Reading Between the Lines:

LLMs Match or Exceed Human Empathic Accuracy Using Text Alone

Noa Oded*[1], Matan Rubin*[1], Shir Genzer[1], Anat Perry[1]

The Hebrew University of Jerusalem

* Equal contribution and joint first-authorship

Author Note

Noa Oded (ORCID 0009-0004-7139-1296)

Matan Rubin (ORCID 0009-0008-7869-4993)

Shir Genzer (ORCID 0000-0003-1633-7882)

Anat Perry (ORCID 0000-0003-2329-856X)

All data and code are available at OSF:

https://osf.io/qtmdr/?view_only=b7a208a91e4e4e14803ac7a1f3a4686b

Corresponding authors:

Matan Rubin or Anat Perry

Matan.rubin@mail.huji.ac.il

Anat.perry@mail.huji.ac.il

Keywords: Empathy, cognitive empathy, empathic accuracy, artificial intelligence, Large language models

Abstract

Empathy plays a central role in human emotional relationships. Empathic accuracy, the ability to accurately infer another person's emotional state, varies by informational modality and, in humans, is often intertwined with emotional and motivational processes. This study examines whether state-of-the-art Large Language Models (LLMs) - GPT-4, Claude, and Gemini - demonstrate empathic accuracy, and how their accuracy compares to that of humans when presented with only the semantic content (transcripts of recorded videos) of ecological, complex autobiographical emotional narratives. We compared the empathic accuracy of LLMs' to that of human participants (N = 127, randomly sampled students, both in-lab and online) who either read the same transcripts or watched the original videos, which enabled them to use facial and bodily expressions, as well as paralinguistic cues, in addition to semantics. LLMs were able to infer emotional states from semantic content alone with a precision that is equal to or surpasses human performance. This was true both generally and when analyzing positive and negative emotions separately. Theoretically, these findings suggest that semantic information alone can support high empathic accuracy, though humans may not fully leverage this potential. Practical implications are discussed regarding the use of LLMs in introspective and emotional contexts, while raising critical concerns about privacy, ethical risks, and the potential reshaping of emotional understanding, intimacy, and human connection in an increasingly AI-mediated world.

1. Introduction

Empathy, the ability to recognize, understand, and share the emotional states of others, is essential for human connection (Batson et al., 1991; Davis, 2017; Genzer et al., 2024; Zaki & Ochsner, 2012). While definitions vary, most scholars agree that empathy consists of three core components: Cognitive empathy, also known as perspective-taking, is the ability to understand another person's internal mental states, and frequently evaluated as an individual's accuracy in discerning another's thoughts and emotions; affective empathy refers to the capacity to share or "feel with" others' emotions; and motivational empathy, or "empathic concern" is a drive to act on behalf of the other's well-being (Genzer et al., 2024; Zaki & Ochsner, 2012).

Research has extensively demonstrated the broad-ranging benefits of human empathy across multiple contexts. Empathy forges social bonds and sustains close relationships (Anderson & Keltner, 2002), positively impacts adolescent peer relationships (Gleason et al., 2009), increases marital satisfaction (Rafaeli et al., 2017; Sened et al., 2017), enhances the effectiveness of medical care (Larson, 2005; Rakel et al., 2009), and improves psychotherapy outcomes (Elliott et al., 2018). Given the myriad of benefits for individual wellbeing and social cohesion, understanding the factors that enable and enhance empathy is essential.

Alongside these benefits, empathy is also taxing for people and frequently avoided (Cameron et al., 2019). This fact has coincided with recent developments in the field of artificial intelligence (AI), leading many to seek emotional and empathic support from AI agents, especially given their accessibility and availability (Alabed et al., 2024; Haensch, 2025; Li & Zhang, 2024). However, there are fundamental differences, both theoretical and practical, between empathy in human relationships and empathy in human–AI relationships.

First, in humans, cognitive, affective, and motivational empathy are intertwined and often align: We tend to care more for those close to us, understand them better, and are more motivated to help them (van den Bedem et al., 2019). This interconnection is further evidenced in vicarious-pain responders, who not only experience others' pain as their own but consequently demonstrate enhanced emotional understanding and stronger helping motivations (Ben Adiva et al., 2024). Furthermore, neuroscience studies on empathic accuracy—a paradigm that examines cognitive empathy—have

revealed that successful perspective-taking engages both cognitive and affective neural networks simultaneously, suggesting that these components operate in concert rather than isolation (Genzer et al., 2022; Schurz et al., 2021; Zaki & Ochsner, 2012). As such, although we can distinguish between the components of empathy theoretically, disentangling them from one another and examining only a single aspect, in ecological contexts where they often co-occur, remains a significant challenge.

To address this challenge, researchers frequently study clinical populations. These include lesion studies (e.g. Perry et al., 2017; Shamay-Tsoory et al., 2004) or individuals with disorders that theoretically affect different aspects of empathy, such as autism, schizophrenia, or psychopathy (Derntl et al., 2012; Keysers & Gazzola, 2014; Song et al., 2019). However, these studies have not shown clear deficits in any one specific aspect of empathy, highlighting how interconnected these aspects are in humans and the difficulty of isolating them in order to clearly study one without the others.

In contrast, AI models, and specifically large language models (LLMs), do not integrate cognitive, affective, and motivational components of empathy. While they are capable of inferring emotions and simulating aspects of cognitive empathy, they lack consciousness, and they do not experience emotions, "feel with" others, or care about others' well-being (Perry, 2023). This difference is experienced by users, with recent research showing that the affective and motivational components of empathy are more valuable to individuals when perceived as human-authored as opposed to AI-generated, without such differences apparent in cognitive empathy (Rubin et al., 2025).

This inherent distinction of aspects of empathy in LLMs presents a unique scientific opportunity: Because LLMs operate without affective or motivational components, they allow researchers to isolate and examine the informational basis of cognitive empathy in a way that is not possible in human studies.

1.1 Communication Channels and the Role of Semantics

Another critical distinction between human–human and human–AI interactions lies in the channels of communication. Human emotional understanding typically relies on multimodal information: verbal content, facial expressions, tone of voice, body language, and paralinguistic cues such as pitch and rhythm (Gunes et al., 2008). In

contrast, as of today, most daily communication with LLMs is limited to semantic textual information (Wang et al., 2024).

Research attempting to dissect the relative contributions of these channels in humans has occasionally used Empathic Accuracy tasks, in which participants observe a target recounting an emotional experience—through video, audio, or text—and are asked to continuously or retrospectively infer the target's emotions. Accuracy is then measured by comparing participants' inferences to the target's own self-reported emotional states. These studies find that while people can identify emotions based on visual cues alone, the combination of auditory semantic information and paralinguistic vocal cues typically yields the highest empathic accuracy (Genzer et al., 2022; Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Jospe et al., 2020; Kraus, 2017; Kraus & Segal, 2015). However, the specific informational value of pure semantic content remains largely unexplored.

In a study directly addressing this, Hall and Schmid Mast (2007) demonstrated that participants reading only transcripts (thus avoiding any paralinguistic or visual cues) achieve relatively high empathic accuracy, though emotional inferences were further improved when paralinguistic cues were available. These results imply that semantic information carries substantial emotional content, but that paralinguistic cues contribute additional information and increase empathic accuracy.

As with the distinction of empathy components, here too the use of LLMs provides an opportunity. Since LLMs operate exclusively through semantic input, they offer a unique methodological solution to this challenge. By utilizing and examining the empathic accuracy of systems that process and generate purely textual information, researchers can more precisely isolate and evaluate the informational contribution of semantics to empathic inference, free from the confounding effects of additional communicative channels.

1.2 Theoretical and Practical Questions

Considering LLMs' lack of emotional experience and their reliance on textual input alone—a central question emerges: *To what extent can LLMs accurately infer human emotional states based purely on semantic information?* The comparison

between humans' and LLMs' empathic accuracy in a naturalistic test, will both show the ability of LLMs to infer emotional states in the present moment, and also presents a unique opportunity to test two fundamental questions in empathy research: First, can LLM's achieve high levels of empathic accuracy compared to humans, without any affective experience or motivational concern? Second, is semantic information alone sufficient for accurate emotional understanding?

Answering these questions will yield valuable insights, and will have important implications, both theoretically and practically. From a *theoretical perspective*, understanding LLMs' and humans' empathic accuracy abilities from semantic information enables researchers to pose fundamental questions about the informational and inferential foundations of cognitive empathy, and to illuminate the distinct contributions of semantic processing in empathic understanding. *Practically*, the findings can inform the development of potential applications of LLMs in the realm of human–AI relationships, including current use cases, their possible value, and important risks that should be taken into account.

1.3 Present Research

Recent studies have begun to explore the capacity of LLMs to understand emotional states, however, accuracy in these studies was evaluated by third-party raters, and not compared to the actual rated emotions of the target (Gandhi et al., 2024; Lee et al., 2024; D. Ong et al., 2022, Tak & Gratch, 2024). Other studies have assessed LLM's accuracy using structured tasks, such as the Reading the Mind in the Eyes Test (RMET), the Movie for the Assessment of Social Cognition (MASC), or vignette-based measures, like the Situational Test of Emotion Management (STEM) or the Geneva EMOtion Knowledge Test—Blends (GEMOK-B). These studies found that LLMs performed as well as humans or outperformed them (Refoua et al., 2024, 2025, Schlegel et al., 2025). However, the stimuli in these tasks are fabricated as opposed to naturalistic or ecological in nature. Moreover, high performance on these tasks may be partially attributable to some correct answers appearing openly in research (See for examples: Baron-Cohen, 2003; MacCann & Roberts, 2008; Schlegel & Scherer, 2018). This potentially makes them part of the LLM training data, thus influencing the answers and preventing a pure analysis of the emotions present in the stimuli. To our knowledge, the only study that specifically measured the empathic accuracy of an LLM in a naturalistic

setting is Yin et al. (2024); but notably, it focused exclusively on negative scenarios in an English-speaking, western culture, examined only a single LLM (Bing Chat), and did not compare the model's performance to that of humans with access to the full video context.

The current study examines LLMs capacity to understand emotional states through an empathic accuracy task using naturalistic videos of autobiographical emotional narratives told by Israeli participants in Hebrew, which included both positive and negative emotions. These videos are part of a strictly-secured dataset, used for experiments on secure platforms and only with explicit permission. This approach ensures that the LLMs had no prior exposure to these specific stimuli (see Methods below). Moreover, the videos' language introduces an additional challenge for LLMs, as the LLMs evaluated were primarily trained on English-language data, heavily influenced by American and other Western cultural contexts (Dey et al., 2024; Roumeliotis & Tselikas, 2023). Therefore, poor performance on this task would not necessarily reflect limitations in the LLMs' empathic accuracy capacities, but may instead reflect a cultural or linguistic gap between their training data and the current stimuli. Alternatively, such findings would support the proposition that semantic information is not enough for empathic accuracy. However, high performance—even on these culturally specific, Hebrew-language personal stories—would point to remarkably robust empathic accuracy abilities, achieved purely from semantic information. This would be especially compelling if the LLMs were to match or even surpass the performance of native Hebrew-speaking Israeli participants, who share the speakers' cultural background and have access to the full videos that include additional sensory information cues, such as tone of voice and facial expressions.

To address these questions, we conducted a study using stimuli from the Israeli Empathic Accuracy Stimuli Set (Jospe et al., 2020). In this dataset, each storyteller rated their own emotional states during the recording, allowing for an objective benchmark against which to compare the empathic accuracy of both humans and LLMs. In our design, we sampled two groups of human participants: One group read the transcripts, and the other watched the full videos, including all semantic, visual, and paralinguistic information. We then gathered ratings from LLMs based only on the textual transcripts of the stories, without any visual or paralinguistic information, and assessed their ability

to infer emotions based solely on this semantic information. We then compared the empathic accuracy of LLMs to that of the human raters.

Because previous research has shown that LLMs can differ in their capacity for empathic responses (Lee et al., 2024; Yongsatianchotet al., 2024), we compared three LLMs alongside the human groups. We systematically examined whether LLMs might extract emotional cues from language more effectively than humans typically do, without a corresponding emotional experience and paralinguistic information.

Lastly, since humans are characterized by a negativity bias, meaning they tend to notice, interpret, and remember negative emotional information more readily than positive information (Norris, 2021; Rozin & Royzman, 2001; Vaish et al., 2008), we further examine whether AI models, trained on human data, perform similar to (or differently from) humans when inferring negative versus positive emotions.

2. Methods

2.1 Transparency and Openness

To protect the privacy of participants who shared autobiographical emotional experiences used as stimuli, the experimental stimuli are not publicly provided, but are available for academic use, and will be sent upon request from the corresponding authors. Data were analyzed using R, Version 4.3.2 (R Core Team, 2023) All studies were approved by the ethics committee of the social sciences faculty of the Hebrew University of Jerusalem. The study design and analysis were not pre-registered, as we had no clear hypotheses ahead of time. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

2.2 Participants

Based on a previously found effect size (Hall & Schmid Mast, 2007, r = .33), we ran a power analysis using the 'pwr' R package to be able to detect a difference between the two human conditions, should one exist. Though this design is not identical to ours, it was the closest we could find, and differences between the LLMs themselves were not the main interest of our paper. We found that a sample of 60 participants in

each group would be highly powered to detect an effect between the two human conditions (power = 0.9).

We recruited 128 Hebrew-speaking undergraduate students from the Hebrew University of Jerusalem through the university's participant recruitment platform. The students received either course credit or 60 NIS (~16 US dollars) in the video condition, or 30 NIS (~8 US dollars) in the text condition, as the latter required less time from the participants. We excluded one participant from the analysis who failed more than 3 attention checks, resulting in a final sample of 127 (n_{text} = 57, 63% female, M_{age} = 25.08 ± 4.15 (SD)).

2.3 Empathic Accuracy Stimuli

We selected 12 videos from the Israeli Empathic Accuracy Stimuli Set (Jospe et al., 2020). These videos featured 10 different storytellers (i.e. targets, 60% female M_{age} = 22.75 ± 1.98 (*SD*)) who shared personal emotional experiences in Hebrew. The videos were mixed in valence, with 4 videos depicting mostly positive emotions, 3 mostly negative emotions and 5 including mixed emotional content. The average duration of the videos was 146.92 seconds (*SD* = 42.03), with a minimum duration of 87 seconds and a maximum of 240 seconds. Once the targets finished narrating all stories, they watched the videos and continuously rated the valence of their emotions in each moment of the video. They were then asked to report on the intensity of eight specific emotions they had felt during their storytelling as a whole, using a scale from 1 (*"not at all"*) to 9 (*"very much"*). The specific emotions included embarrassment, anger, sadness, happiness, disgust, pride, fear, and excitement. We converted the set of videos into written text in order to prompt the LLMs and present them to human participants in the text condition. All stimuli used in the study were not publicly available online, ensuring that LLMs had no prior exposure to them and no opportunity to learn them in advance.

2.4 Procedure

After providing informed consent, participants first were shown the stimuli. In the video condition, participants came into the lab and viewed the videos in a random order, in two blocks of six videos each. A 10-minute break was offered between blocks.

While viewing each video, participants continuously assessed the emotional valence of the narrator, indicating the degree to which the emotions conveyed were positive or negative by moving a cursor along a dynamic scale ranging between 0 (“negative”) and 100 (“positive”) displayed beneath the video. This measurement is not relevant for evaluating text and so was not analyzed further in this research (it is part of a different, ongoing study in the lab, beyond the scope of the current work). At the end of each video, participants rated the intensity of the emotions they perceived the target felt. These were the same specific emotions that the target had rated after the narration of their stories, using a nine-point scale (on a scale of 1 = *not at all* to 9 = *very much*). Additionally, participants answered questions about their familiarity with the narrator and any technical disruptions during video playback. They were then asked one multiple-choice question regarding the content of the story. This was used as an attention check, and all trials where participants failed to answer correctly were removed. Participants who failed more than three of these questions were removed from the analysis. They then completed several other questionnaires, not examined in the current research.

In the text condition, participants took part in an online study in which they read the transcriptions of the same videos presented in the video condition. After each story, they rated the intensity of the emotions experienced by the target while telling the story and answered the same attention-check questions as in the video condition and the same subsequent questionnaires. Trials where participants failed to answer the attention check were removed.

2.4.1 AI Models

Parallel to the human participants, we used three LLMs to gather similar empathic accuracy ratings from AI. These were GPT-4o-2024-08-06, Claude-3.5-sonnet-20240620 and Gemini 1.5 Pro, which were state-of-the-art at the time we generated responses, between October and December 2024. These models were utilized to process the same series of 12 stories in a random order. Each model was tasked with analyzing the emotional content of the stories and providing ratings for the intensity of emotions, using the same scale as the human participants did. The AI models also processed the attention-check question concerning the content of each story. There were no instances where this question was answered incorrectly. The exact prompts are

available in the supplementary section 1(translated into English). To make these iterations as parallel as possible to human participants, we added the full conversation history up until each story before the story itself, including the previous stories and ratings from that iteration. This was parallel to a human participant being asked to rate each story sequentially. We then treated each iteration of an LLM rating all 12 stories as a separate unique participant. It should be noted that this caused some noise, and it required an engineering process until we reached prompts that gave specific responses to each story, despite including the chat history.

In order to generate the ratings by each LLM, the prompts were delivered using an R script to each model through its respective API, which allows direct access to the model without specific third-party guidelines or restrictions. The code used the following packages: claudeR (Yamil Velez, 2024), openai (Iegor Rudnytskyi, 2023), gemini.R (Jinhwan Kim, 2024). All code and data can be found in our OSF project: https://osf.io/qtmdr/?view_only=b7a208a91e4e4e14803ac7a1f3a4686b

Since there is no accepted benchmark for how many iterations to run on LLMs, we opted for matching the number of human samples and ran 60 iterations for each model. All runs were conducted using the same version of the models and identical parameters, specifying a temperature setting of 1 for all models, which is presumed to be a common setting for mainstream platforms using these LLMs at the time of study design. It should be noted that previous research shows temperature does not significantly impact performance for various tasks (Patel et al., 2024; Renze, 2024; Windisch et al., 2024).

2.5 Measures:

2.5.1 Empathic accuracy

The empathic accuracy level of the perceiver was determined by calculating the absolute difference between the perceiver's rating and the storyteller's rating. This scale was then reversed, with 8 points for the highest accuracy and 0 points for the lowest. The total of these reversed scores indicates the overall accuracy in recognizing emotions, with a scale that ranges from 0 to 64. To make it easier to understand, this scale was converted to a range of 0 to 100 (for similar calculation, see Israelashvili et

al., 2020). We also calculated empathic accuracy separately for positive emotions (happiness, excitement, pride) and for negative emotions (embarrassment, sadness, disgust, anger, fear) in a similar manner.

2.6 Analysis:

All analyses were performed using R (R Core Team, 2023). Trials where participants or LLMs scored more than 2.5 standard deviations from the mean empathic accuracy score were excluded from the analysis, leading to the removal of 34 (4.16%) trials from the video condition, 34 (5.04%) trials from the text condition, 3 (.004% trials from GPT, 9 (.012%) trial from Claude, and 4 (.006%) trials from Gemini. No individual participants were removed.

First, we fitted a linear mixed model to predict empathic accuracy scores using the response source, which had five levels: GPT-4o, Claude-sonnet-3.5, Gemini pro 1.5, human participants in the video condition, and human participants in the text condition. The model maintained a random intercept for each participant/iteration and each story, using the lmer function from the lme4 package, with the model's contrasts dummy-coded, using the human participants' mean empathic accuracy in the video condition as the intercept. Thus, we compared all conditions to the most ecological human condition: human raters who received full audiovisual stimuli. We transformed the model effects to a type-III ANOVA to report here, with full model details available in supplementary section 2. We then analyzed specific post-hoc contrasts of interest, comparing the two human conditions to each other and to each AI model. These were conducted using the emmeans package, with Bonferroni corrections applied to adjust for multiple comparisons.

Second, we fitted a linear mixed-effect model that examined empathic accuracy from response source, the rated emotions' valence, and their interaction. Contrasts were dummy-coded, with the human participants' mean empathic accuracy in the video condition as an intercept for response source, and negative emotions as an intercept for the valence. The model maintained a random intercept for each participant/iteration and for each story. We again transformed the model to a type-III ANOVA for reporting purposes, with full details available in supplementary section 3. Post-hoc contrasts

compared the two different human rating conditions to the AI-generated ratings and used Bonferroni corrections to adjust for multiple comparisons.

Third, we explored the distribution of AI-generated responses and compared their statistical characteristics to those of human participants, to see if they behave similarly and are as varied. Specifically, we looked at four characteristics: we used Shapiro-wilk's test to examine the normality of the distribution of each rating source's empathic accuracies; Levene's test to compare the variances of the AI models to human raters; we obtained the percentage of repeated responses by response source; and we subsetted the human and AI data to investigate the variance explained by the random effect of each iteration compared to that explained by the random effect of ratings from the same human rater.

In this examination of the distribution, we found that there was limited variance and a large number of repeated responses in the AI-generated ratings. Aiming to better represent AI ratings, we decided to generate a much larger number of AI responses, and to treat them as a population. To assess whether the larger population of AI-generated responses was more or less varied than the initial sample, we compared the proportion of duplicate responses in each dataset.

Lastly, we repeated our analyses on empathic accuracy. We compared the two different human samples to this larger AI population to examine our hypotheses again while representing as wide a range of AI responses to each story as possible (within reason). We compared the human ratings to the mu of this new population of AI-generated ratings, for total EA, and specifically by valence. We also tested a linear model, predicting whether the differences between human ratings and the mode and mean of the AI ratings are significantly different from 0, while including a random effect for each specific story (Full details for all models are available in supplementary sections 4-7). We then also tested whether the mu of the new population is significantly different to human raters EA (generally and by valence) when averaging the EA of each individual participant across stimuli.

2.6 Data and code availability

All research materials including LLM instructions, anonymized data, and analyses file are available https://osf.io/qtmdr/?view_only=b7a208a91e4e4e14803ac7a1f3a4686b.

3. Results

3.1 Data Cleaning

We removed all observations in which participants failed the attention check (n = 10 in the video condition, n = 10 in the text condition), with no individual participants removed. We also removed observations in the video condition where participants did not complete the continuous rating tasks (n = 5). One participant was removed entirely as they had more than 4 trials removed due to these reasons. We then removed observations in the video condition where participants recognized the target (n = 9), and observations where participants indicated technical difficulties (n = 4 in the video condition), leading to a total of 1491 observations from 127 human participants. There were no incorrect attention checks in the AI models' responses to the text.

3.2 Differences in Empathic Accuracy

The linear mixed model revealed a significant effect of response source on empathic accuracy ($F_{(4, 264.63)} = 38.02$, $p < .001$, $\eta^2 = 0.36$, 95% CI = [.29, 1.00]). Post-hoc contrasts revealed that all LLMs showed significantly higher empathic accuracy than humans in either condition, but humans in the video condition were more accurate than those in the text condition (all *p*s <.001, see Table 1). These results indicate that AI models are more empathically accurate than humans who read transcripts or viewed the full audiovisual stimuli, despite the fact the models received only Hebrew transcripts of the videos.

Table 1. Comparisons of empathic accuracy abilities between different LLMs and human participants.

Contrast	Standardized mean difference	SE	df	*t*	*p*
GPT-4o − Human (Text)	.34	.04	302	7.71	<.001***
GPT-4o − Human (Video)	.19	.04	301	4.44	<.001***
Claude − Human (Text)	.45	.04	304	10.16	<.001***
Claude − Human (Video)	.30	.04	303	7.01	<.001***

Gemini − Human (Text)	.42	.04	302	9.57	<.001***
Gemini − Human (Video)	.27	.04	301	6.38	<.001***
Human (Text) − Human (Video)	-.15	.04	315	-3.59	.003**

Table 1 shows the differences in empathic accuracy between every two response sources in the comparison in standard deviations. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

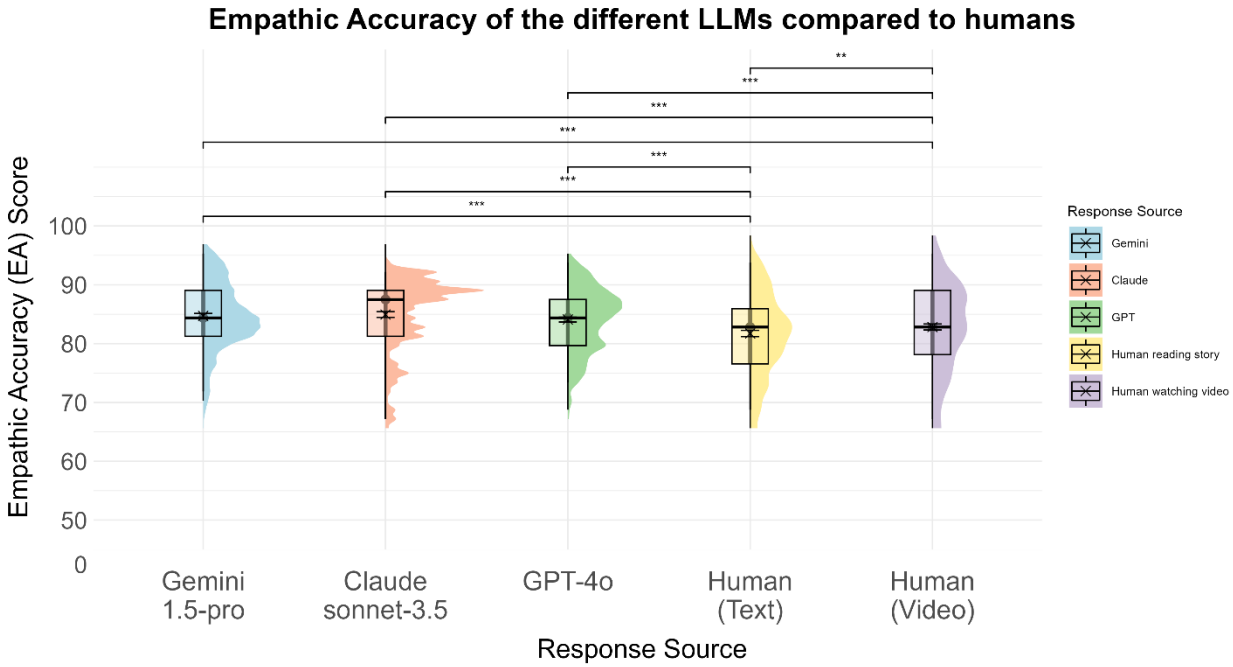

Figure 1: Differences in empathic accuracy between sources. For each source, the box marks the IQR, with the X inside marking the mean empathic accuracy with a 95% CI for the mean around it. The whiskers mark the total range, and the histogram to the side shows the distribution.

We then explored these differences for each valence of emotions. A linear mixed-effect model found a significant main effect of response source ($F_{(4, 7126)} = 21.56$, $p < .001$, $\eta^2_p = .01$, 95% CI = [.01, 1.00]), with no significant effect of emotion valence ($p = .53$), and a significant interaction of response source and emotion valence ($F_{(4, 7126)} = 24.32$, $p < .001$, $\eta^2_p = .01$, 95% CI = [.01, 1.00]). Post-hoc contrasts revealed that in all cases AI models were either similar to human raters or significantly better than them, with Claude and Gemini being more accurate than human raters for negative emotions, and GPT and Gemini being more accurate than human raters for positive emotions. Interestingly, human raters in the video condition were significantly more accurate than human raters in the text condition only for negative emotions, and not for positive emotions (see Table 2).

Table 2. Comparisons of empathic accuracy abilities between different LLMs and human participants' conditions for positive and negative emotions separately.

Valence	Contrast	Standardized mean difference	SE	df	t	*p*
Negative Emotions	GPT-4o − Human (Text)	.15	.04	1174	3.28	.008**
	GPT-4o − Human (Video)	-.04	.04	1143	-1.14	1.00
	Claude − Human (Text)	.43	.04	1176	9.85	<.001***
	Claude − Human (Video)	.24	.04	1145	5.83	<.001***
	Gemini − Human (Text)	.32	.04	1179	7.22	<.001***
	Gemini − Human (Video)	.13	.04	1149	3.05	.02*
	Human (Text) − Human (Video)	-.19	.04	1223	-4.45	<.001***
Positive Emotions	GPT-4o − Human (Text)	.22	.04	1142	5.53	<.001***
	GPT-4o − Human (Video)	.31	.04	1166	7.33	<.001***
	Claude − Human (Text)	-.02	.04	1148	-.37	1.00
	Claude − Human (Video)	.07	.04	1173	1.66	0.68
	Gemini − Human (Text)	.13	.04	1141	2.94	.02*
	Gemini − Human (Video)	.21	.04	1165	5.13	<.001***
	Human (Text) − Human (Video)	.08	.04	1207	2.00	0.32

Table 2 shows the differences in empathic accuracy between every two response sources in the comparison in standard deviations. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

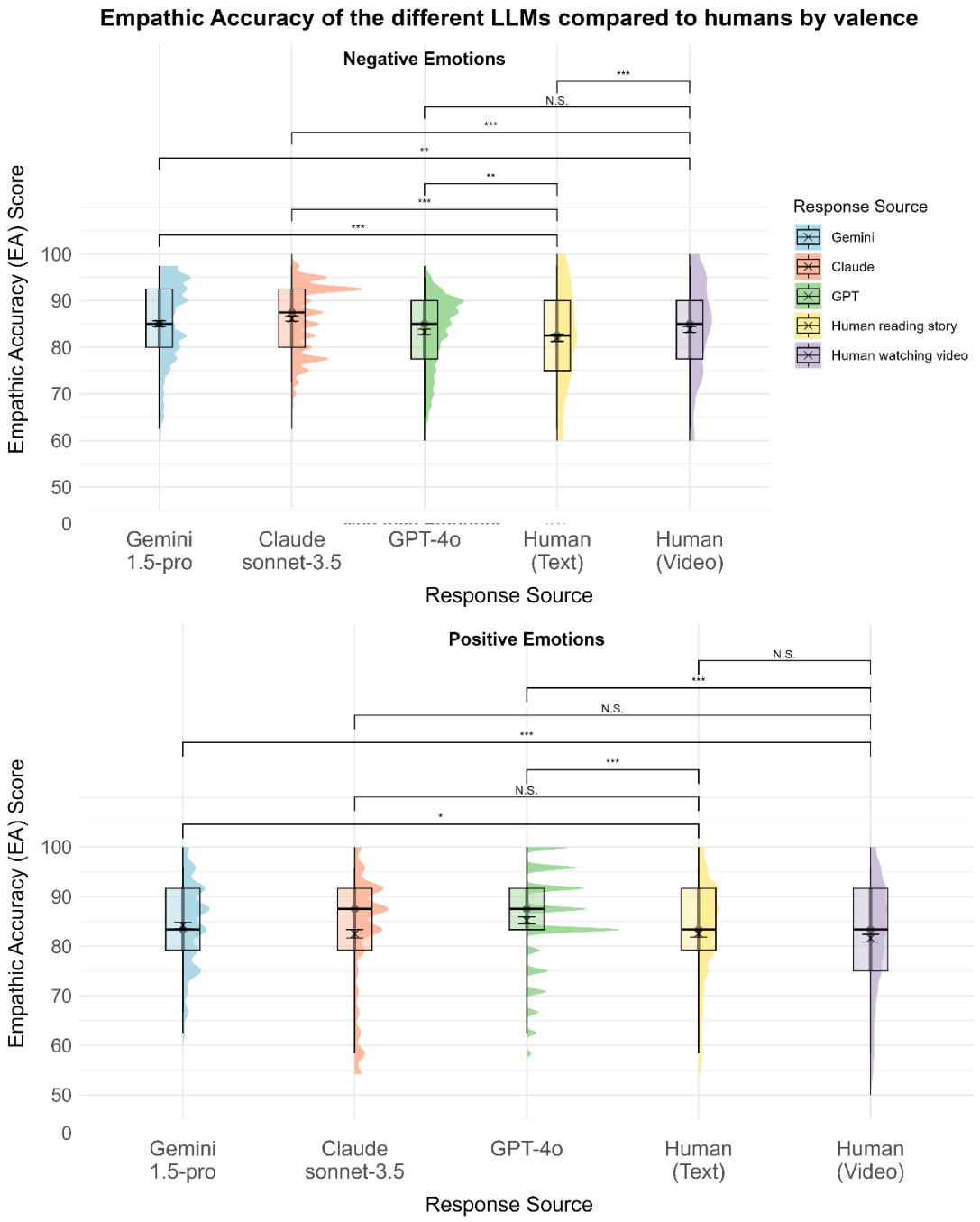

Figure 2: Differences in empathic accuracy between sources for positive and negative emotions. For each source, the box marks the IQR, with the X inside marking the mean empathic accuracy with a 95% CI for the mean around it. The whiskers mark the total range, and the histogram to the side shows the distribution.

3.3 Comparing Human and AI-generated Data Characteristics

In examining the differences between human and AI-generated ratings, we found that all EA ratings were not in a normal distribution (see Table 3), and Levene's test revealed they significantly differed from each other in variance ($F_{(4, 3549)} = 19.82$, $p < .001$). We

also documented two key differences between the AI-generated ratings and the human ones. First, we found that when subsetting the AI and the human data separately, a linear mixed model that predicts only the AI-generated ratings estimated the variance for the random effect of individual LLM iteration at 0 (τ_{ID} = .00), showing no evidence for explaining any variance by assuming dependence within iterations. This was very different from the estimated variance for the random intercept of individual human participants when an identical model instead based its predictions on human ratings (τ_{ID} = .11). We used an ANOVA to compare each subsetted model with the random intercept for specific ID or iteration, and without it. The comparison showed that the model predicting the AI data was not improved by the inclusion of the random intercept for iteration ($\chi^2_{(1)}$ = .00, p = 1.00). However, the model predicting human data was significantly improved ($\chi^2_{(1)}$ = 33.28, p < .001). Second, the AI data included a large number of duplicates, with different iterations of AI ratings of each story being repeated more than once: 43.47% in Gemini, 21.34% in GPT, 63.84% in Claude. This was far more frequent than in ratings provided by different human participants (0.005% in the video condition, 0.007% in the text condition). This reveals, perhaps not surprisingly, that the variance in LLM responses is not similar to that of human raters, nor do AI-generated empathic accuracy scores follow a normal distribution (see Figures 1-2 above). However, the large number of duplicate responses in the AI-generated ratings raises the possibility that our sample of AI responses captured only a subset of possible responses, with very different variability to that of the human raters, and thus our comparison may not be the best one. It is possible that a larger sample of responses, representing a population of AI-generated ratings would provide a better comparison to human ratings and a more detailed account of the differences (see below).

Table 3. Results of Shapiro-Wilk tests for the normality of Empathic Accuracy scores by response source.

Model	W	*p*
GPT-4o	0.97	<.001***
Claude	0.88	<.001***
Gemini	0.98	<.001***
Human (video)	0.98	<.001***
Human (text)	0.99	<.001***

* = $p < .05$, ** = $p < .01$, *** = $p < .001$.

3.4 Dataset b: Validating a Population of AI Ratings

We then sought to address the possibility that our responses were not as varied as AI models could be, and that this variability is very different from that of human ratings, as seen by the large number of duplicate responses we observed. To respond to this potential limitation, we generated 10,000 more ratings using Gemini 1.5-pro for each story, to treat as a population, thus not requiring estimating the variability. Gemini was chosen because it showed the median level of variability with 43.47% identical ratings, with GPT having higher variability but worse performance, and Claude having lower variability but performing identically. We used each story in a prompt separately because, as mentioned above, we found no evidence for dependence between the models' answers in each iteration.

We first examined whether this new data was similar to our initial sample. To do this, we compared the percentage of identical ratings in the initial sample and new population. We found that contrary to our earlier expectation that this population would be more varied, there was instead lower variance in the new sample, which had an average of 99.85% repeated ratings, with the most common ratings being repeated between 36-99% of the time for each story. Generally speaking, LLMs generate very specific responses for each story, very close to providing a singularly unique rating of emotions.

3.5 Differences in Empathic Accuracy between the Human Samples and Gemini Rating Population

We next treated all 10,000 ratings of each story as a population, representing the LLM's empathic accuracy abilities, and compared our previous human sample to it. Upon comparing human participants to the mu of this population, we found that the human participants showed lower empathic accuracy abilities than those of the LLM overall ($t_{(1422)} = –21.51$, $p < .001$, Cohen's $d = .57$, M = 82.33, 95% CI [81.96, 82.70], compared to the $\mu_{LLM} = 86.40$). To account for the dataset including different stimuli, we also calculated the difference between each human rating and the most common rating Gemini 1.5 Pro generated for each story. We then used a mixed-effect linear model to compare these differences to 0, with a random intercept for each story. We

found that the overall difference was significantly lower than 0 (b = –4.95, *SE* = 1.17, $t_{(11.78)}$ = –4.24, p = .001). The same result held when examining the difference between the human rating and the mean empathic accuracy score for each story in the population, instead of the most common one (b = –4.83, SE = 1.03, $t_{(12.06)}$ = –4.68, p < .001). These results show that when compared to the most common rating and to the average score, human raters were still less empathically accurate than the LLM ratings.

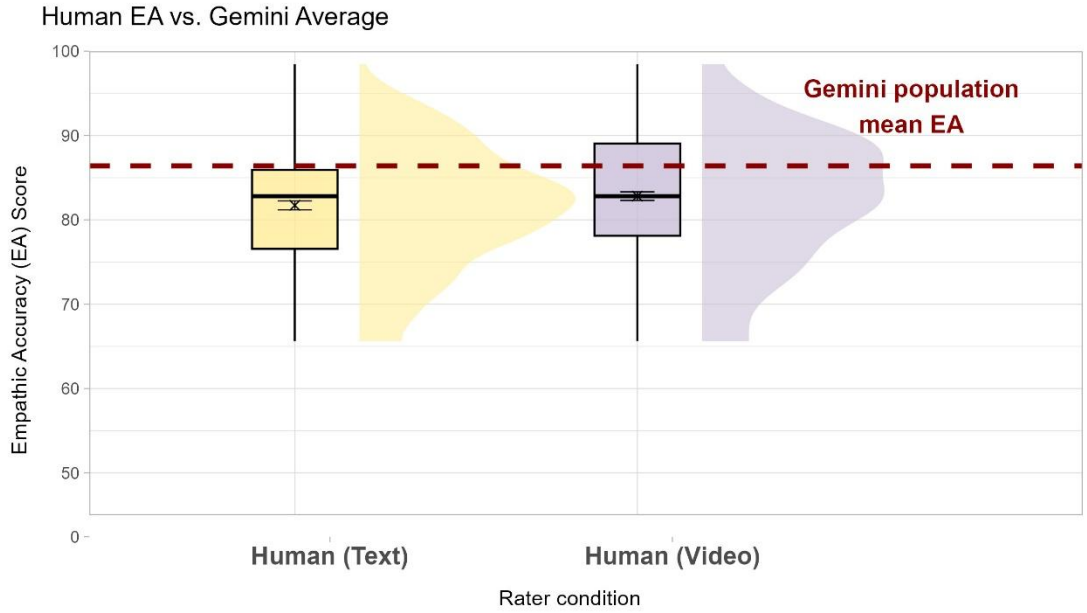

Figure 3: Differences in empathic accuracy between human raters and the mu of the population of gemini ratings. For each condition of human raters, the box marks the IQR, with the X inside marking the mean empathic accuracy with a 95% CI for the mean around it. The whiskers mark the total range, and the histogram to the side shows the distribution. The mu of the gemini population is marked by the dashed red line.

We again examined the differences for positive and negative emotions separately. We found that human raters performed worse than Gemini for both positive emotions ($t_{(1432)}$ = –17.40, p < .001, Cohen's d = .46, M = 82.06, 95% CI [81.51, 82.61], compared to the μ_{LLM} = 86.91) and negative emotions ($t_{(1426)}$ = –10.53, p < .001, Cohen's d = .28, M = 83.38, 95% CI [82.88, 83.88], compared to the μ_{LLM} = 86.09). We found similar results using linear-mixed effect models testing whether the difference between all human ratings and the most common rating by Gemini is different from 0, while accounting for valence. The model showed an intercept that was significantly lower than 0 (b = –4.42, SE = 1.03, $t_{(12.29)}$ = –4.28, p = .001) with significantly greater differences for positive emotions compared to negative emotions (b = –0.92, SE = 0.37, $t_{(2735.97)}$ = –2.46, p = .014, Cohen's d = .07). The same results were significant when

examining the average as opposed to the mode (b = –4.25, SE = 0.98, $t_{(12.54)}$ = –4.33, p = .001), with a similar effect of valence (b = –1.03, SE = 0.37, $t_{(2751.88)}$ = –2.79, p = .005, Cohen's d = .09). These results show that across all measures, human raters perform worse than Gemini in rating both positive and negative emotions, with greater differences in positive emotions.

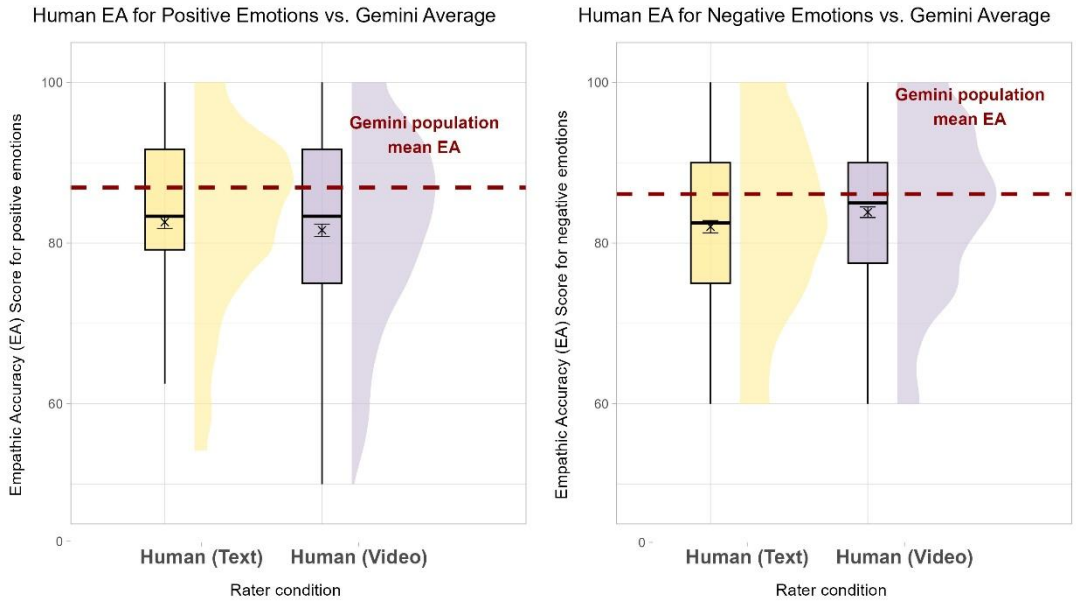

Figure 4: Differences in empathic accuracy between human raters and the mu of the population of gemini ratings for positive (left) and negative (right) emotions separately. For each condition of human raters, the box marks the IQR, with the X inside marking the mean empathic accuracy with a 95% CI for the mean around it. The whiskers mark the total range, and the histogram to the side shows the distribution. The mu of the gemini population for each valence is marked by the dashed red line.

The same results were replicated when we examined participants on an individual level, with Gemini scoring significantly higher than human raters (t(t(122) = –16.86, p < .001, 95% CI [81.04, 82.17], M = 81.60, Cohen's d = 1.52), and higher than 95% of human raters. The same was true for both positive emotions (t(125) = –15.89, p < .001, 95% CI [79.89, 81.45], M = 80.67, Cohen's d = 1.42), with Gemini scoring higher than 96% of participants, and for negative emotions (t(122) = –9.90, p < .001, 95% CI [81.39, 82.96], M = 82.17, Cohen's d = 0.89), with Gemini scoring higher than 78% of participants (see supplementary figure 1).

To summarize, when using transcripts of spoken videos in Hebrew, and using stimuli that could not have been part of the models' training data, LLMs are already significantly more accurate at detecting human emotions than human raters are, both

when humans receive the same text and when they have additional audio-visual channels available.

4. Discussion

The results of this study provide compelling evidence that current LLMs can achieve levels of empathic accuracy that are either similar to or exceed those of human participants, for both positive and negative emotions, even when operating under more constrained conditions. All tested models were either not different from or significantly outperformed humans who read the same textual transcripts or viewed the full audiovisual recordings. Notably, these findings emerged despite the fact that the LLMs were provided with only the Hebrew text of the emotional narratives—without access to tone, facial expressions, or other contextual cues, and without experiencing any emotion themselves. Moreover, they performed similarly to or better than participants who were close in age, culture, and social norms to the targets. This finding shows that LLMs were able to infer emotional states from language alone with a precision that is equal to or surpasses human performance.

The LLMs' high empathic accuracy was true both when response generation included the full experiment, and when prompting the models with each story individually. The models showed higher accuracy with lower variability than humans. In other words, the different distribution of LLM responses from that of human responses is a feature of LLMs, not a bug. These results shed light on mechanisms underlying empathic accuracy in both artificial and human agents.

4.1 Theoretical Implications

First, these results offer converging evidence that semantic information may be sufficient—at least for LLMs—to infer emotional states with high accuracy. LLMs not only matched but mostly outperformed humans in the same text-only condition, suggesting they may extract and weigh linguistic cues differently or more effectively than human readers do. Therefore, whereas LLMs are capable of inferring emotional states accurately through language, humans may do so through integrating additional processes. This is further evidenced by significant differences between the video and text conditions, with participants who received visual information being more accurate than those in the transcript condition, especially for negative emotions.

Second, by examining empathic accuracy in LLMs, our study uniquely isolates cognitive empathy (the ability to infer others' emotional states; Zaki & Ochsner, 2012) without the confounding effects of affective or motivational empathy. While in humans all three components of empathy typically co-occur (Depow et al., 2021; van den Bedem et al., 2019)— our findings suggest that affective or motivational empathy is not a prerequisite for cognitive empathy, as seen with LLMs. This highlights how artificial models can be used to isolate and examine cognitive empathy in ways that are difficult to achieve in typical populations. Previously, such distinctions could only be studied indirectly through clinical populations, whereas LLMs now offer a complementary, non-clinical model for exploring these dynamics.

Moreover, it is possible that AI models are capable of higher levels of cognitive empathy precisely because they do not experience emotions like people do. Unlike humans, in estimating the targets' emotions, LLMs may rely on the common expressions of emotions, while humans may be influenced by their own emotional state, thus adding specific situational and personal biases to their estimations.

This is not to say that AI is unbiased. It has been shown that AI systems' judgements display more bias than humans overall, and can even amplify these biases in humans following computer-human interactions (Glickman & Sharot, 2025). If this amplification is also true in the affective domain, continuously using AI to estimate and validate one's own emotions, or to understand those of others, may result in exaggeration or over-amplification of the actual emotional states. Thus, this exaggeration could potentially lead to further inaccuracies and misunderstandings in human communication (see Genzer et al., under review for already existing amplification effects in humans).

4.2 Practical Implications

Apart from theoretical contributions, the demonstrated ability of LLMs to accurately infer emotional states from naturalistic stimuli, has relevance for a wide range of applications. In clinical and assistive contexts, AI could support emotional reflection, enhance communication (Velagaleti, 2024; Zdravkova et al., 2022), and assist in emotion recognition for individuals who struggle with perspective-taking and interpreting emotional nuance, such as those with autism spectrum disorder (ASD) (see

for example Levy et al., 2024). Similarly, mental health professionals might use AI tools to track emotional patterns in therapy sessions, identify unspoken affect in written materials, or improve documentation of patient experiences (Luxton, 2014; Rebelo et al., 2023). However, such applications also raise important concerns because they may foster overreliance on AI systems, potentially leaving users at a disadvantage in settings where such tools are unavailable; and they pose significant ethical questions related to data privacy and the sensitive nature of emotional information.

In a world increasingly mediated by AI—through digital platforms, messaging systems, and virtual assistants—the capacity of AI to "read" emotions calls for a thoughtful examination of how it might transform interpersonal communication and relationships. AI may become an invisible intermediary, helping to translate, clarify, or even optimize emotional expression in human–human communication. This could, for instance, reduce misunderstandings in online exchanges, flag emotionally sensitive content before it causes harm, or assist individuals in expressing themselves more clearly in moments of distress (though see cautionary note on the risk of amplifying bias, above).

At the same time, such interventions risk altering the essence of human communication, which is often built on miscommunications, reinterpretations, and putting effort into better understanding those we care about (Naaman, 2022; Wilson, 2023). If AI systems routinely "smooth over" emotional misunderstandings, people may begin to rely on these systems to mediate their most intimate connections, which may affect learning, growth, and relationship deepening. In fact, research on AI-mediated communication has already shown that undisclosed AI involvement is considered unacceptable (Purcell et al., 2023). Moreover, any perceived AI involvement reduces the emotional meaning of communication between people (Glikson & Asscher, 2023; Hohenstein et al., 2023; Rubin et al., 2025). As these technologies become increasingly integrated into our relationships, they may inadvertently diminish the quality of interpersonal connection and negatively impact social well-being.

In addition to these interpersonal concerns, the ability of AI to detect emotions—particularly subtle or unintended ones—raises serious ethical questions. Emotion recognition may not always serve the user; it can also serve third parties, such as companies, governments, or platforms that collect, analyze, and act upon this

emotional data (D. C. Ong, 2021; Schaich Borg, 2021). Emotions that individuals may not wish to express outwardly could be inferred through text or speech, potentially without the user's knowledge or consent. This opens the door to new forms of surveillance and manipulation, whether through targeted advertising, political messaging, or more coercive control. In sensitive contexts, such as healthcare, education, or law enforcement, the potential for misuse is especially concerning (Jeyaraman et al., 2023; Leslie, 2019; Weber, 2020). While users might become more emotionally legible to machines, they may simultaneously lose control over how, when, and to whom their emotions are revealed. Ensuring that emotional data is protected, consensually shared, and ethically used must become a central priority in the development and deployment of AI systems capable of empathic inference.

Lastly, it should be noted that empathic accuracy alone does not guarantee meaningful empathic engagement. In human relationships, the subjective feeling of being understood, feeling genuinely heard, seen or cared for, often matters more than objective accuracy (Eyal et al., 2018; Rubin et al., 2025; Yin et al., 2024). Thus, although LLMs demonstrate remarkable capabilities for cognitive empathy, they fundamentally lack the emotional resonance and motivation to care that underlie human connection. As such, they may serve as valuable tools, but not replacements, in domains where emotional depth and interpersonal nuance are essential (Perry, 2023; Rubin et al., 2024).

Moreover, empathic missteps, when acknowledged and repaired, can strengthen relationships and foster emotional growth (Baldwin, 2014; Gordon & Chen, 2016). An AI system that detects emotions perfectly but does not genuinely feel or care may fail to provide a sense of connection or comfort. In contrast, a human who slightly misjudges an emotion but responds with warmth, care, and effort may be perceived as more genuinely empathic. Crucially, what makes an empathic response meaningful to the recipient and valued over time remains an open question that future research should explore more deeply.

4.3 Limitations and Future Directions

This study has several limitations. One is its reliance on Hebrew-language stimuli. On the one hand, this choice represents a methodological strength: Because

Hebrew is underrepresented in the training data of most LLMs, the use of these narratives reduces the likelihood that the models had prior exposure to the specific content or similar linguistic patterns, thereby providing a more ecologically valid test of zero-shot empathic inference. On the other hand, this design choice limits the generalizability of our findings. It remains an open question whether LLMs would demonstrate similar levels of empathic accuracy in other underrepresented languages or cultural contexts.

Second, the study focused on empathic accuracy from video recordings and their transcribed text. Future studies should assess how accurate LLMs are at inferring emotions during live interactions (via text or video), rather than recorded ones.

Lastly, we focused on a specific set of emotional stimuli and a small group of widely used, closed-source commercial LLMs, using a predefined temperature and a specific number of iterations determined by the researchers. These parameters were selected in the absence of established research conventions for evaluating LLMs in comparison to human performance. It is important to note, though, that the primary aim of this study was not to identify the most accurate LLM or optimal temperature setting, nor to explain the internal mechanisms or architectures of these proprietary models—whose exact algorithms, weights, and decision-making processes remain undisclosed, and change on a weekly basis. Rather, this was a proof of concept, designed to demonstrate the potential of LLMs as tools for addressing psychological questions and to explore their far-reaching implications for human relationships—both in terms of their promise and their potential for harm. Future research should extend this work by incorporating a broader range of autobiographical narratives, cultural contexts, and emotional content, as well as evaluating newer or more specialized models, including open-source LLMs.

4.4 Conclusions

This study demonstrates that LLMs are capable of remarkably high empathic accuracy, even when relying solely on text-based input and operating in a language and cultural context underrepresented in their training data. Their performance not only equaled or surpassed that of human participants in comparable conditions but also exceeded the empathic accuracy of humans with access to full audiovisual information.

These findings highlight the rich emotional information embedded in language, even if not fully exploited by humans, and add to the previous debate on the role of different informational cues in empathic accuracy. The ability of LLMs to exhibit high cognitive empathy in the absence of feeling or caring also underscores their potential as novel tools for studying cognitive empathy in isolation—offering a non-clinical complement to traditional research with special populations.

In a world increasingly mediated by AI, emotion-recognition technologies will undoubtedly reshape not only how we interact with machines, but also how we relate to one another—offering both opportunities for support and significant risks for emotional communication. Moreover, the ability to detect and interpret emotions at scale raises serious concerns about privacy, surveillance, and emotional autonomy. Future research in social psychology will be essential for understanding how these technologies influence human connection, trust, and emotional expression—and for guiding their integration in ways that support, rather than disrupt, the fabric of social life.

5. Declaration of generative AI and AI-assisted technologies in the writing process

In writing this paper, the authors utilized ChatGPT-4o during editing to paraphrase sentences for clarity. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article. There was no other use of generative AI.

6. Competing Interests Statement

The authors declare no competing interests.

7. Acknowledgements

Funding: This work was supported by a grant from the Azrieli Israel Center for Addiction and Mental Health to A.P., and a fellowship from the Azrieli Israel Center for Addiction and Mental Health to M.R.

We would also like to thank Dr. Noam Siegelman, for his extremely helpful consultation on the statistical analyses of this novel dataset.

8. References

Alabed, A., Javornik, A., Gregory-Smith, D., & Casey, R. (2024). More than just a chat: a taxonomy of consumers' relationships with conversational AI agents and their well-being implications. *European Journal of Marketing*, *58*(2), 373–409. https://doi.org/10.1108/EJM-01-2023-0037

Anderson, C., & Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behavioral and Brain Sciences*, *25*(1), 21–22. https://doi.org/10.1017/S0140525X02230010

Baron-Cohen, S. (2004). The essential difference: Male and female brains and the truth about autism (1st pbk. ed). Basic Books.

Batson et al. (1991). The Empathy–Altruism Hypothesis. In *The Oxford Handbook of Prosocial Behavior*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.023

Ben Adiva, Y., Genzer, S., & Perry, A. (2024). Beyond physical sensations: investigating empathy and prosocial behavior in vicarious pain responders. *Social Cognitive and Affective Neuroscience*, *19*(1). https://doi.org/10.1093/scan/nsae039

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, *148*(6), 962–976. https://doi.org/10.1037/xge0000595

Davis, M. H. (2017). *Empathy, Compassion, and Social Relationships* (E. M. Seppälä, E. Simon-Thomas, S. L. Brown, M. C. Worline, C. D. Cameron, & J. R. Doty, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190464684.013.23

Depow, G. J., Francis, Z., & Inzlicht, M. (2021). The Experience of Empathy in Everyday Life. *Psychological Science*, *32*(8), 1198–1213. https://doi.org/10.1177/0956797621995202

Derntl, B., Finkelmeyer, A., Voss, B., Eickhoff, S. B., Kellermann, T., Schneider, F., & Habel, U. (2012). Neural correlates of the core facets of empathy in schizophrenia. *Schizophrenia Research*, *136*(1–3), 70–81. https://doi.org/10.1016/j.schres.2011.12.018

Dey, K., Tarannum, P., Hasan, Md. A., Razzak, I., & Naseem, U. (2024). *Better to Ask in English: Evaluation of Large Language Models on English, Low-resource and Cross-Lingual Settings*. http://arxiv.org/abs/2410.13153

Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study. *JMIR Mental Health*, *11*(1). https://doi.org/10.2196/54369

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology*, *114*(4), 547–571. https://doi.org/10.1037/pspa0000115

Gandhi, K., Lynch, Z., Fränken, J.-P., Patterson, K., Wambu, S., Gerstenberg, T., Ong, D. C., & Goodman, N. D. (2024). *Human-like Affective Cognition in Foundation Models*. http://arxiv.org/abs/2409.11733

Genzer, S., Ben Adiva, Y., & Perry, A. (2024). *Empathy*. Cambridge University Press. https://doi.org/10.1017/9781009281072

Genzer, S., Ong, D. C., Zaki, J., & Perry, A. (2022). Mu rhythm suppression over sensorimotor regions is associated with greater empathic accuracy. *Social Cognitive and Affective Neuroscience*, *17*(9), 788–801. https://doi.org/10.1093/scan/nsac011

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology*, *77*(4), 746–761. https://doi.org/10.1037/0022-3514.77.4.746

Gleason, K. A., Jensen-Campbell, L. A., & Ickes, W. (2009). The Role of Empathic Accuracy in Adolescents' Peer Relations and Adjustment. *Personality and Social Psychology Bulletin*, *35*(8), 997–1011. https://doi.org/10.1177/0146167209336605

Glikson, E., & Asscher, O. (2023). AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. *Computers in Human Behavior*, *140*, 107592. https://doi.org/10.1016/j.chb.2022.107592

Gordon, A. M., & Chen, S. (2016). Do you get where I'm coming from?: Perceived understanding buffers against the negative impact of conflict on relationship satisfaction. *Journal of Personality and Social Psychology*, *110*(2), 239–260. https://doi.org/10.1037/pspi0000039

Gunes, H., Piccardi, M., & Pantic, M. (2008). From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities. In *Affective Computing*. I-Tech Education and Publishing. https://doi.org/10.5772/6180

Haensch, A.-C. (2025). *"It Listens Better Than My Therapist": Exploring Social Media Discourse on LLMs as Mental Health Tool*. http://arxiv.org/abs/2504.12337

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion*, *7*(2), 438–446. https://doi.org/10.1037/1528-3542.7.2.438

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Publisher Correction: Artificial

intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 16616. https://doi.org/10.1038/s41598-023-43601-0

Iegor Rudnytskyi. (2023). *Package "openai" Title R Wrapper for OpenAI API*.

Israelashvili, J., Sauter, D. A., & Fischer, A. H. (2020). Different faces of empathy: Feelings of similarity disrupt recognition of negative emotions. *Journal of Experimental Social Psychology*, *87*, 103912. https://doi.org/10.1016/j.jesp.2019.103912

Jeyaraman, M., Balaji, S., Jeyaraman, N., & Yadav, S. (2023). Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. *Cureus*. https://doi.org/10.7759/cureus.43262

Jinhwan Kim. (2024). *Title Interface for "Google Gemini" API*.

Jospe, K., Genzer, S., klein Selle, N., Ong, D., Zaki, J., & Perry, A. (2020). The contribution of linguistic and visual cues to physiological synchrony and empathic accuracy. *Cortex*, *132*, 296–308. https://doi.org/10.1016/j.cortex.2020.09.001

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. In *Trends in Cognitive Sciences* (Vol. 18, Issue 4, pp. 163–166). Elsevier Ltd. https://doi.org/10.1016/j.tics.2013.12.011

Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. American Psychologist, 72(7), 644–654. https://doi.org/10.1037/amp0000147

Kraus, M. W., & Segal, N. (2015). Empathic Accuracy Without Visual Cues. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2628240

Larson, E. B. (2005). Clinical Empathy as Emotional Labor in the Patient-Physician Relationship. *JAMA*, *293*(9), 1100. https://doi.org/10.1001/jama.293.9.1100

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). *Large Language Models Produce Responses Perceived to be Empathic*. http://arxiv.org/abs/2403.18148

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety*. https://doi.org/10.5281/zenodo.3240529

Levy, L., Ambaw, A., Ben-Itzchak, E., & Holdengreber, E. (2024). A real-time environmental translator for emotion recognition in autism spectrum disorder. *Scientific Reports*, *14*(1), 31527. https://doi.org/10.1038/s41598-024-83229-2

Li, H., & Zhang, R. (2024). Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots. *Journal of Computer-Mediated Communication*, *29*(5). https://doi.org/10.1093/jcmc/zmae015

Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, *45*(5), 332–339. https://doi.org/10.1037/a0034559

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. Emotion, 8(4), 540–551. https://doi.org/10.1037/a0012746

Naaman, M. (2022). "My AI must have been broken": How AI Stands to Reshape Human Communication. *Proceedings of the 16th ACM Conference on Recommender Systems*, 1–1. https://doi.org/10.1145/3523227.3555724

Norris, C. J. (2021). The negativity bias, revisited: Evidence from neuroscience measures and an individual differences approach. *Social Neuroscience*, *16*(1), 68–82. https://doi.org/10.1080/17470919.2019.1696225

Ong, D. C. (2021). An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. https://doi.org/10.1109/ACII52823.2021.9597441

Ong, D., Su, J., Chen, B., Luu, A. T., Narendranath, A., Li, Y., Sun, S., Lin, Y., & Wang, H. (2022). *Is Discourse Role Important for Emotion Recognition in Conversation?* www.aaai.org

Patel, D., Timsina, P., Raut, G., Freeman, R., levin, M. A., Nadkarni, G. N., Glicksberg, B. S., & Klang, E. (2024). *Exploring Temperature Effects on Large Language Models Across Various Clinical Tasks*. https://doi.org/10.1101/2024.07.22.24310824

Perry, A. (2023). AI will never convey the essence of human empathy. In *Nature Human Behaviour* (Vol. 7, Issue 11, pp. 1808–1809). Nature Research. https://doi.org/10.1038/s41562-023-01675-w

Perry, A., Saunders, S. N., Stiso, J., Dewar, C., Lubell, J., Meling, T. R., Solbakk, A. K., Endestad, T., & Knight, R. T. (2017). Effects of prefrontal cortex damage on emotion understanding: EEG and behavioural evidence. *Brain*, *140*(4), 1086–1099. https://doi.org/10.1093/brain/awx031

Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N., & Jakesch, M. (n.d.). *LRH: AI-Mediated Communication*.

R Core Team. (2023). *R: A language and environment for statistical computing* (v4.3.2). R Foundation for Statistical Computing .

Rafaeli, E., Gadassi, R., Howland, M., Boussi, A., & Lazarus, G. (2017). Seeing bad does good: Relational benefits of accuracy regarding partners' negative moods. *Motivation and Emotion*, *41*(3), 353–369. https://doi.org/10.1007/s11031-017-9614-x

Rakel, D. P., Hoeft, T. J., Barrett, B. P., Chewning, B. A., Craig, B. M., & Niu, M. (2009). *Practitioner Empathy and the Duration of the Common Cold*.

Rebelo, A. D., Verboom, D. E., dos Santos, N. R., & de Graaf, J. W. (2023). The impact of artificial intelligence on the tasks of mental healthcare workers: A scoping review. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100008. https://doi.org/10.1016/j.chbah.2023.100008

Refoua, E., Meinlschmidt, G., & Elyoseph, Z. (2024). *GENERATIVE AI: EXCELLENT EMOTION RECOGNITION ACROSS ETHNICS 1 Generative Artificial Intelligence Demonstrates Excellent Emotion Recognition Abilities Across Ethnical Boundaries*. https://ssrn.com/abstract=4901183

Renze, M. (2024). The Effect of Sampling Temperature on Problem Solving in Large Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7346–7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, *15*(6), 192. https://doi.org/10.3390/fi15060192

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rubin, M., Arnon, H., Huppert, J. D., & Perry, A. (2024). Considering the Role of Human Empathy in AI-Driven Therapy. In *JMIR Mental Health* (Vol. 11). JMIR Publications Inc. https://doi.org/10.2196/56529

Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the value of perceived human versus AI-generated empathy. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-025-02247-w

Rum, Y., & Perry, A. (2020). Empathic Accuracy in Clinical Populations. In *Frontiers in Psychiatry* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fpsyt.2020.00457

Schaich Borg, J. (2021). Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211040197

Schlegel, K., Sommer, N. R., & Mortillaro, M. (2025). Large language models are proficient in solving and creating emotional intelligence tests. Communications Psychology, 3(1), 1–14. https://doi.org/10.1038/s44271-025-00258-x

Schlegel, K., & Scherer, K. R. (2018). The nomological network of emotion knowledge and emotion understanding in adults: Evidence from two new performance-based tests. Cognition and Emotion. https://www.tandfonline.com/doi/full/10.1080/02699931.2017.1414687

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Sened, H., Lavidor, M., Lazarus, G., Bar-Kalifa, E., Rafaeli, E., & Ickes, W. (2017). Empathic accuracy and relationship satisfaction: A meta-analytic review. *Journal of Family Psychology*, *31*(6), 742–752. https://doi.org/10.1037/fam0000320

Sesha Bhargavi Velagaleti. (2024). Empathetic Algorithms: The Role of AI in Understanding and Enhancing Human Emotional Intelligence. *Journal of Electrical Systems*, *20*(3s), 2051–2060. https://doi.org/10.52783/jes.1806

Shamay-Tsoory, S. G., Tomer, R., Goldsher, D., Berger, B. D., & Aharon-Peretz, J. (2004). Impairment in cognitive and affective empathy in patients with brain lesions: Anatomical and cognitive correlates. *Journal of Clinical and Experimental Neuropsychology*, *26*(8), 1113–1127. https://doi.org/10.1080/13803390490515531

Song, Y., Nie, T., Shi, W., Zhao, X., & Yang, Y. (2019). Empathy Impairment in Individuals With Autism Spectrum Conditions From a Multidimensional Perspective: A Meta-Analysis. In *Frontiers in Psychology* (Vol. 10). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2019.01902

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

van den Bedem, N. P., Willems, D., Dockrell, J. E., van Alphen, P. M., & Rieffe, C. (2019). Interrelation between empathy and friendship development during (pre)adolescence and the moderating effect of developmental language disorder: A longitudinal study. *Social Development*, *28*(3), 599–619. https://doi.org/10.1111/sode.12353

Wang, Y., Sun, Z., Fan, J., & Ma, H. (2024). On the Uses of Large Language Models to Design End-to-End Learning Semantic Communication. *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6. https://doi.org/10.1109/WCNC57260.2024.10570717

Weber, A. (2020). *ETHICS CONCERNS IN ARTIFICIAL INTELLIGENCE USE IN EDUCATION*. 4539–4544. https://doi.org/10.21125/inted.2020.1262

Weisz, E., & Zaki, J. (2017). *Empathy-Building Interventions* (E. M. Seppälä, E. Simon-Thomas, S. L. Brown, M. C. Worline, C. D. Cameron, & J. R. Doty, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190464684.013.16

Wilson, S. R. (2023). Rethinking Communication in the Era of Artificial Intelligence: An *HCR* Special Issue. *Emerging Media*, *1*(1), 40–45. https://doi.org/10.1177/27523543231188791

Windisch, P., Dennstädt, F., Koechli, C., Schröder, C., Aebersold, D. M., Förster, R., & Zwahlen, D. R. (2024). The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. *Cureus*. https://doi.org/10.7759/cureus.75748

Yamil Velez. (2024). *claudeR* (version 0.0.0.9000).

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(14). https://doi.org/10.1073/pnas.2319112121

Yongsatianchot, N., Thejll-Madsen, T., & Marsella, S. (2024). *Exploring Theory of Mind in Large Language Models through Multimodal Negotiation*. *24*. https://doi.org/10.1145/3652988.3673960

Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls and promise. In *Nature Neuroscience* (Vol. 15, Issue 5, pp. 675–680). https://doi.org/10.1038/nn.3085

Zdravkova, K., Krasniqi, V., Dalipi, F., & Ferati, M. (2022). Cutting-edge communication and learning assistive technologies for disabled children: An artificial intelligence perspective. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.970430