

# How Do Large Language Models Acquire Theory of Mind?

**Team: DZS**

Zuzanna Drużyńska

Sofia Ivanko

Dilyara Katasheva

Poznań, 2026

# Contents

<b>1 Topic and Task Description</b>	<b>3</b>
1.1 Task Definition . . . . .	3
<b>2 Dataset Description</b>	<b>3</b>
2.1 Source and Motivation . . . . .	3
2.2 Dataset Structure . . . . .	3
2.3 Question Types . . . . .	4
<b>3 Exploratory Data Analysis</b>	<b>4</b>
3.1 Sentiment Distribution . . . . .	5
3.2 Observed vs Unobserved Scenarios . . . . .	6
3.3 Answer Frequency Analysis . . . . .	6
3.4 Intentions Word Analysis . . . . .	7
<b>4 Experiment 1: Zero-Shot Baseline</b>	<b>8</b>
4.1 Experimental Setup . . . . .	8
4.2 Results . . . . .	8
4.3 Analysis . . . . .	8
<b>5 Experiment 2: Multi-Model Comparison</b>	<b>8</b>
5.1 Motivation . . . . .	8
5.2 Models Evaluated . . . . .	8
5.3 Scoring Method . . . . .	8
5.4 Results . . . . .	9
5.5 Interpretation . . . . .	9
<b>6 Conclusion</b>	<b>9</b>
<b>7 References</b>	<b>9</b>

# 1 Topic and Task Description

Theory of Mind (ToM) is the cognitive ability to reason about the beliefs, intentions, emotions, and perspectives of other agents. It allows humans to understand that others may have knowledge different from their own, including incorrect or false beliefs. This ability is essential for social interaction, cooperation, and communication.

With the rapid development of Large Language Models (LLMs), an important research question has emerged: whether models trained purely on large-scale text data can acquire ToM-like reasoning abilities without explicit symbolic modeling or grounding.

Recent work by Kosinski (2024) suggests that LLMs may demonstrate ToM-like behavior under specific experimental conditions. However, it remains unclear whether these behaviors reflect genuine reasoning or are the result of surface-level statistical correlations learned from text.

## 1.1 Task Definition

The task of this project is to evaluate whether LLMs can correctly reason about mental states in structured social scenarios. We focus on sentiment-based attitude questions from the OpenToM dataset and formulate the problem as a three-class classification task:

- negative
- neutral
- positive

The models are evaluated in a zero-shot setting using Natural Language Inference (NLI), without any task-specific fine-tuning.

# 2 Dataset Description

## 2.1 Source and Motivation

We use the OpenToM dataset, a large-scale benchmark designed specifically to evaluate Theory of Mind reasoning in language models. The dataset is publicly available at:

<https://github.com/seacowx/OpenToM>

OpenToM was selected because it explicitly targets mental-state reasoning rather than surface-level linguistic patterns.

## 2.2 Dataset Structure

The dataset contains 13,708 samples. Each sample consists of a narrative scenario describing interactions between agents, followed by a structured question.

Each entry includes:

- **plot**: a short description of the situation
- **narrative**: an extended story context
- **intention**: inferred intention of an agent
- **personality**: personality traits
- **true\_sentiment**: ground truth label

- **observed**: whether the action was witnessed
- **preferences**: likes and dislikes of agents
- **question**: question text, type, and correct answer

### 2.3 Question Types

Although OpenToM includes multiple question types, this project focuses on *attitude* questions, which require the model to infer emotional or evaluative states of agents based on the scenario.

## 3 Exploratory Data Analysis

Before running experiments, we performed exploratory data analysis to understand the structure and limitations of the dataset.

### 3.1 Sentiment Distribution

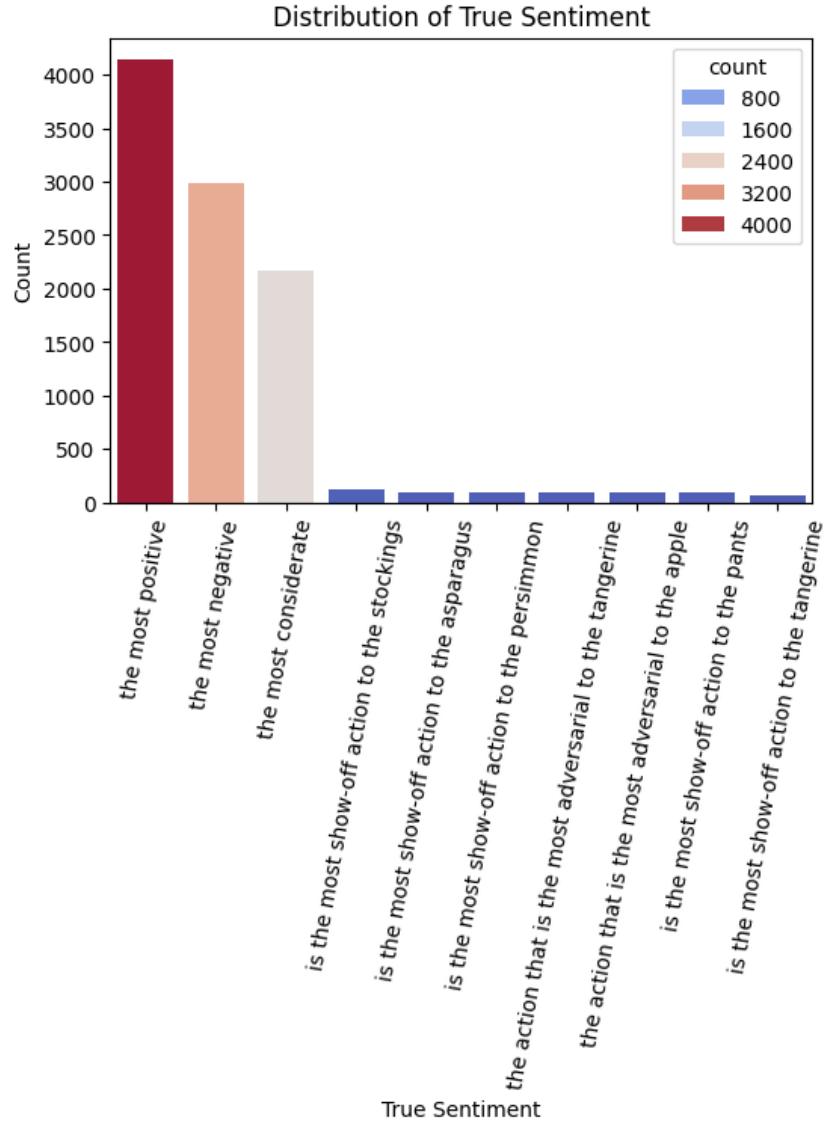


Figure 1: Distribution of true sentiment labels

The dataset is highly imbalanced. Extreme sentiment categories occur significantly more often than neutral ones. This imbalance makes classification more difficult and increases the likelihood of biased predictions.

### 3.2 Observed vs Unobserved Scenarios

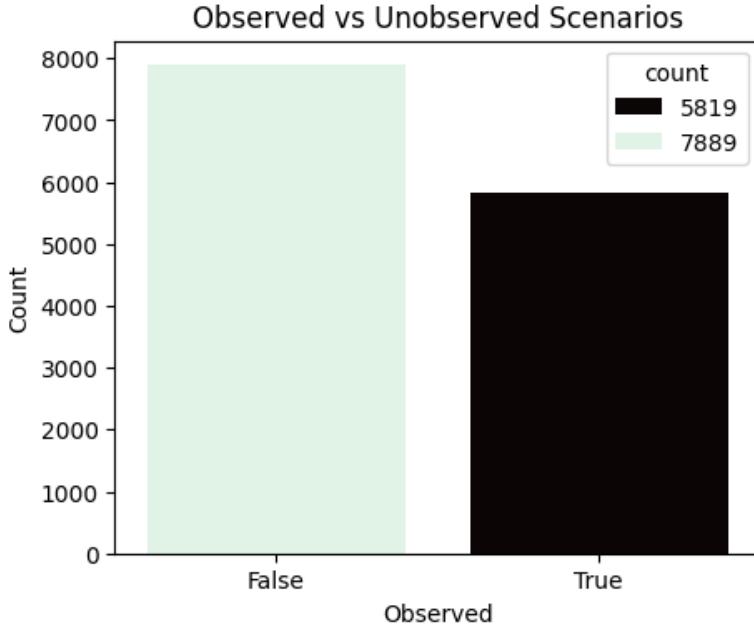


Figure 2: Observed and unobserved scenarios

This graph shows how different sentiment labels (like 'positive', 'negative', or 'neutral') are distributed across the dataset. Most entries are strongly negative, suggesting the dataset contains many situations involving negative emotions or evaluations.. This confirms that OpenToM is not a trivial dataset and explicitly targets core ToM abilities.

### 3.3 Answer Frequency Analysis

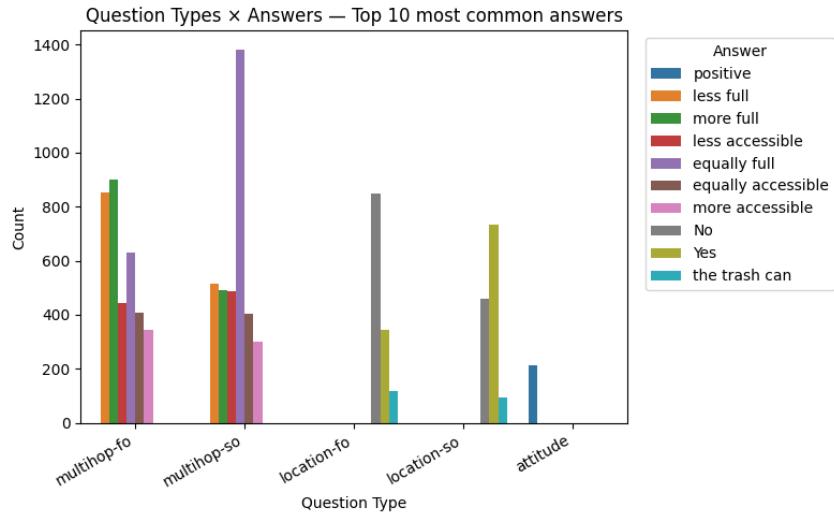


Figure 3: Most common answers

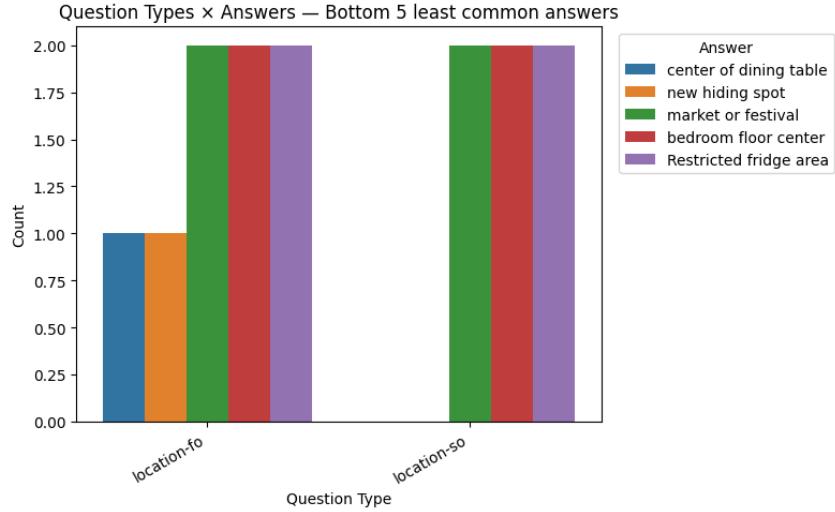


Figure 4: Least common answers

The answer distribution is skewed, with a small set of labels dominating the dataset. Rare answers often correspond to highly specific spatial or contextual reasoning. It highlights patterns in how the model or participants answered standard ToM questions.

### 3.4 Intentions Word Analysis

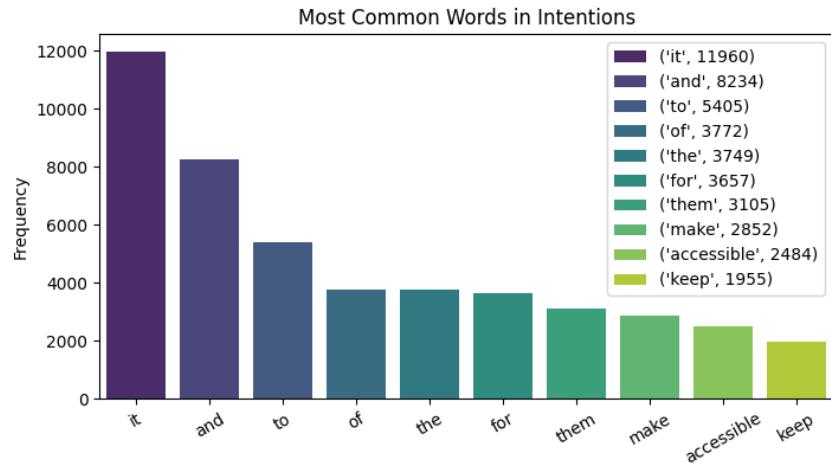


Figure 5: Most frequent words in intention descriptions

This bar chart highlights the most frequently used words in intention descriptions. Common words like 'it', 'make', and 'accessible' suggest the dataset often describes purposeful, goal-driven actions. Intentions are often described using goal-oriented language, indicating that reasoning about purpose and outcome is central to the dataset.

## 4 Experiment 1: Zero-Shot Baseline

### 4.1 Experimental Setup

The first experiment evaluates a strong NLI baseline using the pretrained model `roberta-large-mnli`. This model was selected because it is commonly used for zero-shot classification tasks and performs well on natural language inference benchmarks.

Due to computational constraints, we randomly sampled 500 instances from the dataset.

```
MODEL_NAME = "roberta-large-mnli"
batch_size = 4
max_length = 512
```

Each instance was converted into a premise–hypothesis pair. The premise contained the narrative and question, while the hypothesis followed the template:

```
"The attitude is {}."
```

### 4.2 Results

Accuracy: 0.3260

Macro F1-score: 0.2408

### 4.3 Analysis

The performance is close to random guessing. This suggests that even a large NLI model struggles to infer emotional attitudes from ToM-heavy scenarios without task-specific adaptation.

## 5 Experiment 2: Multi-Model Comparison

### 5.1 Motivation

Since the baseline model performed poorly, we evaluated several alternative NLI models to determine whether architecture choice significantly affects ToM performance.

### 5.2 Models Evaluated

- microsoft/deberta-large-mnli
- facebook/bart-large-mnli
- roberta-large-mnli
- MoritzLaurer/DeBERTa-v3-large-mnli

### 5.3 Scoring Method

We modified the evaluation logic by computing the difference between entailment and contradiction logits:

```
score = entailment_logit - contradiction_logit
```

## 5.4 Results

Model	Accuracy	Macro F1
DeBERTa-large-MNLI	0.4220	0.3421
BART-large-MNLI	0.3800	0.3218
RoBERTa-large-MNLI	0.3080	0.2061
DeBERTa-v3	0.3000	0.2899

## 5.5 Interpretation

DeBERTa-large-MNLI achieved the highest accuracy, but overall performance remains far from reliable. This indicates that zero-shot NLI models are not well suited for complex ToM reasoning.

## 6 Conclusion

This project examined whether Large Language Models demonstrate Theory of Mind abilities using the OpenToM dataset. Through detailed data analysis and two experimental setups, we showed that:

- The dataset requires genuine mental-state reasoning.
- Zero-shot NLI models perform poorly on ToM tasks.
- Even the best-performing model achieves only moderate accuracy.

These results suggest that Theory of Mind does not reliably emerge from standard language modeling alone and likely requires explicit modeling or task-specific training.

## 7 References

Kosinski, M. (2024). "Testing theory of mind in large language models and human". Article: [https://www.nature.com/articles/s41562-024-01882-z?utm\\_source](https://www.nature.com/articles/s41562-024-01882-z?utm_source)  
OpenToM Dataset: <https://github.com/seacowx/OpenToM>  
Project Repository: [https://github.com/Sophy333/Sem\\_Web\\_DZS](https://github.com/Sophy333/Sem_Web_DZS)