Scientific Reports

https://doi.org/10.1038/s41598-026-39292-y

Article in Press

Evaluation of cross-ethnic emotion recognition capabilities in multimodal large language models using the reading the mind in the eyes test

Received: 22 July 2025

Accepted: 4 February 2026

Published online: 20 February 2026

Cite this article as: Refoua E., Elyoseph Z., Piterman D. *et al.* Evaluation of cross-ethnic emotion recognition capabilities in multimodal large language models using the reading the mind in the eyes test. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39292-y

Elad Refoua, Zohar Elyoseph, David Piterman, Alon Geller, Gunther Meinlschmidt & Dorit Hadar Shoval

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Evaluation of Cross-Ethnic Emotion Recognition Capabilities in Multimodal Large Language Models Using the Reading the Mind in the Eyes Test

Elad Refoua+*[1], Zohar Elyoseph+* [2,7], David Piterman[3], Alon Geller[4], Gunther Meinlschmidt+*[5,6] and Dorit Hadar-shoval+*[3]

+These authors contributed equally to this work and share first/last authorship

[1] Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel

[2] School of Counseling and Human Development, University of Haifa , Haifa, Israel

[3] Department of Psychology, Tel - Hai, University of Kiryat Shmona and the Galilee, Israel

[4] Ruppin Academic Center, Emek Hefer, Israel

[5] Clinical Psychology and Psychotherapy – Methods and Approaches, Department of Psychology, Trier University, Trier, Germany

[6] University of Basel and University Hospital Basel, Department of Digital and Blended Psychosomatics and Psychotherapy, Psychosomatic Medicine, Basel, Switzerland

[7] Imperial College London, London, United Kingdom

*corresponding authors:

Zohar Elyoseph: zohar.j.a@gmail.com

Elad Refoua: eladrefoua@gmail.com

Abstract

Background: Accurate emotion recognition is a foundational component of social cognition, yet human biases can compromise its reliability. The emergent capabilities of multimodal large language models (MLLMs) offer a potential avenue for objective analysis, but their performance has been tested mainly with ethnically homogenous stimuli. This study provides a systematic cross-ethnic evaluation of leading MLLMs on an emotion recognition task to assess their accuracy and consistency across diverse groups.

Methods: We evaluated three leading MLLMs: ChatGPT-4, ChatGPT-4o, and Claude 3 Opus. Performance was tested twice using three "Reading the Mind in the Eyes Test" (RMET) versions featuring White, Black, and Korean faces. We analyzed accuracy against chance (25%) and compared scores to established human normative data for each ethnic version.

Results: ChatGPT-4o achieved performance significantly above chance levels across all tests (p < .001), with large effect sizes indicating robust performance (Cohen’s h = 1.253-1.619; RD = 0.583-0.694). The model obtained a mean accuracy of 83.3% (30/36) on the White RMET, 94.4% (34/36) on the Black RMET, and 86.1% (31/36) on the Korean RMET, placing it in the 85th, 94th, and 90th percentiles of human norms, respectively. This high accuracy remained consistent across ethnic stimuli. In contrast, ChatGPT-4 performed near the human average, while Claude 3 Opus performed near chance level.

Conclusion: These preliminary findings highlight the rapid evolution of MLLMs, highlighting a significant performance leap between consecutive versions.

This study suggests that ChatGPT-4o demonstrated performance scores exceeding average human accuracy on this specific task in recognizing complex emotions from static images of the eye region, with its performance remaining consistent across different ethnic groups. While these results are notable, the pronounced performance gaps between models and the inherent limitations of the RMET task underscore the need for continuous validation and careful, ethical consideration to fully understand the capabilities and boundaries of this technology.

Keywords: Generative Artificial Intelligence (GenAI), Emotion Recognition, Cross-Cultural Psychology, Psychiatric Diagnosis, Reading the Mind in the Eyes Test (RMET), Bias, Mental Health.

Evaluation of Cross-Ethnic Emotion Recognition Capabilities in Multimodal Large Language Models Using the Reading the Mind in the Eyes Test

Introduction

Recent advancements in Artificial Intelligence (AI), particularly the emergence of multimodal large language models (MLLMs), have demonstrated emergent capabilities in complex psychological reasoning [1, 2]. This has prompted a wave of research evaluating their performance on established psychological assessments, moving beyond simple task completion to detailed psychometric profiling. For instance, studies have successfully utilized LLM-driven conversational agents to conduct Big Five personality tests [3], applied clinical diagnostic tools to profile AI models for maladaptive traits analogous to human personality disorders [4], and explored gamified interactions with multi-personality agents to assess user traits [5]. One of the most intriguing frontiers is social cognition, the ability to perceive and interpret subtle emotional and social cues, which is fundamental to human interaction [6, 7]. Beyond basic emotion recognition, recent research highlights MLLMs' growing proficiency in complex social cognition tasks, such as Theory of Mind reasoning and understanding sarcasm [1, 2, 5]. However, alongside these capabilities, concerns have emerged regarding their potential to propagate deep-seated human biases. Studies have shown that AI models can inadvertently mirror societal prejudices, including racial and ethnic biases, which may affect their judgment in social contexts

[8, 9]. This duality—advanced social reasoning coupled with the risk of inherent bias—underscores the need for rigorous evaluation across diverse demographic stimuli. As these models are increasingly explored for various applications, a critical need arises to systematically evaluate their performance, consistency, and potential biases on subtle socio-emotional tasks. Fields such as psychology and psychiatry, which rely heavily on interpreting such cues, provide a relevant context for assessing these advanced capabilities [10].

For decades, the field of affective computing has explored the potential of AI to provide objective and scalable solutions for emotion recognition, traditionally a resource-intensive process prone to subjective biases [11 – 13]. Building on this, recent studies indicate that MLLMs can outperform the average human on standardized emotional intelligence tests [14] and recognize basic facial emotions with an accuracy comparable or even superior to human judges [15, 16]. These findings establish a strong basis for investigating the utility of these models in more socially complex assessment contexts.

The present study builds directly upon a pilot investigation that used the "Reading the Mind in the Eyes" Test (RMET), a key instrument for assessing Theory of Mind (ToM) [17], to probe the capabilities of an earlier MLLM. The RMET challenges individuals to infer complex mental states from cropped photographs of the eye region, thereby avoiding the ceiling effects common in simpler ToM tasks [17, 18].

The pilot study found that an early version of ChatGPT (GPT-4) performed significantly above chance level when presented with the original RMET, which features exclusively White stimuli [19]. This promising initial result demonstrated the model's potential in this domain and served as the impetus for the current, more comprehensive investigation.

However, the pilot study's reliance on a single, ethnically homogenous stimulus set represents a significant methodological limitation. Extensive research documents cross-cultural differences in the perception of facial expressions [20, 21] and the existence of an "other-race effect" in human face processing, where individuals are often less accurate at interpreting faces from different ethnic backgrounds [21, 22]. This limitation is particularly critical for MLLMs, given documented instances of models perpetuating Western-centric values [23, 24] and racial biases [8, 9]. To ensure fairness and a robust evaluation, it is imperative to validate these models using stimuli that reflect demographic diversity [16, 25].

To address this gap, researchers have developed and validated ethnically adapted versions of the RMET, including those featuring Black (B-RMET) [16] and Korean faces (K-RMET) [26]. These instruments provide a crucial opportunity for a more rigorous and equitable evaluation of MLLMs' capabilities. While some recent studies on facial perception suggest that certain models may not exhibit a strong racial bias in judging social traits or basic emotions

[16, 27], these findings are preliminary and require systematic replication across different models and more complex cognitive tasks like the RMET. The present study addresses this need by conducting the first cross-ethnic evaluation of advanced ToM in leading MLLMs.

This study, therefore, evaluates and compares the performance of three leading models, ChatGPT-4 [28], the more recent ChatGPT-4o [29], and Claude 3 Opus [30], across the White, Black, and Korean versions of the RMET. These specific models were selected to represent the forefront of commercially available multimodal capabilities at the time of the study, enabling a comparison between different iterations of the industry-standard GPT architecture and a leading competitor, Claude 3, to assess consistency across different systems. By systematically assessing both the accuracy and the cross-ethnic consistency of these models, we aim to determine the extent to which their advanced ToM abilities generalize across diverse social stimuli, thereby providing crucial insights for their responsible development and future research. Despite these documented biases, regarding cross-ethnic consistency, our hypothesis was grounded in two expectations. First, we posited that advanced MLLMs possess robust visual generalization capabilities, allowing them to analyze facial morphology and emotional expression independent of specific demographic features, rather than relying on memorization of the widespread original RMET images. Second, recent advancements in safety alignment and RLHF aim to mitigate historical algorithmic biases. Therefore, we hypothesized that these models would demonstrate consistent performance across ethnic groups.

Based on the foregoing literature, this study tested the following hypotheses:

RH1: (Performance Above Chance and Human Average): We hypothesized that all cutting-edge MLLMs will demonstrate statistically significant emotion recognition performance above chance levels on all RMET versions. Furthermore, it is hypothesized that ChatGPT-4o will achieve above-average human performance on all RMET versions.

RH2: (Cross-Ethnic Consistency): We hypothesized that the emotion recognition performance of the MLLMs will remain consistent across different ethnic versions of the RMET (White, Black, and Korean faces).

RH3: (Inter-Model Comparison): We hypothesized that there will be discernible differences in emotion recognition performance among the various MLLMs tested (ChatGPT-4, ChatGPT-4o, and Claude 3 Opus).

Results

We provide an example of ChatGPT-4o’s responses in Figure 1. In Table 1, we present the descriptive statistics, performance comparison and effect sizes of ChatGPT-4, ChatGPT-4o and Claude 3 Opus, compared to human samples [17, 18, 26].

Performance above chance levels (RH1)

Binomial tests, incorporating Bonferroni corrections for multiple comparisons across 18 measures, revealed significant differences from chance performance ($0.001 < p < 0.05$) for ChatGPT-4 and ChatGPT-4o across all administrations. The effect sizes, reported as Cohen's h and Risk Difference (RD), demonstrated particularly robust performance for ChatGPT-4o, with Cohen's h ranging from 1.253 to 1.619 and RD values between 0.583 and 0.694, indicating substantial deviation from chance performance. ChatGPT-4 showed moderate to strong effects (Cohen's h: 0.524 to 0.923; RD: 0.25 to 0.444). In contrast, Claude 3 Opus demonstrated more limited capabilities, achieving statistical significance in only one instance (B-RMET first administration: 19/36 correct, h = 0.579, RD = 0.278, $p < .05$).

Table 2 presents the performance of the MLLMs in comparison to the percentiles of human samples across the different RMET versions. The performance of ChatGPT-4o quantitatively exceeded the median of the human samples, achieving percentiles of 85, 94, and 90 for the RMET versions with White, Black, and East Asia faces, respectively. This indicates that ChatGPT-4o performed better than 85% of the human sample on the White face version of the RMET, better than 94% of the human sample on the Black face version, and better than 90% on the Korean face version when compared to established human normative samples. Conversely, ChatGPT-4 demonstrated lower performance, with ChatGPT-4 obtaining percentiles of 9.6, 43, and 17, and Claude 3 Opus achieving substantially lower percentiles of 0.012, 4.00, and 0.045 for the White, Black, and Korean face version, respectively.

In summary, the findings support RH1, demonstrating that both ChatGPT-4 and ChatGPT-4o exhibited emotion recognition performance significantly above chance levels across all RMET versions. Specifically, ChatGPT-4o consistently achieved above-average human performance across all ethnic RMET versions, supporting the latter part of the hypothesis.

Consistency and generalizability across ethnic groups (RH2)

Notably, the comparison between the RMET scores across versions revealed minimal differences in the percentile rankings for each model, suggesting consistent performance across different ethnic contexts of the RMET. Specifically, the small differences between the percentiles between each RMET version indicate that the MLLMs' abilities to interpret emotional states from the eyes were similarly accurate, or inaccurate, irrespective of the ethnic adaptation of the test.

Regarding the level of agreement between two evaluations of the same MLLMs (ChatGPT-4, ChatGPT-4o or Claude 3 Opus), when responding to the RMET version with White face stimuli, ChatGPT-4 showed moderate test-retest agreement (0.69), ChatGPT-4o showed the highest test-retest agreement (0.94) and Claude 3 Opus showed a substantial test-retest agreement (0.83). When responding to the B-RMET, ChatGPT-4 showed substantial test-retest agreement (0.94), ChatGPT-4o showed similar test-retest agreement (0.94) and Claude 3 Opus showed moderate test-retest agreement (0.77). When responding to the RMET version with Korean faces, ChatGPT-4 showed moderate test-retest agreement (0.75) , ChatGPT-4o again showed the highest test-retest agreement (0.83) and Claude 3 Opus showed moderate test-retest agreement (0.58). These results suggest that ChatGPT-4o consistently showed the highest agreement between repeated testing across ethnic contexts, as compared to the other models.

In summary, the consistent percentile rankings and high test-retest agreement levels across different ethnic versions of the RMET for each LLM support RH2, indicating that their emotion recognition capabilities generalize across various ethnic contexts.

Comparison between MLLMs (RH3)

As detailed above, ChatGPT-4o consistently achieved the highest performance, with percentile rankings that substantially exceeded those of ChatGPT-4 and Claude 3 Opus across all RMET versions. While ChatGPT-4 also showed significant performance above chance, its percentile rankings were considerably lower than ChatGPT-4o. Claude 3 Opus exhibited the most limited capabilities among the tested models, reaching statistical significance in only one instance and generally achieving substantially lower percentiles.

In summary, the findings support RH3, revealing clear discernible differences in emotion recognition performance among the LLM models, with ChatGPT-4o consistently outperforming both ChatGPT-4 and Claude 3 Opus.

Item difficulty and thematic error analysis

To further explore the models' performance beyond overall accuracy, we conducted an item difficulty analysis to identify which mental states were systematically easier or more difficult for the MLLMs to recognize. Error rates were calculated for each of the 36 items across all models and test versions. A clear pattern emerged, distinguishing items that were consistently challenging from those that were consistently easy for the models. Table 3 displays the ten most difficult and ten least difficult items for the MLLMs, based on their aggregated error rates.

A descriptive model-by-model analysis revealed that this pattern of difficulty was most pronounced in Claude 3 Opus and ChatGPT-4. In contrast, ChatGPT-4o, successfully overcame many of these challenges, demonstrating lower error rates on the same items that were difficult for the other models. This suggests a notable evolution in capability between model versions.

A thematic analysis of these error patterns revealed a distinct conceptual divide. The models demonstrated high accuracy in identifying internal cognitive and emotional states, especially those with negative or neutral valence (e.g., *contemplative*, *uneasy*, *worried*). Conversely, they exhibited significantly higher error rates for socially complex states that imply a behavioral or interpersonal intention, particularly those with a positive valence (e.g., *playful*, *friendly*, *flirtatious*).

Discussion

This study evaluated the emotion recognition capabilities of three MLLMs across ethnically diverse stimuli using adapted versions of the Reading the Mind in the Eyes Test. The results provide insights into the current capabilities and limitations of these rapidly evolving technologies. The study yielded three principal findings. First, the most advanced model, ChatGPT-4o, demonstrated performance significantly above the human average, consistently scoring in the upper percentiles of human normative samples.

Second, this high level of performance was consistent across the RMET versions featuring White, Black, and Korean faces, suggesting an absence of ethnic bias in this specific visual recognition task. Third, there were significant performance disparities among the models, with the newer ChatGPT-4o showing significantly higher performance than its predecessor, ChatGPT-4, and the alternative architecture, Claude 3 Opus, which highlights the rapid evolution of these capabilities.

The notably poor performance of Claude 3 Opus, which consistently operated near chance level, warrants specific consideration. While a direct comparison of proprietary models is not possible, recent literature suggests a confluence of potential factors. First, studies indicate that while Claude models are highly capable, they can underperform in tasks requiring the detection of highly subtle or implicit sentiments, such as sarcasm, compared to competitors [5, 31]. The RMET, being a test of inference from minimal cues, may thus fall into a category of tasks where Claude is less proficient. Second, this tendency may be rooted in the model's foundational "Constitutional AI" safety alignment, which prioritizes cautiousness and can lead to a refusal to engage with ambiguous prompts [32, 33]. In a forced-choice task like the RMET, this conservative approach could result in performance that does not significantly differ from random guessing. Finally, a model's visual process is not monolithic; for instance, studies in radiology have found that different models excel at distinct visual tasks, such as Claude 3.5 Sonnet being more accurate in identifying anatomy while GPT-4o is superior in detecting fractures [34]. It is therefore plausible that Claude's visual architecture is optimized for different types of analysis than the fine-grained interpretation of human facial expressions.

Additionally, our exploratory item difficulty analysis suggests that this performance gap may be linked to a discernible thematic pattern. A possible interpretation of the results is that the MLLMs performed well when identifying internal cognitive and emotional states, particularly those with a negative or neutral valence (e.g ,. *worried* ,*contemplative* ,*uneasy*). Conversely, the models appeared to struggle more with socially complex states that imply a behavioral or interpersonal intention, especially those with a positive valence (e.g ,.*playful* ,*friendly* ,*flirtatious*). While the improved performance of ChatGPT-4o on these specific items could indicate an evolution in this capability, further research is required to confirm this pattern and explore its underlying causes.

Performance and cross-ethnic consistency of MLLMs

A key finding of this study is that MLLMs can achieve accuracy on the RMET that is not only significantly above the human average but also consistent across ethnic groups. This cross-ethnic consistency is particularly noteworthy. While humans often exhibit an "other-race effect," leading to diminished accuracy when interpreting faces from different ethnic backgrounds [21, 22], ChatGPT-4o's performance did not show such a discrepancy. This suggests that, for this specific task, the model's visual processing does not appear to be constrained by the same biases that affect human perception. This finding provides a valuable benchmark for the development of fair AI models and suggests that they may not necessarily replicate specific human cognitive biases in visual emotion recognition tasks [35, 36].

Model evolution and the shifting landscape of AI capabilities

The stark performance difference between ChatGPT-4o and the other models is a critical finding. It demonstrates that advanced visual-perceptual ability is not an inherent feature of all MLLMs but rather a rapidly evolving capability.

While the exact technical mechanisms remain proprietary, the performance gap suggests that ChatGPT-4o possesses a significantly enhanced visual sensitivity compared to its predecessor. From a functional perspective, the newer model appears better equipped to process subtle facial cues (such as the intricate muscle movements around the eyes) and integrate them with semantic knowledge. This leap in accuracy also raises cautious optimism regarding cognitive remediation. If future research confirms that such models can provide reliable, real-time feedback on social cues, they might eventually serve as accessible 'training partners' for individuals with social-cognitive deficits, supplementing traditional therapeutic interventions.

Beyond these clinical possibilities, the rapid evolution of these models has practical implications for the development of any AI-driven application that relies on visual interpretation. An application built on an MLLM today could be rendered functionally inferior by a newer model released months later.

This "moving target" nature of MLLM performance creates a technological imperative. The selection of the underlying model (e.g., GPT-4 vs. GPT-4o) is not a trivial implementation detail but a critical variable that directly impacts performance accuracy [37, 38]. Therefore, we recommend that applications leveraging MLLMs be designed with a "dynamic" or "pluggable" architecture. Such a design would allow systems to be updated with the latest, most validated version with minimal friction, ensuring that they benefit from the most accurate technology available and do not become rapidly obsolete [37, 38].

Ethical considerations

As these technologies approach and potentially surpass human performance on specific tasks, the ethical conversation moves beyond general concerns about bias and privacy into new territory. The findings raise two ethical dilemmas:

The Dilemma of the High-Performing Tool: If a tool is demonstrably more accurate and less biased than a human on a specific task, what are the ethical ramifications for professional standards? This questions traditional notions of expertise and liability. For fields that rely on human judgment, at what point might it become ethically questionable not to use such a tool as a secondary check, especially in complex cases? This finding presses professional communities to define new standards of care that incorporate these advanced capabilities [25].

The Dilemma of the Objective Tool in a Biased System: While ChatGPT-4o showed no ethnic bias on this task, it would be deployed within healthcare systems or other societal structures that may have deep-seated systemic biases. There is a significant risk that the objective outputs of the model could be filtered through the biased lens of a human user, or worse, used to "tech-wash" decisions that remain biased [9]. Ensuring that this potential for unbiased performance translates into equitable outcomes requires more than just a validated algorithm; it requires robust implementation protocols, user training, and systems of accountability that address the entire decision-making pipeline, not just the tool itself [8].

Limitations

Our study has several limitations. First and foremost, our findings must be interpreted with significant caution due to the choice of the "Reading the Mind in the Eyes Test" (RMET) as our primary instrument. While widely used, the RMET's validity has been subject to increasing and substantial critique. Recent psychometric work raises serious concerns about the test's structural properties and ecological validity. Specifically, a large-scale analysis by Higgins et al. [39] found that RMET scores demonstrate poor structural properties, failing to conform to a unidimensional factor structure, which complicates the interpretation of a single sum score as a measure of a unified social-cognitive ability.

Furthermore, this critique extends to the very nature of the task. Higgins et al. [39] highlight evidence suggesting that performance on the RMET may rely less on direct emotion perception and more on other cognitive skills, such as vocabulary, abductive reasoning, and a process of elimination among the forced-choice options. This is supported by findings that individuals rarely generate the "correct" mental state term in free-response versions of the test. This constitutes a significant challenge to the RMET's ecological validity, as it suggests a fundamental gap between the test's demands and the dynamic, context-rich nature of real-world social cognition [40, 41]. Therefore, while our study provides a valuable assessment of MLLMs performance on this specific, constrained task, these profound methodological limitations of the RMET itself mean that our results cannot be readily generalized to the broader domain of real-world "mind-reading" or clinical empathy.

Second, we cannot entirely rule out the possibility that some RMET stimuli may overlap with the training data of the MLLMs. This consideration is critical when evaluating AI performance on standardized tests, as stimulus overlap could theoretically inflate performance metrics. However, our finding of consistent performance across versions suggests the models' capabilities extend beyond potential training data familiarity. Given documented imbalances in online facial datasets favoring White faces [42, 43], any training data bias would likely manifest as superior performance on the original RMET. The absence of ceiling effects and the comparably strong results on less widely published RMET variants makes it less likely that memorization alone accounts for the observed performance. Nevertheless, this remains an empirical question warranting further investigation. Third, our study was confined to three MLLMs; the rapid evolution of the field implies that a broader spectrum of models could yield more comprehensive insights. Additionally, given the exploratory nature of this study, the comparisons between the models’ performance rankings are primarily descriptive and should be interpreted as preliminary trends rather than definitive statistical superiorities. Fourth, the potential impact of more detailed prompting strategies, such as ‘chain of thought’ instructions, was not examined.

Finally, our focus was narrow. To attain a fuller understanding of MLLMs' capabilities, future studies should include a wider array of tasks, such as those involving voice, body movement, and other non-verbal cues. Furthermore, while we explored ethnic variations, our stimuli's cultural diversity was limited, highlighting the need to distinguish between cross-ethnic and cross-cultural generalization and include a broader range of ethnicities in future work.

Potential implications and future research directions

The findings of this study, while preliminary, carry several potential implications for the future development and evaluation of MLLMs. The demonstration of high-accuracy, cross-ethnically consistent performance by ChatGPT-4o provides a valuable benchmark for fairness in AI, suggesting that it is technologically feasible to mitigate certain human-like biases. Future research should investigate the architectural or training data differences that contribute to this robustness.

Furthermore, while acknowledging the significant limitations of the RMET, it is worth considering the potential long-term trajectory of this technology for fields like psychology and psychiatry. Should future research first establish the validity of these capabilities using more ecologically valid, dynamic assessments [43], advanced MLLMs might one day serve as valuable tools to augment human judgment. For instance, such validated models could potentially assist in identifying subtle social-cognitive deficits [44, 45] or provide an objective baseline for monitoring treatment efficacy [46]. However, it is critical to stress that this remains a distant prospect, contingent on extensive further research and validation.

Building on this work, we propose several additional directions for future research. The most critical next step is to move beyond static images and evaluate MLLMs using video-based tasks that better simulate real-world social interactions [43]. Additionally, studies should systematically analyze error patterns to understand which specific mental states are most frequently misidentified by these models. A particularly valuable direction would be to collect data that includes the specific incorrect foils chosen by the models, which would enable the construction of confusion matrices to reveal systematic error patterns (e.g., whether "pensive" is consistently misread as "anxious").

Conclusion

In conclusion, this study provides the first cross-ethnic evaluation of advanced MLLMs on a standardized emotion recognition task. Our findings indicate that a leading model, ChatGPT-4o, can achieve performance significantly above the human average with consistency across diverse ethnic stimuli. This marks a significant advancement in benchmarking MLLM performance on established social cognition assessments, highlighting both the rapid pace of technological advancement and the potential for these models to process complex social stimuli within a constrained, standardized context without exhibiting certain human biases. The pronounced performance gaps between models and the inherent limitations of the assessment tool itself call for a cautious interpretation of these findings. This work underscores the critical importance of rigorous, multi-faceted, and cross-demographic validation to ensure the responsible and equitable development of these advanced technologies.

Methods

Study design

This study employed a comparative design to evaluate the emotion recognition abilities of three leading MLLMs. The models for this study were selected based on several criteria aligned with our psychological research questions. We chose three popular and widely used commercial MLLMs with image analysis capabilities that were publicly available during the research period (ChatGPT-4, ChatGPT-4o, and Claude 3 Opus). Our primary goal was not to compare their technical architecture, but rather to investigate a psychological phenomenon: whether these common, accessible AI systems could recognize complex emotions and if they exhibited human-like biases, such as the 'other-race effect'. The inclusion of multiple models from different developers allowed us to examine the consistency of this phenomenon across different systems and to explore potential performance differences. Access to the MLLMs was obtained manually via their publicly available user interfaces during the study period. The study's hypotheses were not preregistered but were formulated based on the existing literature on AI's emerging psychological capabilities and documented cross-ethnic biases in computational models.

Measures

We assessed emotion recognition abilities using three well-established measures:

Reading the Mind in the Eyes Test [17]. The RMET consists of 36 photographs depicting eye regions of faces of White individuals expressing various mental states. Participants are asked to select which of the four options best describes the mental state portrayed in each photograph. Scores range from 0 to 36, with higher scores indicating better emotion recognition abilities.

The Black version of the Reading the Mind in the Eyes Test [18]. The B-RMET is an ethnically adapted version of the original RMET, featuring 36 photographs of eye regions from faces of Black individuals. The scoring and administration procedures are identical to those of the original RMET.

The Korean version of the Reading the Mind in the Eyes Test [26] The K-RMET is an ethnically adapted version of the RMET, featuring 36 photographs of eye regions from faces from Korean individuals. The scoring and administration procedures are identical to those of the original RMET.

Each version demonstrates robust psychometric properties: the original RMET shows strong test-retest reliability (ICC = .833) [45], the K-RMET exhibits comparable reliability (ICC = 0.758, 95% CI: 0.462, 0.892) [26], and the B-RMET demonstrates equivalent psychometric characteristics with no significant difficulty differences from the original version (t(35) = 0.552, p = 0.584) [16].

Procedure

From March to June 2024, all MLLMs (ChatGPT-4, ChatGPT-4o, Claude 3 Opus) were administered each RMET version twice to establish test-retest reliability. Each version of the RMET was presented in a separate and independent session, ensuring that the MLLMs had no memory of prior tests or responses. The second administration was conducted following the completion of the first administration, again in a new and independent session. The tests were presented to the MLLMs in the same order as they appear in the original test materials and with the same prompts, and the correctness of the MLLMs' responses (correct/incorrect) was recorded for later analysis. After providing the general instructions, each of the 36 images was presented, followed by four response options. The MLLMs selected one option, and immediately afterward, the next question was presented.

Data analysis

For each MLLMs and each administration of the RMET (White, Black, Korean), we conducted a binomial test to assess whether performance (proportion correct out of 36 items) exceeded the fixed chance level of 25%. To determine whether this performance deviated significantly from chance-level we implemented a Bonferroni correction to adjust the p-value threshold across the 18 measures (3 models × 3 RMET versions × 2 administrations), thereby controlling for familywise error rate. A corrected p-value less than 0.05 was considered statistically significant. We supplemented this test of statistical significance with two effect-size metrics. First, we computed the difference from chance (observed proportion minus 0.25), which indicates the absolute magnitude of improvement above random guessing. Second, to facilitate comparison to commonly recognized guidelines, we calculated Cohen’s h [47], a measure appropriate for comparing single proportions to a known reference proportion.

To ensure the methodological validity of aggregating scores across the two administrations, we conducted a preliminary stability analysis. A pooled paired-samples t-test (N=9) revealed no significant performance differences between the first and second administrations ($t(8)=0.88$, $p=.407$), ruling out systematic learning or inconsistency effects. Furthermore, the Intraclass Correlation Coefficient (ICC) indicated excellent test-retest agreement (ICC=.899, $p<.001$). Given these results, all subsequent statistical analyses and comparisons to human norms were conducted using the averaged scores of the two administrations.

Then, we calculated the percentiles of the MLLMs' performance by using each model's mean score from their two evaluations and comparing these means to the well-established, published neurotypical human sample results of the different RMET versions. The White RME results were sourced from Baron-Cohen and colleagues [17]; the Black RME results were derived from Handley and colleagues [18]; and the East Korean RME results were derived from Koo and colleagues [26]. This methodology was also employed to examine whether the models' performance remained consistent across ethnic groups. Specifically, bias was indicated if one of the results was at or above the human average while the other results was below the human average.

Our two-stage analysis first establishes performance above chance through binomial testing and its effect size, then contextualizes performance relative to human normative samples through percentile rankings. This approach, while innovative in the context of MLLMs evaluation, builds upon methodological frameworks previously validated in psychological assessment research [19].

Finally, to further characterize the models' performance, we conducted an exploratory item difficulty analysis to identify which mental states were systematically easier or more difficult for the MLLMs to recognize.

Declarations

Ethics approval and consent to participate: Not applicable. This study evaluated the performance of publicly available Generative AI models and did not involve human participants, and therefore did not require ethics approval or consent to participate.

Consent for publication: Not applicable This study did not involve human participants.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article.

Clinical trial registration: Not applicable. This study is not a clinical trial and did not require registration.

Availability of data and materials: In line with open science principles, the full prompts, study materials, and raw data are publicly available at **http://osf.io/6rh8m**. A detailed description of the manual data collection procedure is also provided in the repository.

References

1. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl Acad. Sci. USA* **121**, e2405460121 (2024). https://doi.org/10.1073/pnas.2405460121
2. van Duijn, M. J. et al. Theory of mind in large language models: examining performance of 11 state-of-the-art models vs. children aged 7–10 on advanceagud tests. Preprint at https://doi.org/10.18653/v1/2023.conll-1.25 (2023).
3. Lee, J., Choi, Y., Song, M. & Park, S. ChatFive: enhancing user experience in Likert scale personality test through interactive conversation with LLM agents. *Proc. 6th ACM Conf. Conversational User Interfaces* 1–8 (2024). https://doi.org/10.1145/3640794.3665572
4. Hamalwa, G. D. Mind the (AI) gap: psychometric profiling of GPT models for bias exploration. (2024).
5. Zhang, Y, Zou, C., Lian, Z., Tiwari, P. & Qin, J. SarcasmBench: towards evaluating large language models on sarcasm understanding. *IEEE Trans. Affect. Comput.* (2025). https://doi.org/10.1109/taffc.2025.3604806

6. Hall, J. A., Harrigan, J. A. & Rosenthal, R. Nonverbal behavior in clinician–patient interaction. *Appl. Prev. Psychol.* **4**, 21–37 (1995). https://doi.org/10.1016/S0962-1849(05)80049-6

7. Tian, Y., Kanade, T. & Cohn, J. F. Facial expression recognition. In *Handbook of Face Recognition* 487–519 (Springer London, 2011). https://doi.org/10.1007/978-0-85729-932-1_19

8. Hofmann, V., Kalluri, P. R., Jurafsky, D. & King, S. AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024). https://doi.org/10.1038/s41586-024-07856-5

9. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024). https://doi.org/10.1016/S2589-7500(23)00225-X

10. Atzil-Slonim, D. et al. Therapists' empathic accuracy toward their clients' emotions. *J. Consult. Clin. Psychol.* **87**, 33 (2019). https://doi.org/10.1037/ccp0000354

11. Kashner, T. M. et al. Impact of structured clinical interviews on physicians' practices in community mental health settings. *Psychiatr. Serv.* **54**, 712–718 (2003). https://doi.org/10.1176/appi.ps.54.5.712

12. Mobbs, R., Makris, D. & Argyriou, V. Emotion recognition and generation: a comprehensive review of face, speech, and

text modalities. Preprint at https://doi.org/10.48550/arXiv.2502.06803 (2025).

13. Picard, R. W. *Affective Computing* (MIT Press, Cambridge, 2000).

14. Schlegel, K., Sommer, N. R. & Mortillaro, M. Large language models are proficient in solving and creating emotional intelligence tests. *Commun. Psychol.* **3**, 80 (2025). https://doi.org/10.1038/s44271-025-00258-x

15. Kramer, R. S. Comparing ChatGPT with human judgements of social traits from face photographs. *Comput. Hum. Behav. Artif. Hum.* **4**, 100156 (2025).

16. Nelson, B. et al. Evaluating the performance of large language models in identifying human facial emotions: GPT 4o, Gemini 2.0 Experimental, and Claude 3.5 Sonnet. Preprint at https://doi.org/10.31234/osf.io/pxq5h_v1 (2025).

17. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The 'Reading the Mind in the Eyes’ test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* **42**, 241–251 (2001). https://doi.org/10.1017/s0021963001006643

18. Handley, G., Kubota, J. T., Li, T. & Cloutier, J. Black ‘Reading the Mind in the Eyes’ task: the development of a task assessing mentalizing from black faces. *PLoS One* **14**,

e0221867 (2019). https://doi.org/10.1371/journal.pone.0221867

19. Elyoseph, Z. et al. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. *JMIR Ment. Health* **11**, e54369 (2024). https://doi.org/10.2196/54369

20. Scherer, K. R., Clark-Polner, E. & Mortillaro, M. In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *Int. J. Psychol.* **46**, 401–435 (2011). https://doi.org/10.1080/00207594.2011.626049

21. Yan, X., Andrews, T. J., Jenkins, R. & Young, A. W. Cross-cultural differences and similarities underlying other-race effects for facial identity and expression. *Q. J. Exp. Psychol.* (2016). https://doi.org/10.1080/17470218.2016.1146312

22. Flade, F. & Imhoff, R. Closing a conceptual gap in race perception research: a functional integration of the other-race face recognition and “who said what?” paradigms. *J. Pers. Soc. Psychol.* **127**, 1 (2024). https://doi.org/10.1037/pspa0000388

23. Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y. & Elyoseph, Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz’s theory of basic values. *JMIR Ment. Health* **11**, e55988 (2024). https://doi.org/10.2196/55988

24. Hadar-Shoval, D., Asraf, K., Shinan-Altman, S., Elyoseph, Z. & Levkovich, I. Embedded values-like shape ethical reasoning of large language models on primary care ethical dilemmas. *Heliyon* **10**, e38056 (2024). https://doi.org/10.1016/j.heliyon.2024.e38056

25. Fiske, A., Henningsen, P. & Buyx, A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J. Med. Internet Res.* **21**, e13216 (2019). https://doi.org/10.2196/13216

26. Koo, S. J. et al. "Reading the mind in the eyes test": translated and Korean versions. *Psychiatry Investig.* **18**, 295 (2021). https://doi.org/10.30773/pi.2020.0289

27. Kramer, R. S. Comparing ChatGPT with human judgements of social traits from face photographs. *Comput. Hum. Behav. Artif. Hum.* **4**, 100156 (2025).

28. OpenAI. GPT-4. https://openai.com/product/gpt-4 (2023).

29. OpenAI. ChatGPT-4o. https://openai.com/index/hello-gpt-4o/ (2024).

30. Anthropic. Claude AI. https://www.anthropic.com/claude (2023).

31. Jiao, J. & Chang, A. Evaluating sentiment and spatial patterns of EV charging station user experience with AI-

agents. *Int. J. Urban Sci.* 1–29 (2025). https://doi.org/10.1080/12265934.2025.2547792

32. Fisher, J. et al. Political neutrality in AI is impossible – but here is how to approximate it. Preprint at arXiv:2503.05728 (2025).

33. Huang, S. et al. Collective constitutional AI: aligning a language model with public input. *Proc. 2024 ACM Conf. Fairness, Accountab. Transparency* 1395–1417 (2024). https://doi.org/10.1145/3630106.3658979

34. Nguyen, C., Carrion, D. & Badawy, M. K. Comparative performance of Claude and GPT models in basic radiological imaging tasks. *MedRxiv* 2024-11 (2024). https://doi.org/10.1101/2024.11.16.24317414

35. Atreides, K. & Kelley, D. J. Cognitive biases in natural language: automatically detecting, differentiating, and measuring bias in text. *Cogn. Syst. Res.* **88**, 101304 (2024). https://doi.org/10.2139/ssrn.4568851

36. Luczak, A. How artificial intelligence reduces human bias in diagnostics? *AIMS Bioeng.* **12**, 69–89 (2025). https://doi.org/10.3934/bioeng.2025004

37. Gupta, M., Virostko, J. & Kaufmann, C. Large language models in radiology: fluctuating performance and decreasing discordance over time. *Eur. J. Radiol.* **182**, 111842 (2025). https://doi.org/10.1016/j.ejrad.2024.111842

38. Kocak, B. et al. Radiology AI and sustainability paradox: environmental, economic, and social dimensions. *Insights Imaging* **16**, 88 (2025). https://doi.org/10.1186/s13244-025-01962-2

39. Higgins, W. C., Kaplan, D. M., Deschrijver, E. & Ross, R. M. Why most research based on the Reading the Mind in the Eyes Test is unsubstantiated and uninterpretable: a response to Murphy and Hall (2024). *Clin. Psychol. Rev.* **115**, 102530 (2025). https://doi.org/10.1016/j.cpr.2024.102530

40. Cuff, B. M., Brown, S. J., Taylor, L. & Howat, D. J. Empathy: a review of the concept. *Emotion Rev.* **8**, 144–153 (2016). https://doi.org/10.1177/1754073914558466

41. Yager, J., Kay, J. & Kelsay, K. Clinicians' cognitive and affective biases and the practice of psychotherapy. *Am. J. Psychother.* **74**, 119–126 (2021).

42. Sumsion, A., Torrie, S., Lee, D. J. & Sun, Z. Surveying racial bias in facial recognition: balancing datasets and algorithmic enhancements. *Electronics* **13**, 2317 (2024). https://doi.org/10.3390/electronics13122317

43. Refoua, E. et al. The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli. *Comput. Hum. Behav. Rep.* 100702 (2025). https://doi.org/10.1016/j.chbr.2025.100702

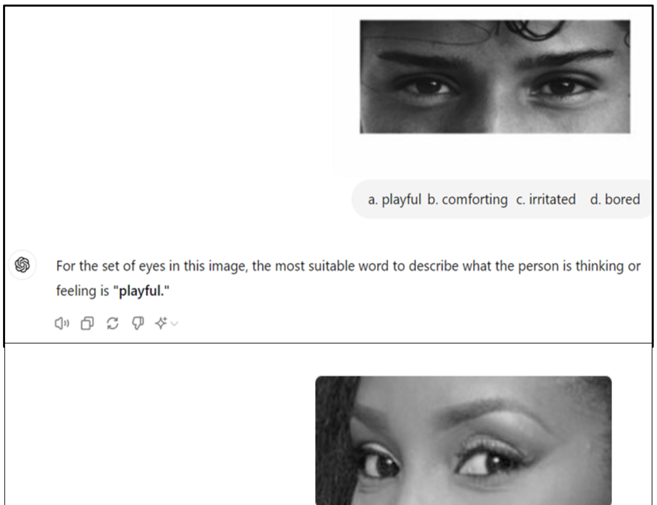
44. Konstantin, G. E., Nordgaard, J. & Henriksen, M. G. Methodological issues in social cognition research in autism spectrum disorder and schizophrenia spectrum disorder: a systematic review. *Psychol. Med.* **53**, 3281–3292 (2023). https://doi.org/10.1017/S0033291723001095

45. Vellante, M. et al. The ‘Reading the Mind in the Eyes' test: systematic review of psychometric properties and a validation study in Italy. *Cogn. Neuropsychiatry* **18**, 326–354 (2013). https://doi.org/10.1080/13546805.2012.721728

46. Hamdoun, S., Monteleone, R., Bookman, T. & Michael, K. AI-based and digital mental health apps: balancing need and risk. *IEEE Technol. Soc. Mag.* **42**, 25–36 (2023). https://doi.org/10.1109/MTS.2023.3241309

47. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013). https://doi.org/10.4324/9780203771587

Tables and Figure



a. playful b. comforting c. irritated d. bored

For the set of eyes in this image, the most suitable word to describe what the person is thinking or feeling is **"playful."**

Figure 1 | An illustration of the ChatGPT-4o interface for the RMET.

All MLLMs were prompted with the original test prompt: “*For each set of eyes, choose and circle which word best describes what the person in the picture is thinking or feeling. You may feel that more than one word is applicable but please choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words. You should try to do the task as quickly as possible, but you will not be timed. If you really don’t know what a word means you can look it up in the definition handout".* Abbreviations: MLLM, multimodal large language models.

Table 1 | Performance comparison and statistical analysis of MLLM models on emotion recognition tasks.

LLMs	RMET First Administration	RMET Second Administration	B-RMET First Administration	B-RMET Second Administration	K-REMT First Administration	K-REMT Second Administration
ChatGPT-4	*18 (h = 0.524, RD = 0.25)	**25 (h = 0.923, RD = 0.444)	**25 (h = 0.923, RD = 0.444)	**25 (h = 0.923, RD = 0.444)	**25 (h = 0.923, RD = 0.444)	**22 (h = 0.748, RD = 0.361)
ChatGPT-4o	**30 (h = 1.253, RD = 0.583)	**30 (h = 1.253, RD = 0.583)	**34 (h = 1.619, RD = 0.694)	**34 (h = 1.619, RD = 0.694)	**32 (h = 1.415, RD = 0.639)	**30 (h = 1.253, RD = 0.583)
Claude 3 Opus	15 (h = 0.356, RD = 0.167)	11 (h = 0.124, RD = 0.056)	*19 (h = 0.579, RD = 0.278)	15 (h = 0.356, RD = 0.167)	17 (h = 0.468, RD = 0.222)	14 (h = 0.3, RD = 0.139)
Human Sample	26.2 ± 3.6 (Baron-Cohen and colleagues[17])		25.92±5.19 (Handley and colleagues[18])		26.72 ± 3.38 (Koo and colleagues[24])	

Statistical significance levels: *p < .05, **p < .001. Effect sizes are reported using Cohen's h (h) and Risk Difference (RD). All p-values are Bonferroni-corrected for multiple comparisons. The Standard Deviations (SD) reported represent between-subject variability in the human normative samples, as published in the original validation studies. All tests consist of 36 items. The tested models include ChatGPT driven by GPT-4 (ChatGPT-4), ChatGPT driven by GPT-4o (ChatGPT-4o), and Claude 3 Opus. Abbreviations: RMET = Reading the Mind in the Eyes Test; B-RMET = Black Reading the Mind in the Eyes Test; K-REMT = Korean Reading the Mind in the Eyes Test.

Table 2 | Performance of MLLMs on RMET versions presented as percentiles of human norms.

Model	RMET	B-RMET	K-RMET
ChatGPT-4o	85.00	94.00	90.00
ChatGPT-4	9.60	43.00	17.00
Claude 3 Opus	0.01	4.00	0.04

The table displays the performance of each model as a percentile rank compared to human normative data for each version of the test. RMET = Reading the Mind in the Eyes Test (White version); B-RMET = Black version of the RMET; K-RMET = Korean version of the RMET.

Table 3 | Item difficulty analysis for MLLMs.

Most Difficult Items	Aggregated Error Rate (%)	Claude 3 Opus (%)	ChatGPT-4 (%)	ChatGPT-4o (%)
Nervous	83.3	100.0	100.0	66.7
Fantasizing	61.1	91.7	91.7	0.0
Playful	61.1	100.0	66.7	16.7
Friendly	55.6	100.0	66.7	16.7
Insisting	55.6	100.0	33.3	33.3
Flirtatious	50.0	83.3	50.0	16.7
Defiant	50.0	16.7	83.3	50.0
Distrustful	50.0	100.0	16.7	33.3
Interested	50.0	66.7	50.0	33.3
Cautious	44.4	58.3	50.0	25.0
Least Difficult Items	**Aggregated Error Rate (%)**	**Claude 3 Opus (%)**	**ChatGPT-4 (%)**	**ChatGPT-4o (%)**
Contemplative	0.0	0.0	0.0	0.0
Uneasy	0.0	0.0	0.0	0.0
Worried	0.0	0.0	0.0	0.0
Concerned	11.1	33.3	0.0	0.0
Doubtful	11.1	33.3	0.0	0.0
Suspicious	11.1	33.3	0.0	0.0
Reflective	16.7	50.0	0.0	0.0
Serious	16.7	50.0	0.0	0.0
Pensive	16.7	50.0	0.0	0.0
Despondent	22.2	0.0	50.0	16.7

The table displays the error rates for the ten most difficult and ten least difficult items, aggregated and broken down by each MLLM.