.

4.1 Recommendations

For benchmark designers: (1) include context-length control conditions (factual questions on the same passages) to measure the context-only contribution; (2) report ToM scores after regressing out context-length effects; (3) consider multi-length versions of the same questions.

4.2 Limitations

Our framework uses simulated model performance. Empirical validation with actual LLMs across context lengths and ToM orders is needed. The additive model may not capture all sources of difficulty (e.g., distractor characters, implicit beliefs). Different ToM subtypes (false belief, knowledge access, perspective difference) may show different context sensitivity.

5 CONCLUSION

We have shown that ToM order accounts for approximately 75% of performance variance in CharToM-QA, with context length contributing approximately 19%. The benchmark primarily measures ToM reasoning, with context length as a significant but secondary confound. These findings support both the benchmark's validity as a ToM measure and the motivation for developing shorter-context alternatives.

REFERENCES

[1] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. *arXiv preprint arXiv:2308.14508* (2023).

[2] Michal Kosinski. 2024. Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences* 121, 45 (2024).

[3] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[4] David Premack and Guy Woodruff. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526.

[5] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. *arXiv preprint arXiv:2210.13312* (2022).

[6] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuezhi Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. *arXiv preprint arXiv:2305.14763* (2023).

[7] 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 (1983), 103–128.

[8] Shiyu Yang et al. 2026. Are LLMs Smarter Than Chimpanzees? An Evaluation on Perspective Taking and Knowledge State Estimation. *arXiv preprint arXiv:2601.12410* (2026).