Automatic Scoring of an Open-Response Measure of Advanced Mind-Reading Using Large Language Models

Yixiao Wang[1], Russel Dsouza[1], Robert Lee[2], Ian Apperly[2], Rory T. Devine[2], Sanne W. van der Kleij[2], Mark Lee[1]
[1]School of Computer Science, University of Birmingham, UK
[2]School of Psychology, University of Birmingham, UK
{y.wang.37, r.s.dsouza, r.lee.5, i.a.apperly, r.t.devine, s.w.vanderkleij, m.g.lee}@bham.ac.uk

Abstract

A rigorous psychometric approach is crucial for the accurate measurement of mind-reading abilities. Traditional scoring methods for such tests, which involve lengthy free-text responses, require considerable time and human effort. This study investigates the use of large language models (LLMs) to automate the scoring of psychometric tests. Data were collected from participants aged 13 to 30 years and scored by trained human coders to establish a benchmark. We evaluated multiple LLMs against human assessments, exploring various prompting strategies to optimize performance and fine-tuning the models using a subset of the collected data to enhance accuracy. Our results demonstrate that LLMs can assess advanced mind-reading abilities with over 90% accuracy on average. Notably, in most test items, the LLMs achieved higher Kappa agreement with the lead coder than two trained human coders, highlighting their potential to reliably score open-response psychometric tests.

1 Introduction

Theory of Mind (ToM), commonly referred to as mind-reading, is a crucial social cognitive skill that enables individuals to understand, analyze, and use mental states to predict and explain the behavior of others (Apperly, 2010). Researchers have extensively studied the emergence and development of mind-reading abilities in young children, focusing on how they begin to grasp concepts such as perspective-taking and intention recognition (Perner et al., 1987; Wimmer and Perner, 1983; Gopnik and Astington, 1988). There is growing evidence (Apperly et al., 2011; Devine, 2021) to suggest that ToM continues to develop throughout middle childhood and adolescence and that there are individual differences in mind-reading across this age range.

Individual differences in a child's ability to understand others' perspectives remain stable over time, are frequently disrupted in clinical and mental health conditions, and have a significant impact on long-term outcomes. (Hughes and Devine, 2015). These outcomes include the quality of peer relationships, experiences of loneliness, mental health, overall well-being, and success in educational settings. Given its importance in mental health, individual differences in mind-reading offer a target for intervention. Such interventions can be tailored for individuals in therapeutic settings or applied broadly to larger populations by improving social environments. It is plausible that mind-reading will be equally important to the mental health and well-being of older adolescents and adults. However, researchers currently lack reliable and valid tools to study individual differences in mind reading beyond middle adolescence to adulthood (Yeung et al., 2024). This work addresses the significant challenges of creating sufficiently difficult mind-reading tasks that are scalable to large samples.

To create a sufficiently difficult task we reasoned that a core challenge for performing advanced mindreading is to apply mindreading abilities across a variety of people and contexts. Building on established theoretical frameworks, as outlined in previous research (Dziobek et al., 2006), we collected authentic social narratives from a demographically diverse group of individuals aged 17-18 to serve as test items, ensuring that the assessment effectively measures mind-reading ability. Story authors' interpretation of the mental states of characters in their story became the ground-truth against which mindreading accuracy was assessed. To maximize the potential for individual differences in performance, participants were asked to provide open-ended responses explaining their reasoning. This approach generated rich qualitative data that were graded by trained human coders who evaluated answers based on predefined rubrics. While this approach ensures a nuanced understanding of participants' mental state inferences, it is labour-intensive, time-

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 79–89

consuming, and prone to variation due to subjective interpretation (Devine et al., 2023).

Automation to overcome the need for human coding is needed for employing the new task at scale. However, automated coding of such responses poses challenges because, by design, the mindreading involved is highly sensitive to the story context, and the expression of correct and incorrect answers is highly variable. Recent advancements in natural language processing, particularly the development of large language models (LLMs), present a promising solution to automate this process. LLMs have demonstrated impressive capabilities in understanding, generating, and evaluating human language (Achiam et al., 2023; Dubey et al., 2024). They have been successfully used to grade free-text responses in educational settings (Xiao et al., 2024; Nilsson and Tuvstedt, 2023), making them strong candidates for evaluating individual differences in advanced mind-reading ability. However, unlike standard text classification, scoring advanced mind-reading responses is particularly challenging due to the complexity of following and applying the coding scheme consistently. Even for human coders, extensive training is required to achieve reliable scoring.

In this study, we explore the potential of LLMs to address these challenges and improve the automation of mind-reading assessment. Specifically, we investigate the following key questions:

1. How well do state-of-the-art LLMs measure advanced mind-reading ability compared to human coders?

2. What prompting strategies optimize the grading performance of these models?

3. To what extent does fine-tuning improve LLMs' grading accuracy?

To address these questions, we designed a set of mind-reading tests based on 10 selected social narratives, collected and coded responses from 1733 participants aged 13-30 before benchmarking several LLMs against human-coded scores. In particular, we assessed the impact of various prompting techniques and fine-tuning strategies on model performance. To further enhance the models, we applied data augmentation to expand the dataset, improving the effectiveness of fine-tuning. Our results show that LLMs, particularly those fine-tuned on the augmented dataset, achieve high accuracy and consistency, significantly reducing the effort required for human grading while maintaining reliability. This automated scoring approach provides clinicians with a fast, scalable, and reliable tool for assessing mind-reading ability. By addressing the scalability limitations of human-coded evaluations, it improves screening for conditions such as autism spectrum disorder and social communication disorders, where difficulties in mind-reading are prevalent (Dziobek et al., 2008; Happé, 2015). Our contributions can be summarized as follows

- We designed and implemented innovative psychological tests to measure advanced mind-reading abilities, addressing a critical need for robust and scalable assessment tools in psychometrics.

- We collected a unique dataset from participants aged 13 to 30 years and will publicly release this dataset, along with our code and fine-tuned models[1] .

- We systematically optimized the performance of LLMs through various prompting strategies and fine-tuning based on data augmentation, achieving over 90% accuracy in scoring psychometric tests.

2 Related Work

Automated grading of psychometric tests with open-ended responses has attracted significant interest in recent years. Early efforts focused on rule-based systems, which relied on manually defined patterns and logic to assess responses (Williamson et al., 2006). While these systems provided consistency in scoring, they struggled to handle the variability and nuance of open-ended responses (Burrows et al., 2015).

Over time, machine learning techniques have gained prominence as a more versatile and adaptable solution to address the problem (Mohler et al., 2011). Machine learning models frequently employ supervised learning methods, which rely on annotated datasets containing labeled examples (Bailey and Meurers, 2008; Nielsen et al., 2008; Madnani et al., 2013). These datasets enable the models to train classifiers that learn patterns and relationships between input features and their corresponding outcomes, allowing them to predict scores or make

[1]All code and data to replicate our experiments is available at https://github.com/YixiaoWang/ToM-automatic-scoring-using-LLMs/.

informed decisions when presented with unseen data. Additionally, machine learning also incorporates unsupervised learning approaches (Alfonseca and Pérez, 2004; Pérez et al., 2005; Mohler and Mihalcea, 2009). These methods identify hidden patterns, groupings, or structures within the data itself, such as clustering similar items or detecting anomalies. However, the performance of these machine learning models remained constrained by the quality and size of the training data, as well as their limited ability to capture deeper semantic understanding.

The advent of pre-trained language models marked a significant leap forward in automating text-based assessments. The model DistilBERT (Sanh, 2019), have been applied to scoring the open-response for mind-reading, where they have shown promise in scoring standardized tests of children's mind-reading (Devine et al., 2023). To further enhance the effectiveness of fine-tuning in these language models, data augmentation techniques can be employed to artificially expand the training dataset, thereby improving model generalization and robustness (Kovatchev et al., 2021). Methods such as synonym replacement, back-translation, and paraphrasing introduce variability in training samples, reducing the risk of overfitting to limited datasets.

The emergence of foundation models (Bommasani et al., 2021) trained on larger datasets with substantially more parameters to capture deeper contextual relationships has significantly enhanced performance in text-based tasks. Research efforts have successfully used LLMs to develop automatic grading systems in education setting (Xiao et al., 2024; Nilsson and Tuvstedt, 2023), enabling accurate evaluation of student writing and essay grading, often matching human evaluators in accuracy.

Assessing mind-reading ability poses significant challenges, as it requires the interpretation of nuanced psychological cues that are often deeply context-dependent, extending beyond surface-level or factual knowledge. Recent studies (Strachan et al., 2024; Kosinski, 2023; He et al., 2023) have demonstrated that LLMs are capable of making mental inferences, highlighting their suitability for this task. Although the application of LLMs to evaluate advanced mind-reading assessments remains underexplored in the broader literature, prior work by Devine et al. (2023) has made notable progress by automating the scoring of mind-reading ability using DistilBERT. This study builds on that foundation while advancing it in two key directions: (1) introducing a novel open-ended test designed for adults, which requires inferring more subtle and context-dependent mental states, and (2) leveraging LLMs instead of lightweight models to enable more sophisticated evaluations. By applying LLMs to assess responses in advanced mind-reading tests, this study seeks to further explore their potential in assessing complex human cognition.

3 Methodology

3.1 Data

The mind-reading test included 10 social narratives, each followed by a question that asked participants to interpret the mental states of the characters. An example of one such narrative, along with the corresponding mind-reading question and coding scheme, is presented in the table 1. A total of 1,733 participants aged 13–30 provided free-text responses after completing these psychometric test either in schools or online via Prolific.co. The labeling process was conducted by one lead coder and four trained coders. After an initial training phase, during which coders achieved inter-rater reliability (Cohen's Kappa > 0.7) with the lead coder, they independently coded different portions of the dataset. To ensure consistency and accuracy, each coder periodically re-coded responses from another coder. Discrepancies were resolved by the lead coder, ensuring high reliability throughout the process.

The final labeled data confirmed that the designed task was sufficiently challenging. The table 2 below shows the percentages of participants who successfully completed the mind-reading test for each of the 10 social narratives. This rigorous, multi-step process generated a high-quality, gold-standard dataset for training and evaluating LLMs. We assessed LLMs by comparing their predictions with labels of the dataset, using accuracy as the evaluation metric. Through systematic benchmarking, we aim to identify the most effective LLM for automated grading.

3.2 Model Selection

To assess the suitability of different LLMs for the mind-reading evaluation, we select a diverse set of state-of-the-art models as shown in Table 3. By comparing these models, we aim to analyze the trade-offs between model size, computational cost, and task performance for automating the mind-reading evaluation process.

Story	It was October last year, and I went to a theme park that had extra attractions for Halloween. One highlight was the "Dungeon Experience". This had actors playing characters who interact with you as you pass through it. I went in. It was really fun, but I have sensory needs and I couldn't believe how loud it was. For the first half of the experience, I had to keep my fingers in my ears, and I felt really self-conscious. I got to the bit of the experience where you get to ride on a boat through the 'Black River'. A Ferryman was wearing dark robes, limping, and carrying a lantern. He greeted us in a raspy voice then started warning us about the journey to come. He saw how I looked and put his finger up for us to wait. He hobbled off to one side, then returned a moment later and pressed a small package into my hand. It was a pack of earplugs. I put them in my ears and the ferryman caught my eye and raised an eyebrow. I gave him a thumbs up back and he grinned then returned to warning us about the journey in his raspy voice, before giving my sister two riddles to solve. He didn't even really break character!
Question	Why did she appreciate that the actor stayed in character?
Partial Coding Scheme	Correct Responses: 1 part required for a 'correct response': She appreciated that the actor was able to help her without drawing excessive attention to her sensory needs/differences (about which she felt self-conscious) (the Ferryman drew less unwanted attention by staying in character, but it was the not drawing unnecessary attention rather than specifically staying in character per se) This may be phrased in a number of different ways, for example: • He did not make her feel 'different', 'strange', or 'weird' through his actions did not make a big deal of it did not make it into an emergency (whilst also meeting her needs) • The Ferryman was able to help discretely helped without making an unnecessary fuss scene • He did not make her feel embarrassed/awkward/ self-conscious, • He did not make her feel like an inconvenience or 'a nuisance', • He did not treat her differently (other than by supporting her needs) • It meant that the actor helped her discretely (without drawing unwanted attention). • It did not make her feel more conspicuous (and therefore more self-conscious). • ... Incorrect or Incomplete Mindreading responses: Fail to mention or indicate that she was glad that the actor was able to help in a way that did not draw unwanted attention to her needs/differences • For example, responses that just mention that it didn't break the immersion for herself/others (without considering the context that made this important) [e.g. this may be expressed as 'it didn't ruin the magic'] would be incomplete mindreading responses.(The actor staying in character was his way of not drawing unnecessary attention, and not making her feel embarrassed but the not breaking immersion for others was only an add-on, not the key reason that the author appreciated him staying in character) • "It made her feel included/not left out"/"it included her in the experience" are incomplete responses, since the actor giving her the earplugs would help with this, regardless of whether he stayed in character. • Responses that focus on how helping her 'didn't ruin the experience/immersion for others' • ... Non-mindreading responses: Express an opinion on the situation, rather than trying to take the author's perspective. Or just describes the general situation without linking this to the author's experience, e.g. 'it kept it fun', 'it didn't break immersion' (for whom?)
Correct Sample	She appreciated that the actor was able to help her without drawing excessive attention to her sensory needs.
Incorrect Sample	Because the immersion wasn't totally ruined for the author and the other people in the experience.

Table 1: One Example from the 10 Test Items

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10
Proportion of Correct Answers (%)	56.7	58.3	56.8	42.1	16.1	51.4	30.7	35.7	20.8	42.2

Table 2: Percentages of participants (out of 1,733 total) accurately performing mind-reading tasks across 10 distinct test items

Model Name	Reference
allenai/longformer-base-4096	Beltagy et al. (2020)
meta/llama-3.2-3B-instruct	Dubey et al. (2024)
microsoft/phi-3.5-mini-instruct	Abdin et al. (2024a)
mistralai/mistral-7B-v0.3-Instruct	Jiang et al. (2023)
microsoft/phi-4	Abdin et al. (2024b)
openai/gpt-4o-2024-08-06	Achiam et al. (2023)
openai/gpt-4o-mini-2024-07-18	Achiam et al. (2023)
BERT	Devlin (2018)
RoBERTa	Liu (2019)

Table 3: List of Models and References

3.3 Prompt Strategies

These experiments explore the influence of various prompting strategies on the performance of LLMs in the task. We conducted a series of experiments to assess how different input formats, grading schemes, and prompting techniques impact a LLM performance. First, we compared the effect of different input formats, including plain text, XML, and JSON, on the task results. The goal was to determine whether more structured formats, such as XML and JSON, yield better results than plain text input. Next, we evaluated the impact of different grading schemes included in the prompts. The original grading scheme, which is highly detailed but often difficult for humans to interpret, was compared to two alternative formulations: a rephrased version and a summarized version generated by GPT-4o. This comparison aimed to identify which grading scheme provided the clearest and most effective guidance for mind-reading responses evaluation.

Based on the findings from the input format and grading scheme experiments, we selected the most effective combination of syntax format and grading scheme for the remaining experiments. Following the tradition in prompt engineering (Ouyang et al., 2022), the LLM to be tested was given two prompts as components: a system prompt and a user prompt. The system prompt provided the LLM with basic instructions for the task, while the user prompt contained the specific mind-reading context, including the narrative, question, corresponding grading schemes, and the participants' responses to be graded. The LLM was expected to provide a binary response (0 or 1), indicating the true value of the given response. This process is visualized in Figure 1.

Finally, we compared the performance of LLMs under zero-shot and few-shot prompting conditions. In the few-shot condition, the user prompt included a small number of labeled responses for each test. In contrast, no labeled responses from the dataset were included in the user prompt under the zero-shot condition. This comparison aimed to assess whether providing labeled responses within the prompt improves the model's performance on the mind-reading evaluation task.

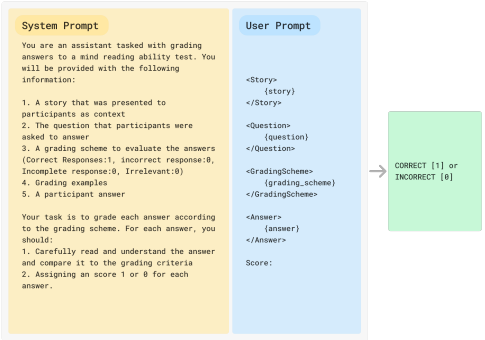

Figure 1: Prompt structure

3.4 Fine-Tuning

To enhance the performance of LLMs in grading free-text responses, we fine-tuned selected models using a labeled dataset. The dataset was split into 80% for training, 10% for validation and 10% for testing. The models to be fine-tuned are listed in Table 3. Fine-tuning aimed to align these models with the grading rubric, improving their ability to interpret and assess responses accurately. Since different models have distinct architectures and constraints, we adopted three fine-tuning strategies:

- Propreitary Models (GPT-4o, GPT-4o-mini): Fine-tuning was conducted using OpenAI's API service. Due to token limitations, we could include a maximum of 50 labeled responses per test, totaling 500 responses across 10 test items.

- Open-source Models (Llama, Mistral, and Phi): Given our computational constraints, we employed LoRA (Low-Rank Adaptation) (Hu et al., 2021) instead of full-parameter fine-tuning. LoRA performs on par with full fine-tuning, but requires significantly less memory. However, LoRA requires careful hyperparameter tuning, which we select using Bayesian search to achieve the best performance.

- BERT and RoBERTa: Fine-tuning these models differs significantly from other LLMs. Unlike other models, it involves training a binary

classifier for each test item. Each binary classifier receive individual response from its corresponding test and predicts its truth values without considering contextual elements like the question, narrative, or grading rubric.

After fine-tuning, we evaluated the models on the test set using accuracy as the primary metric.

3.5 Data Augmentation

A key challenge in fine-tuning LLMs is the limited availability of labelled training data. To address this, we investigated the role of data augmentation in enhancing fine-tuning performance. Specifically, we used GPT-4o to generate paraphrased versions of all responses in the training split of our gold-standard dataset. These paraphrased response preserved the original meaning while varying in vocabulary and sentence structure. To ensure labelling consistency, a human coder randomly selected 50 paraphrased responses per story test item for labeling. The coder's labels were then compared to those of the original responses, achieving an agreement rate of over 90%, indicating a high level of consistency. After generating the paraphrased responses, we incorporated them into the training split of the original dataset, effectively doubling the size of the available training data. This augmented dataset was then used to fine-tune the LLMs.

4 Results and Analysis

Syntax Format	Accuracy
Plain	0.82
XML	**0.84**
JSON	0.83

Table 4: Results testing the effect of syntax format in prompting GPT-4o in terms of grading accuracy.

Scheme	Accuracy
Original Grading Scheme	**0.88**
Grading Scheme Summary	0.86
Paraphrase Summary Scheme	0.85

Table 5: Results testing the effect of grading scheme on GPT-4o prompt in terms of grading accuracy.

4.1 Result of Prompt Engineering

We first analyze how different input formats affected the performance of the LLM (GPT-4o) in the mind-reading evaluation task. As shown in the table 4, structured formats, XML and JSON, slightly outperform plain text in terms of accuracy. Then, we compared the effect of different grading schemes incorporated into the prompts. As shown in the Table 5, the original grading scheme, although highly detailed and challenging for human coders to employ consistently, surprisingly produced the better results, outperforming both the rephrased version and summarized version generated by GPT-4o. This finding suggests that, despite the complexity of the original scheme, LLMs are capable of capturing the relevant information embedded in highly detailed text. Based on these results, we use XML as prompt syntax and original grading scheme as default coding rubric to prompt all LLMs, both in zero-shot setting and few-shot setting. The detailed performance of zero-shot prompting and few-shot prompting are included in the Table 6 and Table 7.

In the zero-shot condition, each model's performance was assessed by comparing the results to those assigned by trained human coders and calculating its accuracy in scoring answers for each test. Overall performance was determined by averaging accuracy rates across all 10 test items. Among the models tested, GPT-4o achieved the highest accuracy at 89.4 %, significantly outperforming the others. Phi-4 followed with a strong 81.5%, while Mistral-7B and Phi-3.5 scored 77.1% and 73.5%, respectively. Llama-3.2 trailed at 64.3%, and Longformer, the smallest model in the table, lagged further at just 50%—likely due to its limited capacity to process complex information. These results indicate that larger language models tend to perform better on mind-reading ability scoring task.

Building on the observation from zero-shot results, we evaluated the performance of the best-performing model, GPT-4o, along with GPT-4o-mini, under few-shot conditions. These models were selected due to their outstanding performance in the zero-shot evaluation and their larger capacity to handle more complex prompts. In the first few-shot test, where 10 labelled answers from the dataset were provided for each test, we observed a slight improvement in performance for both models. GPT-4o achieved an accuracy rate of 89.5%, marginally outperforming its zero-shot result of 89.4%. Similarly, GPT-4o-mini saw an increase, with its accuracy rate rising to 81.4% from 79.7% in the zero-shot condition. However, when the number of labelled answer was increased to 50 for each test, the results shifted, GPT-4o's accuracy

rate decreased to 88.1%, and GPT-4o-mini's accuracy rate dropped to 80.1%. These results highlight an important insight in few-shot prompting: While providing a certain number of examples can enhance model performance, increasing this number beyond a certain threshold does not always lead to improved outcomes.

4.2 Result of LLMs fine-tuning

As is shown in the Table 8, all models in the evaluation show significant improvements after fine-tuning, highlighting the effectiveness of this approach for the task of psychometric scoring. GPT-4o achieves the best results, with its accuracy increasing from 89.4% to 92.8%. Notably, its performance is further supported by a kappa value of 0.83, indicating strong agreement that far exceeds what would be expected by chance. GPT-4o-mini benefits greatly from fine-tuning, rising from 79.7% to 90.5%. This success is particularly remarkable considering that GPT-4o-mini was fine-tuned on only 50 examples per test. Longformer, initially starting at 50.0%, shows a remarkable jump to 86.7%, and Llama moves from 64.3% to 91.1%. Models like Mistral and Phi-4, which started with strong zero-shot accuracy, also see significant improvements. These results underscore the substantial benefits of fine-tuning in improving model accuracy.

Notably, the BERT family of models has demonstrated impressive performance despite their smaller sizes. BERT-base and BERT-large achieved accuracies of 90.2% and 90.5%, respectively, matching or even surpassing larger models like GPT-4o-mini. This is particularly remarkable given BERT's more compact architecture, highlighting its competitive edge when fine-tuned for specific test items. However, fine-tuning BERT models differs significantly from that of other LLMs. Unlike LLMs, which are fine-tuned as single scoring systems to handle all test items, BERT and RoBERTa are trained into 10 distinct classifiers, each dedicated to a specific test item. These classifiers are test-specific and cannot be transferred to other test items, so while their specialization enhances performance on individual test, it limits their flexibility across a range of tests. Additionally, BERT and RoBERTa fall short of LLMs in providing explanations or feedback to justify the scores they assign, making their high performance both impressive and somewhat constrained in comparison.

4.3 Effect of Data Augmentation

Data augmentation has a positive effect on performance for most models, although the improvements are not consistent across all of them. Longformer sees a notable gain, increasing from 86.7% to 91.6%, demonstrating the clear benefit of augmented data. Mistral and Phi-3.5 also benefit, with Mistral rising from 88.7% to 91.6% and Phi-3.5 improving from 83.8% to 90.1%. However, Llama experiences a slight drop, from 91.1% to 90.5%, and Phi-4 shows only a small increase, from 87.5% to 87.6%. These results indicate that while data augmentation often enhances model accuracy, its impact can vary depending on the model and test item.

4.4 Comparison between human coders and LLMs

Building on the previous findings that LLMs can grade psychometric tests with high accuracy, we now compare their performance to human coders in both accuracy and efficiency. Initially, all four trained human coders demonstrated adequate Inter-Rater Reliability (Cohen's Kappa > .7) with the lead coder across 10 test items before being assigned different portions of the main dataset to code. However, the following spot checks revealed that two trained coders drifted in their application of the marking criteria for certain test items. To address this, the fine-tuned GPT-4o was used to reassess all participant responses for those cases. Whenever the LLM and the trained coder disagreed, the lead coder made the final decision. The table below summarizes the relative accuracy of human coders and the fine-tuned GPT-4o under this procedure. Importantly, the Kappa agreement score was calculated only for cases where the LLM and the human coder initially disagreed. The results indicate a clear trend: except for test item 5, fine-tuned GPT-4o consistently showed higher agreement with the lead coder than the trained coders did. This suggests that, for the majority of test items (1, 2, 3, 6, 7, 8, and 10), the LLM provided more reliable coding across many cases.

In terms of time efficiency, training a single human coder requires at least 14 hours before they can pass the Inter-Rater Reliability check. With four human coders trained, this amounts to a total of 56 hours of training time. After passing the check, each coder takes an average of 33 seconds to grade a single response. Given 10 test items and 1,733

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	AVG
gpt-4o	**88.5**	**86.7**	**90.2**	**91.3**	**87.3**	**91.9**	**94.2**	78.1	**95.9**	**90.2**	**89.4**
gpt-4o-mini	80.4	83.3	83.9	82.1	80.1	83.9	81	73.5	65.5	83.3	79.7
longformer-4096-base	68.5	63.3	66.5	49.7	22.3	60.2	42.6	47.1	28.5	48	50
Llama-3.2-3B	72.7	78.6	71.5	66.2	43.9	70.7	65.5	61.4	48.8	61.7	64.3
mistral-7b-v0.3-instruct	81.1	78.6	77.8	74.5	82	78.2	87.1	70	65.8	74.7	77.1
phi-3.5-mini-instruct	81.8	79.4	72.1	65.6	81.3	73.7	75	72.8	72.3	63.6	73.5
phi-4	86	78.6	79.1	83.4	87.1	82.7	71	**80**	84.5	83.1	81.5

Table 6: Evaluation results of LLMs on 10 psychometric tests using zero-shot prompting. Results are reported in terms of accuracy (%)

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	AVG
gpt-4o (10 shots)	83.9	87.3	**90.2**	91.3	**89.6**	**91.9**	**90.8**	**85**	**96.5**	**89**	**89.5**
gpt-4o-mini (10 shots)	78.7	85.6	82.7	86.2	81	84.4	83.4	77.5	68.9	85.6	81.4
gpt-4o (50 shots)	**86.2**	**89.6**	86.2	**93.6**	89	86.7	89	82.7	94.8	83.9	88.17
gpt-4o-mini (50 shots)	80.4	87.3	85	81	80.2	83.9	79.8	77.5	64.3	82.1	80.1

Table 7: Grading results of LLMs on 10 psychometric tests using few-shot prompting. Results are reported in terms of accuracy rate (%)

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	AVG
BERT-base	86.2	83.3	90.2	93.1	94.8	91.9	89	87.3	94.2	90.2	90.2
BERT-large	**89.6**	83.3	90.8	93.6	93.1	88.5	91.9	88.5	92.5	92.5	90.4
RoBERTa-base	88.5	86.7	93.6	91.9	91.9	91.9	91.3	83.3	94.2	90.8	90.4
RoBERTa-large	89.6	87.9	**94.8**	92.5	94.8	91.3	92.5	87.3	95.4	94.2	92
gpt-4o	89	**91.3**	93.1	94.2	**94.8**	**93.6**	**94.2**	85	**97.1**	96.5	**92.8**
gpt-4o-mini	86.7	83.3	91.9	94.8	93.1	92.5	90.8	79.3	97.1	95.9	90.5
longformer-4096-base	87.0	80.9	91.7	94.9	77.7	89.5	87.1	89.3	81.3	87.6	86.7
Llama-3.2-3B	86.7	85.5	91.1	**96.8**	90.6	90.2	87.8	**90**	95.1	**97.4**	91.1
mistral-7b-v0.3-instruct	86.7	80.9	89.2	94.9	84.9	91.7	89.2	82.1	91.9	95.4	88.7
phi-3.5-mini-instruct	76.2	86.3	89.9	69.3	87.1	80.4	85.8	88.6	90.2	85.7	83.8
phi-4	87.4	84	84.8	90.4	89.9	83.5	85.8	86.4	91.9	90.9	87.5

Table 8: Results of evaluation of fine-tuned LLMs in the 10 psychometric tests. Results are reported in terms of accuracy (%)

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	AVG
longformer-4096-base	**91.6**	83.9	**94.3**	94.9	89.9	91	**90.5**	90.7	93.5	96.1	91.6
Llama-3.2-3B	85.3	86.2	93	**94.9**	89.2	89.4	89.8	89.3	93.5	94.1	90.5
mistral-7b-v0.3-instruct	90.9	**87**	90.5	94.3	91.4	90.2	89.2	**91.4**	94.3	**97.4**	**91.6**
phi-3.5-mini-instruct	88.8	81.7	87.3	93.6	**92.8**	90.2	87.8	87.9	**96.7**	94.2	90.1
phi-4	86	85.5	84.8	91.7	89.2	**91**	88.5	81.4	88.6	88.9	87.6

Table 9: Results of evaluation of fine-tuned LLMs on augmented train split in the 10 psychometric tests. Results are reported in terms of accuracy (%)

Participant Numbers	ID number assigned to human coder	Item (Story and Question)	Kappa agreement of human coder with lead coder	Kappa agreement of GPT-4o with lead coder
1325–2013	1	2	0.791	0.906
		5	0.873	0.786
		6	0.758	0.837
2014–2157	3	1	0.929	0.971
		2	0.756	0.878
		3	0.889	0.845
		5	0.974	0.638
		8	0.580	0.928
		10	0.833	0.889

Table 10: Kappa agreement of trained human coder and GPT-4o with lead coder.

participants, the entire dataset requires approximately 158 hours to label. In contrast, fine-tuning LLMs (e.g., Llama-3.2-3B) takes approximately 16-24 hours, including 8-16 hours for hyperparameter tuning. Once fine-tuned, the LLM can score each response in milliseconds, a dramatic reduction compared to the time required by human coders. This highlights the LLM's exceptional efficiency in processing speed.

5 Discussion

Our findings highlight the transformative potential of LLMs in automating the scoring of open-ended responses in complex mind-reading tests. Fine-tuning, particularly when paired with augmented training data, enables LLMs to better grasp the test-specific nuances of intricate coding manuals, resulting in more accurate evaluation. Despite the inherent complexity of the task, LLMs demonstrated an impressive ability to interpret and apply these detailed coding guidelines effectively. This adaptability suggests that LLMs could be valuable tools for automating the scoring of other psychometric tests, particularly those that involve open-ended responses. Such applications could help overcome the ceiling effect often seen in closed-ended questions, making it possible to quantify reliably the abilities of more developmentally advanced participants (i.e. older adolescents and adults) than has previously been possible.

Our exploration of prompt strategies further revealed that a relatively small number of examples led to noticeable improvements in performance. However, increasing the number of examples beyond a certain point did not produce gains. As our results show, fine-tuning is a more effective strategy than prompting, particularly when leveraging a larger set of examples to enhance model performance. This highlights that fine-tuning, rather than prompting, is the more powerful tool for maximizing LLM capabilities in psychometric task scoring.

Furthermore, the BERT family of models continues to be a highly effective and practical choice for scoring open-response psychometric tasks. While a BERT classifier trained on one test may not directly transfer to others due to the distinct nature of each test, its strength lies in its simplicity and computational efficiency. BERT models are relatively easy to implement and require fewer computational resources compared to other LLMs, making them an ideal option for users with limited computational resources or specific task requirements.

6 Conclusion

This paper demonstrates the effectiveness of LLMs in scoring psychometric tests designed to assess advanced mind-reading ability. By optimizing prompting strategies and fine-tuning models, we achieve results that not only align closely with human evaluations but also surpass the performance of some trained human coders on most of test items. This highlights LLMs' potential to reliably assess complex cognitive processes, offering a scalable, efficient, and consistent approach to psychometric testing. While current methods use LLMs to evaluate responses against pre-defined answers, LLMs also excel at analyzing patterns in mind-reading responses. This goes beyond identifying performance gaps in individuals with neurodevelopmental or psychiatric conditions, allowing researchers to explore whether they mind-read in systematically different ways. Such insights could transform our understanding of individual differences in mind-read processes. Future work should explore these applications, further expanding the utility of LLMs in psychometric research.

7 Limitations

While the performance of LLMs in scoring mind-reading responses is impressive, it raises the question of what enables them to excel in this task. Are LLMs inherently skilled at mind-reading, allowing them to assess responses reliably, or do they simply follow the complex coding manual with high accuracy? This study does not provide a definitive

answer, and further research is needed to explore the underlying mechanisms of LLM judgment.

8 Ethical Considerations

The project gained ethical review and approval from the Science Technology Engineering and Mathematics ethical review panel at the University of Birmingham UK, project approval ID: ERN_2311-Jun2024.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Enrique Alfonseca and Diana Pérez. 2004. Automatic assessment of open ended questions with a bleu-inspired algorithm and shallow nlp. In *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4*, pages 25–35. Springer.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Ian A Apperly, Frances Warren, Benjamin J Andrews, Jay Grant, and Sophie Todd. 2011. Developmental continuity in theory of mind: Speed and accuracy of belief–desire reasoning in children and adults. *Child development*, 82(5):1691–1703.

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 107–115.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Rory T Devine. 2021. Individual differences in theory of mind in middle childhood and adolescence. In *Theory of mind in middle childhood and adolescence*, pages 55–76. Routledge.

Rory T Devine, Venelin Kovatchev, Imogen Grumley Traynor, Phillip Smith, and Mark Lee. 2023. Machine learning and deep learning systems for automated measurement of "advanced" theory of mind: Reliability and validity in children and adolescents. *Psychological Assessment*, 35(2):165.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Isabel Dziobek, Stefan Fleck, Elke Kalbe, Kimberley Rogers, Jason Hassenstab, Matthias Brand, Josef Kessler, Jan K Woike, Oliver T Wolf, and Antonio Convit. 2006. Introducing masc: a movie for the assessment of social cognition. *Journal of autism and developmental disorders*, 36:623–636.

Isabel Dziobek, Kimberley Rogers, Stefan Fleck, Markus Bahnemann, Hauke R Heekeren, Oliver T Wolf, and Antonio Convit. 2008. Dissociation of cognitive and emotional empathy in adults with asperger syndrome using the multifaceted empathy test (met). *Journal of autism and developmental disorders*, 38:464–473.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

FRANCESCA Happé. 2015. Autism as a neurodevelopmental disorder of mind-reading. *Journal of the British Academy*, 3(1):197–209.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Claire Hughes and Rory T Devine. 2015. Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child development perspectives*, 9(3):149–153.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Venelin Kovatchev, Phillip Smith, Mark Lee, and Rory Devine. 2021. Can vectors read minds better than experts? comparing data augmentation strategies for the automated scoring of children's mindreading ability. *arXiv preprint arXiv:2106.01635*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated scoring of summary-writing tasks designed to measure reading comprehension. *Grantee Submission*.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 752–762.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.

Rodney D Nielsen, Wayne H Ward, and James H Martin. 2008. Learning to assess low-level conceptual understanding. In *FLAIRS*, pages 427–432.

Filippa Nilsson and Jonatan Tuvstedt. 2023. Gpt-4 as an automatic grader: The accuracy of grades set by gpt-4 on introductory programming assignments.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Diana Pérez, Enrique Alfonseca, Pilar Rodríguez, Alfio Gliozzo, Carlo Strapparava, and Bernardo Magnini. 2005. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista signos*, 38(59):325–343.

Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

David M Williamson, Robert J Mislevy, and Isaac I Bejar. 2006. *Automated scoring of complex tasks in computer-based testing*. Lawrence Erlbaum Associates Mahwah, NJ.

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):103–128.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Elaine Kit Ling Yeung, Ian A Apperly, and Rory T Devine. 2024. Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, 157:105481.