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, each capturing different aspects of linguistic and cognitive processing requirements. Our analysis employs natural language processing techniques including dependency parsing, named entity recognition, and lexical analysis to quantify these complexity dimensions systematically.

Syntactic complexity (C_{syn}) measures the structural linguistic complexity of questions through dependency tree analysis and grammatical construction identification. The syntactic complexity score is computed as a weighted combination of three key metrics:

$$C_{syn} = 0.3 \cdot D_{max} + 0.3 \cdot T_{dep} + 0.4 \cdot N_{clause} \tag{1}$$

where D_{max} represents the maximum dependency tree depth, calculated as the longest path from the root to any leaf node in the syntactic parse tree; T_{dep} denotes the number of unique dependency relation types present in the question, indicating grammatical diversity; and N_{clause} counts the number of clausal structures, approximated by identifying subordinating conjunctions and relative pronouns that introduce embedded clauses.

Semantic complexity (C_{sem}) quantifies the meaning-level complexity through lexical and conceptual analysis. This dimension captures the cognitive demands associated with processing diverse vocabulary, abstract concepts, and semantic relationships:

$$C_{sem} = 0.4 \cdot TTR + 0.3 \cdot N_{ent} + 0.3 \cdot R_{content} \tag{2}$$

where TTR represents the type-token ratio (lexical diversity), computed as the number of unique lemmatized words divided by the total word count; N_{ent} is the count of named entities identified in the question, reflecting the presence of specific referents that require world knowledge; and $R_{content}$ denotes the content word ratio, calculated as the proportion of nouns, verbs, adjectives, and adverbs relative to the total word count.

Theory of Mind complexity (C_{ToM}) assesses the cognitive demands specific to mental state reasoning and perspective-taking. This dimension captures the complexity arising from multiple agents, mental state attributions, and perspective shifts:

$$C_{ToM} = 0.4 \cdot N_{mental} + 0.3 \cdot N_{perspective} + 0.3 \cdot N_{third} \tag{3}$$

where N_{mental} counts mental state verbs (think, believe, know, feel, want, etc.) that indicate explicit mental state attributions; $N_{perspective}$ identifies perspective markers such as

"he thinks," "her view," or "from his point of view" that signal perspective-taking requirements; and N_{third} represents third-person pronoun usage, which correlates with the presence of multiple agents whose mental states must be tracked.

Reasoning complexity (C_{reas}) evaluates the logical and inferential demands imposed by the question structure. This dimension captures the complexity arising from logical operations, temporal reasoning, and conditional structures:

$$C_{reas} = 0.3 \cdot N_{logical} + 0.3 \cdot N_{temporal} + 0.4 \cdot N_{conditional} \tag{4}$$

where $N_{logical}$ counts logical operators (and, or, but, because, etc.) that require logical reasoning; $N_{temporal}$ identifies temporal references (before, after, when, while, etc.) that introduce temporal reasoning demands; and $N_{conditional}$ captures conditional structures (if-then statements, counterfactuals) that require hypothetical reasoning.

The overall Question Complexity Score integrates these four dimensions through equal weighting, reflecting our assumption that each dimension contributes equally to overall cognitive demand:

$$C_{total} = \frac{1}{4}(C_{syn} + C_{sem} + C_{ToM} + C_{reas}) \tag{5}$$

This multi-dimensional approach enables fine-grained analysis of the specific cognitive demands imposed by different types of Theory of Mind questions, allowing us to identify which aspects of complexity most strongly influence model performance across different ability categories.

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. Let T denote the RST parse tree for a given text, where each node $n \in T$ has an associated rhetorical relation type relation(n). Our RST implementation extracts five key metrics that quantify different aspects of discourse complexity relevant to Theory of Mind reasoning.

The first metric, Elementary Discourse Units (EDUs), represents the minimal building blocks of discourse, typically clauses or clause-like segments that express single propositions or events. The number of EDUs in a text is formally defined as:

$$\text{num\_edus} = |\{n \in T : \text{relation}(n) = \text{“elementary”}\}| \tag{6}$$

This metric captures the granularity of discourse segmentation, with higher values indicating more complex propositional structures that require greater cognitive resources to process and integrate.

The second metric, tree depth, quantifies the hierarchical complexity of discourse organization by measuring the maximum depth of the RST parse tree:

$$\text{tree\_depth} = \max\{\text{depth}(n) : n \in \text{leaves}(T)\} \tag{7}$$

where depth(n) represents the number of edges from the root to node n, and leaves(T) denotes the set of leaf nodes in tree T. Greater tree depth indicates more embedded rhetorical

relations, suggesting increased demands on working memory and hierarchical processing capabilities.

We focus on three specific rhetorical relation types that are particularly relevant to Theory of Mind reasoning. Attribution relations capture instances where mental states, beliefs, or speech acts are attributed to agents, formally defined as:

$$\text{rel\_attribution} = |\{n \in T : \text{relation}(n) = \text{“attribution”}\}| \tag{8}$$

These relations are crucial for tracking perspective and mental state attributions in ToM scenarios, such as identifying when the text indicates "John thinks that..." or "Mary believes that...".

Causal relations identify cause-effect relationships that are essential for understanding motivations and consequences in Theory of Mind contexts:

$$\text{rel\_causal} = |\{n \in T : \text{relation}(n) = \text{“causal”}\}| \tag{9}$$

These relations help capture the logical chains of reasoning that connect mental states to actions and outcomes, which are fundamental to ToM understanding.

Explanation relations mark discourse segments that provide explanatory information about agents' mental states or behaviors:

$$\text{rel\_explanation} = |\{n \in T : \text{relation}(n) = \text{“explanation”}\}| \tag{10}$$

These relations are particularly important for identifying passages that elaborate on the reasoning behind characters' actions or mental states.

The general formula for counting any rhetorical relation type r excludes elementary relations, as these serve as discourse unit markers rather than actual rhetorical connections:

$$\text{rel\_}r = |\{n \in T : \text{relation}(n) = r \wedge \text{relation}(n) \neq \text{“elementary”}\}| \tag{11}$$

These RST metrics are constrained by the structural properties of discourse trees. Specifically, the number of EDUs must be at least one ($\text{num\_edus} \geq 1$), tree depth is non-negative ($\text{tree\_depth} \geq 0$), all relation counts are non-negative, and the total number of rhetorical relations cannot exceed the maximum possible internal nodes in a binary tree: $\sum_r \text{rel\_}r \leq \text{num\_edus} - 1$.

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. By quantifying these discourse-level features, we can examine how the organizational complexity of narratives influences large language models' ability to reason about mental states and social interactions.

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.