Methods: Multi-Dimensional Analysis of Theory of Mind in Large Language Models

1 Methods

1.1 Dataset and Language Models

We evaluated six state-of-the-art large language models on a comprehensive Theory of Mind (ToM) dataset containing 2,860 samples across six ability categories: Emotion, Belief, Desire, Intention, Knowledge, and Non-Literal Communication. The evaluated models included:

- Meta Llama 3.1 70B Instruct
- Qwen 2.5 32B Instruct
- Allen Institute OLMo 2.0 13B Instruct
- Mistral 7B Instruct v0.3
- Microsoft Phi-3 Mini 4K Instruct
- InternLM 2.5 1.8B Chat

Model performance was evaluated using binary accuracy (correct/incorrect) on multiple-choice Theory of Mind questions spanning diverse cognitive abilities and complexity levels.

1.2 Multi-Dimensional Question Analysis

To understand the relationship between question characteristics and model performance, we conducted three complementary analyses that capture different aspects of linguistic and cognitive complexity.

1.2.1 Idea Density Analysis

Idea density measures the propositional content per unit of text, providing insight into the informational complexity of ToM questions. We employed the DEPID (Density Estimation of Propositional IDeas) algorithm, which:

1. Performs part-of-speech tagging and syntactic parsing

2. Identifies propositional ideas based on predicate-argument structures
3. Calculates density as the ratio of propositions to total words
4. Normalizes scores to account for text length variations

The idea density metric captures how much semantic information is packed into each question, with higher values indicating more cognitively demanding content that requires processing multiple interconnected concepts simultaneously.

1.2.2 Question Complexity Analysis

We developed a comprehensive question complexity framework that decomposes cognitive demands into four distinct dimensions:

Syntactic Complexity (C_{syn}): Measures structural linguistic complexity through:

- Parse tree depth and branching factor
- Clause embedding levels
- Syntactic dependency distances
- Presence of complex grammatical constructions

Semantic Complexity (C_{sem}): Quantifies meaning-level complexity via:

- Lexical diversity and semantic field breadth
- Abstract concept density
- Polysemy and ambiguity measures
- Conceptual relationship complexity

Theory of Mind Complexity (C_{ToM}): Assesses ToM-specific cognitive demands:

- Mental state attribution depth (first-order, second-order, etc.)
- Number of agents and their mental states
- Temporal reasoning about belief changes
- False belief and perspective-taking requirements

Reasoning Complexity (C_{reas}): Evaluates logical and inferential demands:

- Inference chain length and complexity
- Causal reasoning requirements
- Counterfactual thinking demands
- Integration of multiple information sources

The overall Question Complexity Score is computed as:

$$C_{total} = \alpha C_{syn} + \beta C_{sem} + \gamma C_{ToM} + \delta C_{reas} \tag{1}$$

where α, β, γ, and δ are empirically determined weights reflecting the relative contribution of each dimension.

1.2.3 Rhetorical Structure Theory (RST) Analysis

RST analysis captures the discourse-level organization of ToM questions by analyzing how textual segments relate to form coherent narratives. Our RST implementation identifies:

Elementary Discourse Units (EDUs): The minimal building blocks of discourse, typically clauses or clause-like segments that express single propositions or events.

Tree Depth: The hierarchical depth of the RST parse tree, indicating the complexity of discourse organization and the number of embedded rhetorical relations.

Rhetorical Relations: We focus on three key relation types particularly relevant to ToM reasoning:

- **Attribution Relations:** Capture instances where mental states, beliefs, or speech acts are attributed to agents (e.g., "John thinks that...", "Mary said that...")
- **Causal Relations:** Identify cause-effect relationships that are crucial for understanding motivations and consequences in ToM scenarios
- **Explanation Relations:** Mark segments that provide explanatory information about agents' mental states or behaviors

The RST analysis provides insight into how discourse structure affects the cognitive load of processing ToM questions, with more complex rhetorical structures potentially requiring greater working memory and integration capabilities.

1.3 Statistical Analysis

1.3.1 Correlation Analysis

We computed Pearson correlation coefficients between model performance and each analysis metric to identify significant relationships between question characteristics and model accuracy. Statistical significance was assessed using $p < 0.05$ with appropriate corrections for multiple comparisons.

1.3.2 Performance Ranking

Models were ranked by overall accuracy across all ToM questions, allowing us to examine whether the relationship between question characteristics and performance varies systematically with model capability.

1.3.3 Ability Group Analysis

We analyzed model performance separately for each of the six ToM ability categories to identify domain-specific patterns and determine whether certain types of mental state reasoning are more challenging than others.

1.4 Visualization and Interpretation

Our analysis employed several visualization techniques to reveal patterns in the high-dimensional relationship between question characteristics and model performance:

- **Circular Radar Charts:** Display model performance across ToM ability groups, allowing direct comparison of strengths and weaknesses
- **Correlation Matrices:** Show the relationship between model performance and analysis metrics, with significance highlighting
- **Scatter Plots:** Reveal how performance varies across different levels of question complexity
- **Submeasure Correlation Analysis:** Examine relationships between different analysis dimensions

This multi-dimensional approach provides a comprehensive characterization of the factors that influence Theory of Mind performance in large language models, enabling both theoretical insights into the nature of machine ToM reasoning and practical guidance for model development and evaluation.