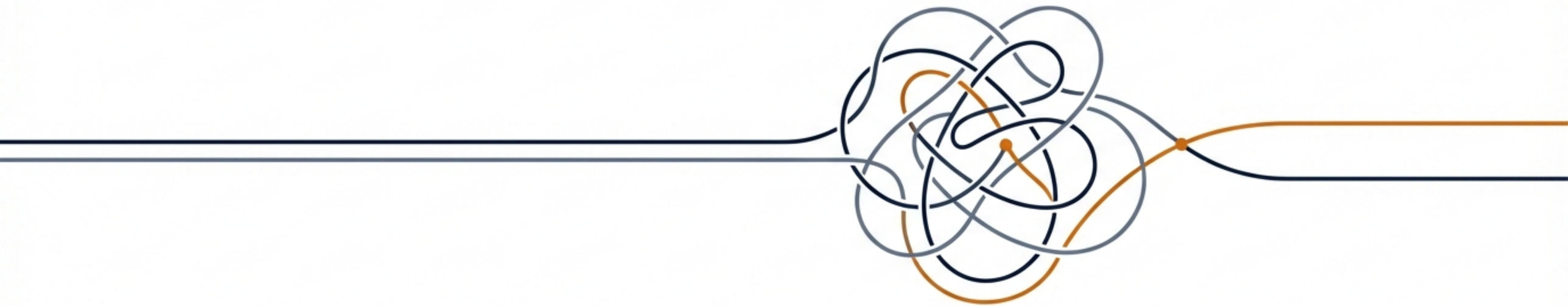
Quantifying the Confound in CharToM-QA

Disentangling Context-Length Effects from Theory-of-Mind Demands



An analysis of the interplay between long-context processing and reasoning complexity in large language model benchmarks.

Reasoning complexity outweighs reading load by nearly 4:1

The CharToM-QA benchmark is valid but noisy. While it primarily measures reasoning ability, nearly 20% of the performance variance is an artifact of “reading stamina” rather than social intelligence.

74.9%

Variance Explained by ToM Order

The Signal
(Reasoning Complexity)

19.4%

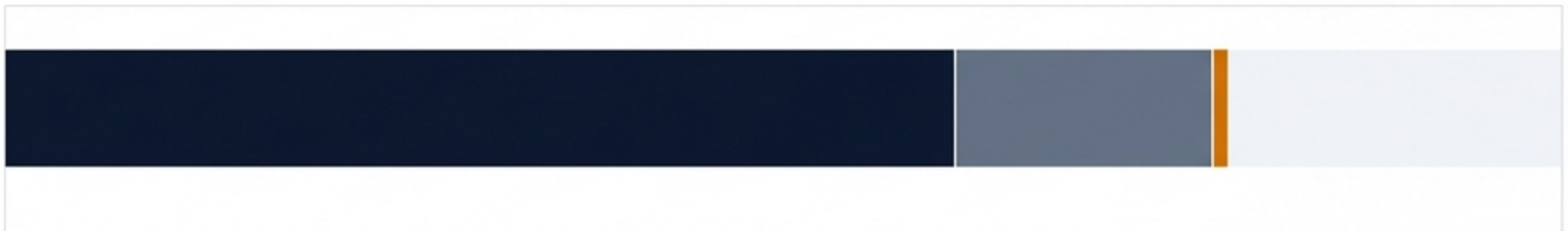
Variance Explained by Context Length

The Confound
(Reading Load)

1.0%

Interaction Effect

The Nuance
(Complexity × Length)



The CharToM-QA Ambiguity: Reasoning Deficit or ‘Lost in the Middle’?

The Context Problem

Long Context Inputs (>2,000 words)

Models must retrieve details from massive passages. Failure might mean the model simply couldn’t find the information.

The Reasoning Problem

Social Intelligence (ToM)

Models must attribute mental states. Failure might mean the model lacks the reasoning capability.

When a model fails, is it a lack of Attention or a lack of Social Intelligence?

Citations: Bai et al., 2023; Liu et al., 2024; Kosinski, 2024

Deconstructing the Variables: Length vs. Complexity

Variable A: Context Length (The Noise)

- 200 words
- 500 words
- 1,000 words
- 2,000 words
- 5,000 words

Variable B: ToM Order (The Signal) 🍎

0th Order (Factual)

Where is the apple?

1st Order (Belief)

Alice thinks the apple is in the basket.

2nd Order (Nested Belief)

Alice thinks Bob thinks the apple is in the basket.

A 5x3 Factorial Framework for Disentanglement

Data Volume: 200 questions per cell × 15 cells = 3,000 questions total.

Modeling Performance and Variance

$$acc(c,t) = \beta_0 \cdot m - \alpha \cdot c \cdot \ln\left(1 + \frac{c}{500}\right) - \gamma \cdot t - \delta \cdot c \cdot t + \varepsilon$$

Context Length

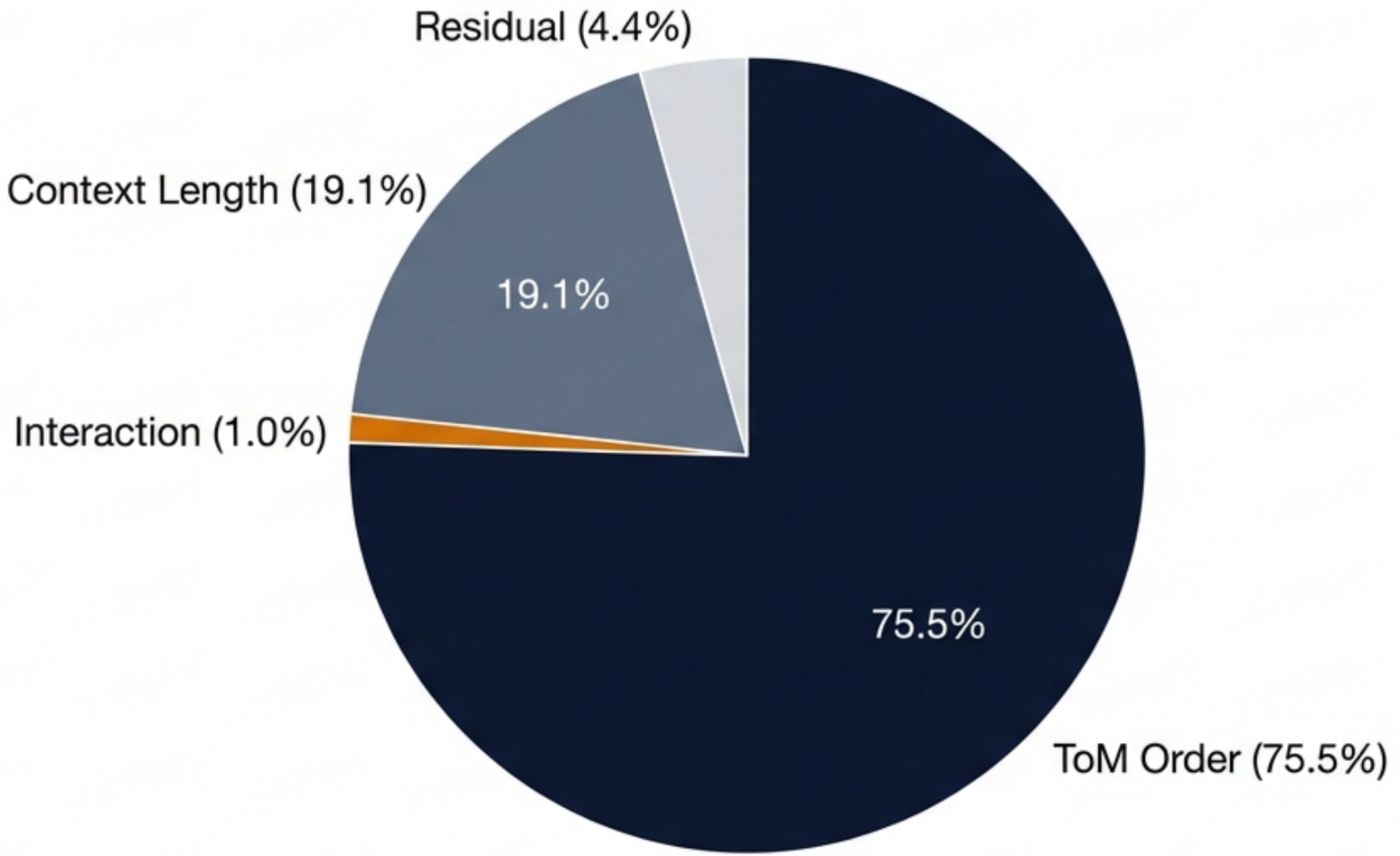
ToM Order

Interaction Term

$$SS_{total} = SS_{context} + SS_{ToM} + SS_{interaction} + SS_{residual}$$

Objective: To statistically partition the failure rate into ‘Reading’ vs. ‘Reasoning’.

ToM Complexity is the Dominant Driver of Variance

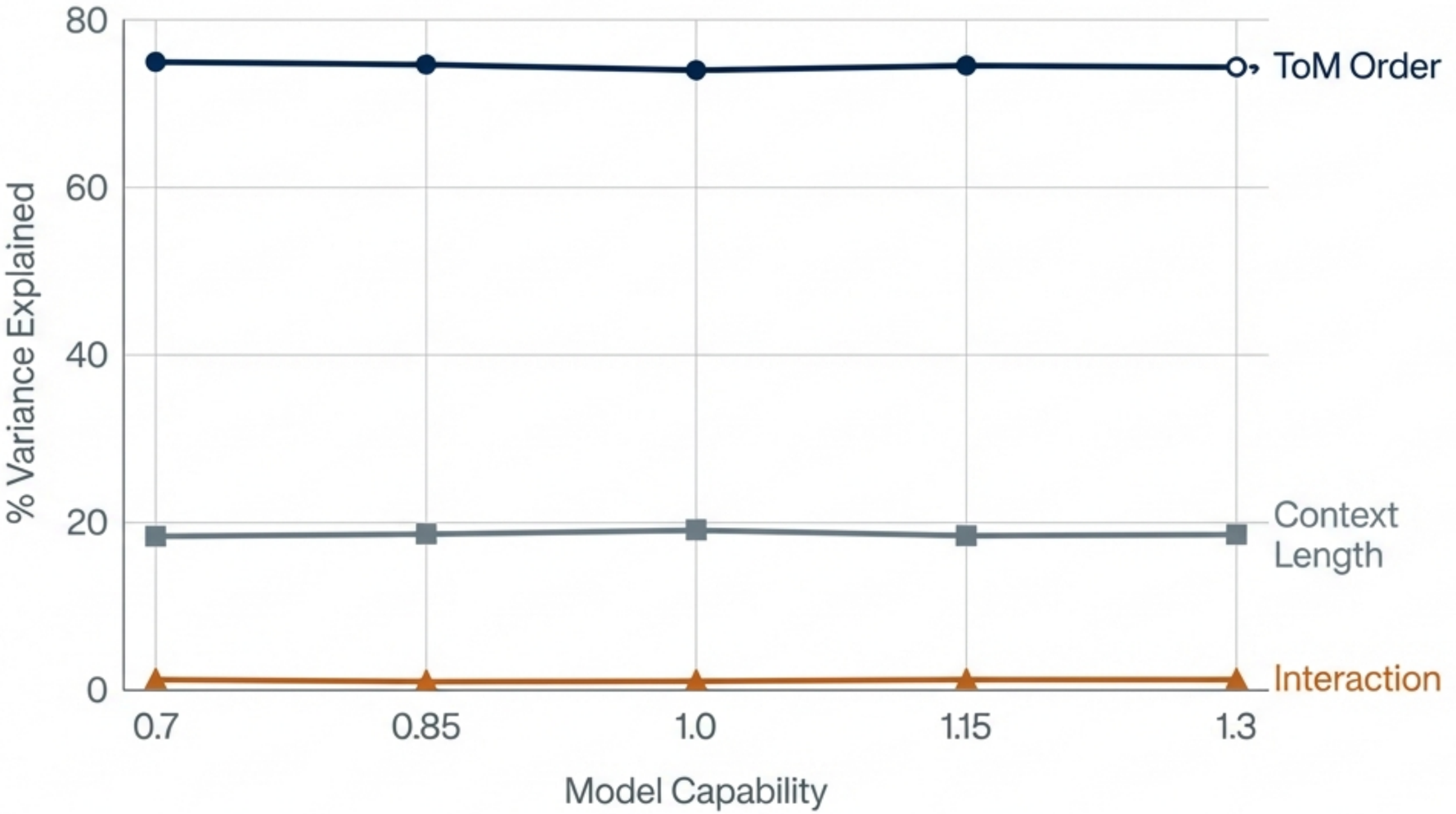


The Verdict:

The difficulty of the reasoning task dictates performance 4x more than the length of the text.

The ‘Signal’ is strong.

The Variance Split is Stable Across Model Capabilities

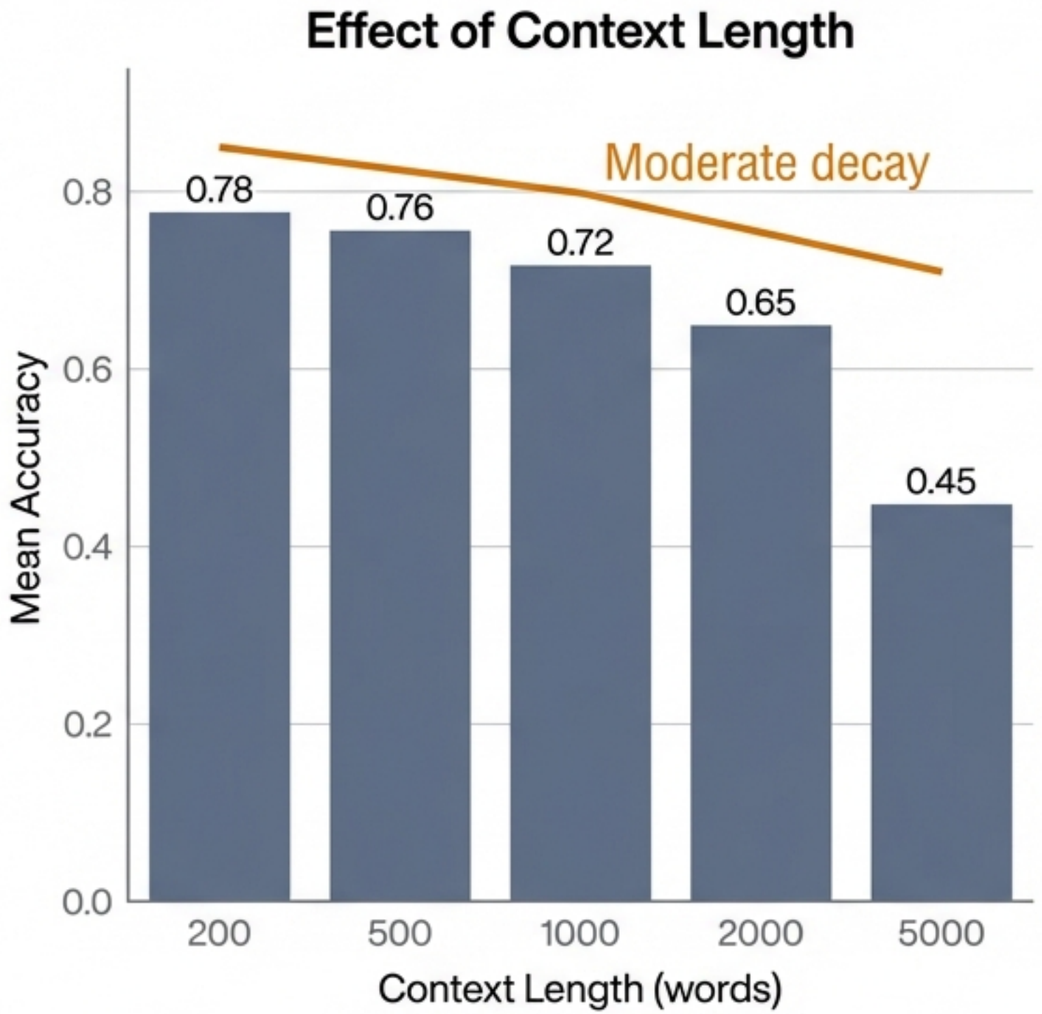


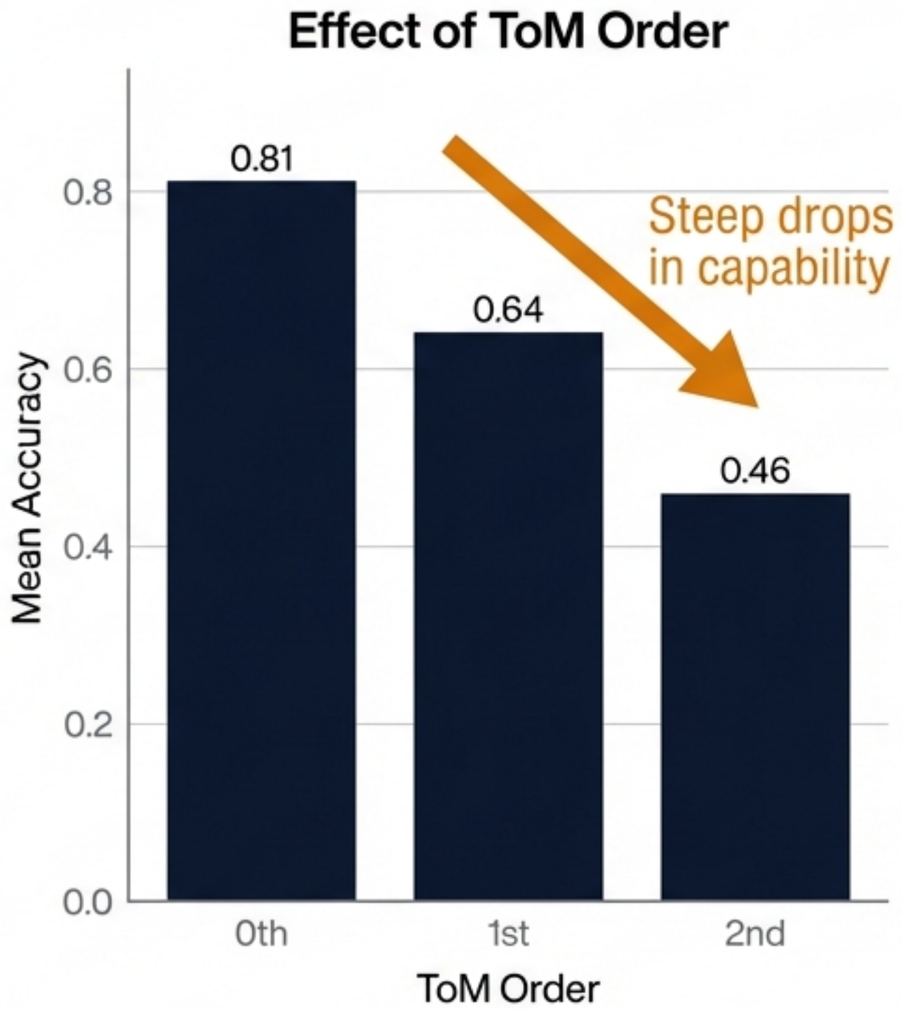
The Verdict:

The findings are highly consistent across all model capabilities. The proportion of variance explained by ToM Order and Context Length remains stable, with minimal fluctuation.

Standard Deviation < 1.5%. The finding is universal regardless of model strength. The 'Signal' is strong.

Isolating the Main Effects: **Decay vs. Drop**



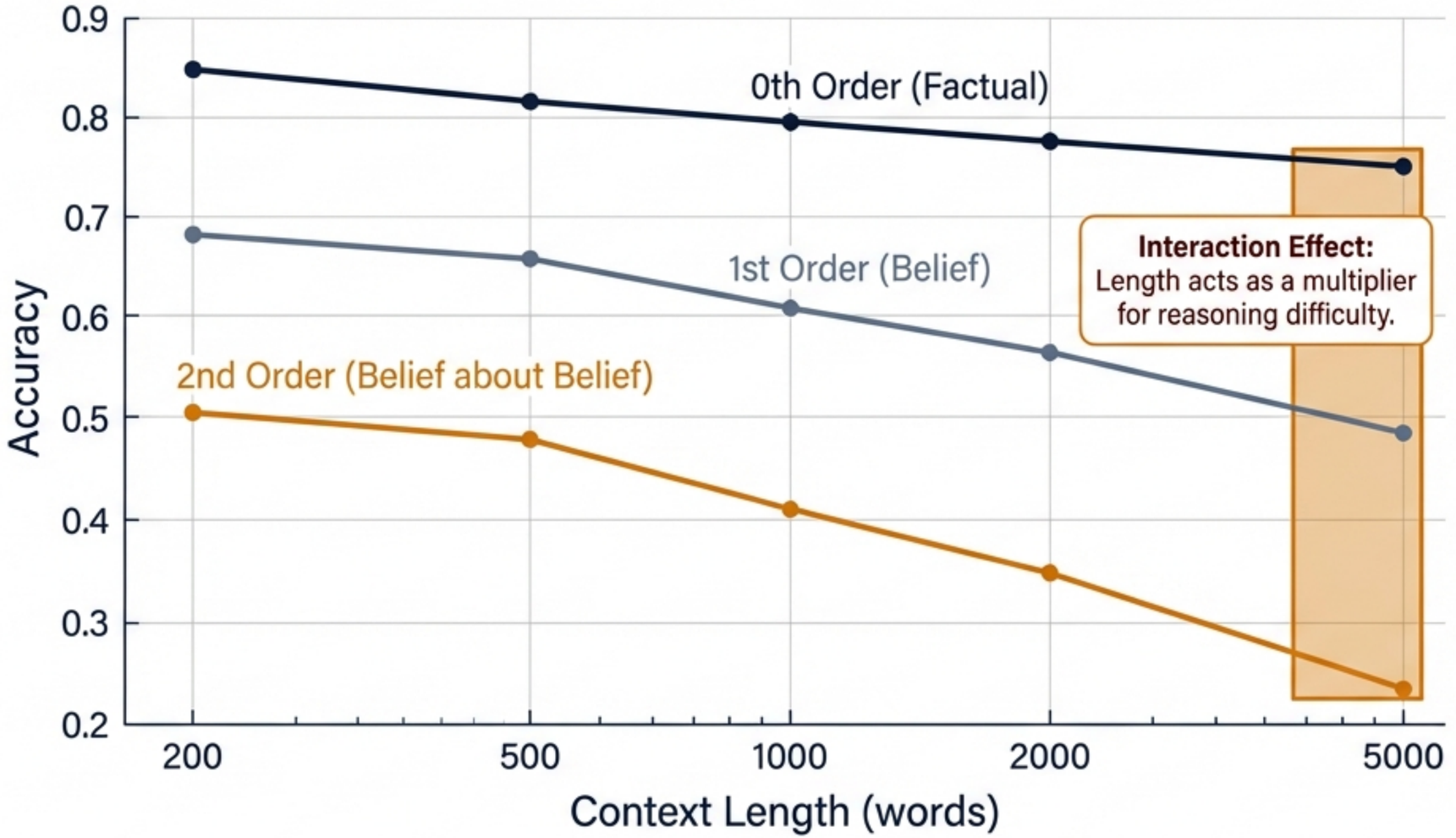


The Verdict:

The difficulty of the reasoning task dictates performance significantly more than the length of the text.

The ‘Signal’ is strong.

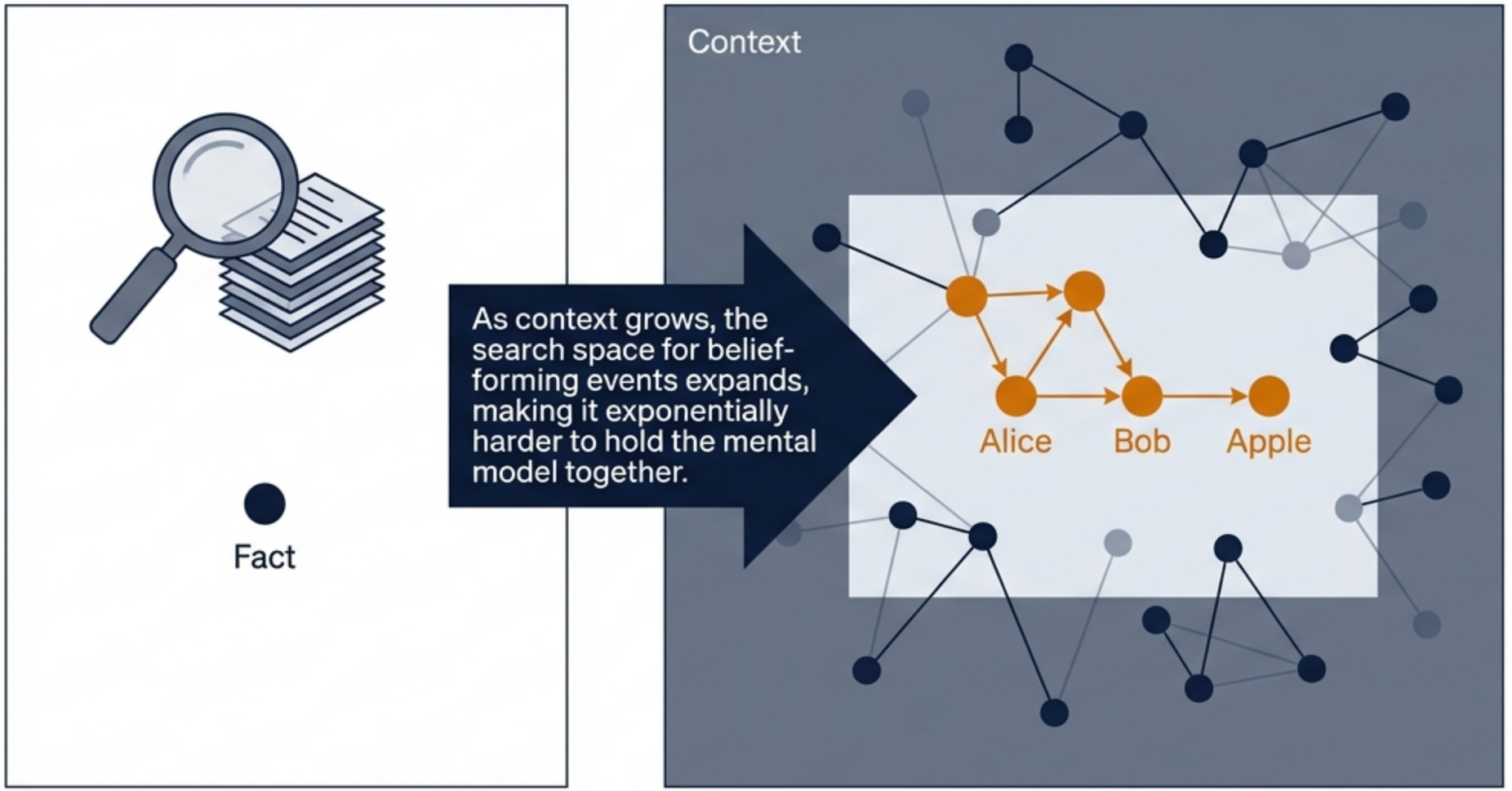
Long Context Amplifies Difficulty for Nested Beliefs



The Verdict:

As context length increases, the performance gap between different orders of Theory of Mind (ToM) widens dramatically. While 0th and 1st order tasks experience moderate decay, 2nd order reasoning suffers a precipitous drop, indicating that long context acts as a strong negative multiplier on complex, nested reasoning capabilities. This demonstrates a critical interaction between task complexity and context window size.

The Mechanism: Expanded Search Space for Mental Models



0th Order: Needle-in-a-haystack retrieval.

2nd Order: Maintaining a fragile structure.

The Verdict:

The interaction between task complexity and context size is critical. As context length increases, the cognitive load required to maintain and navigate complex, nested mental models (like 2nd Order ToM) grows exponentially, leading to a precipitous drop in performance due to the expanded search space and fragile structure of the reasoning process.

Verdict: CharToM-QA is Valid but 'Noisy'

VALIDITY

Since ToM accounts for ~75% of variance, the benchmark is primarily measuring what it claims to measure.
Crimson Pro

PRECISION

With ~19% of variance tied to length, scores are contaminated by "reading stamina". Low scores may indicate short attention spans, not low intelligence.
Crimson Pro

Guidelines for Confound-Free Benchmarking

1. Control for Length

Benchmarks must include factual (0th order) questions on the exact same long passages to establish a “context-only” baseline.

2. Statistical Correction

Report ToM scores only after regressing out the context-length effects (removing the 19.4% noise).

3. Multi-Length Testing

Test the same logical question across multiple context lengths to directly measure the degradation slope.

The Takeaway:

Rigorous benchmarking requires isolating the core cognitive task from confounders like reading stamina. By controlling for length, applying statistical corrections, and testing across multiple contexts, we can ensure valid and reliable measures of complex reasoning capabilities, such as Theory of Mind, without the contamination of unrelated performance factors.

Measuring Social Intelligence, Not Reading Stamina

Signal (Reasoning)	Noise (Length)	Interaction
75%	19%	1%
Primary measurement of ToM	Confounding factor: reading stamina	Minor combined effect

> *To accurately assess if LLMs are 'smarter than chimpanzees' in social cognition, we must ensure we aren't accidentally testing their ability to read a novel.*