Empathy Simulation in Large Language Models: Evaluation, Enhancement, and Challenges

Authors: Zhou Qianyi, Cai Yaqi, Zhang Ya, Zhang Ya

Date: 2025-06-14T23:58:19+00:00

Abstract

With the technological advancements of Large Language Models (LLMs) in natural language generation and affective computing, their capacity for empathy simulation in domains such as psychological counseling, doctor-patient communication, and customer service has attracted considerable attention. The empathy simulation exhibited by LLMs is predominantly cognitive empathy simulation rather than emotional empathy simulation, principally comprising emotion recognition, empathetic response, and contextual adaptation. Current methodologies for evaluating LLMs' empathy simulation encompass human evaluation, automated evaluation, and task-driven evaluation, each presenting distinct advantages and disadvantages with varying applicable scenarios. Compared to human empathic capabilities, LLMs demonstrate remarkable performance in empathy generation tasks, yet still confront limitations in affective understanding; to further enhance the empathy simulation capabilities of LLMs, approaches such as data augmentation, model framework and architecture optimization, reinforcement learning, and prompt optimization may be adopted for improvement. Concurrently, ethical guidelines and potential risks associated with model deployment warrant continued attention.

Full Text

Empathy Simulation in Large Language Models: Evaluation, Enhancement, and Challenges

ZHOU Qianyi[1], **CAI Yaqi**[1], **ZHANG Ya**[1,2]
(1 Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China)
(2 Key Laboratory of Philosophy and Social Science of Anhui Province on Adolescent Mental Health and Crisis Intelligence Intervention, Hefei Normal University, Hefei 230601, China)

Abstract

With advances in large language models (LLMs) in natural language generation and affective computing, their capacity for empathy simulation has attracted widespread attention in domains such as psychological counseling, doctor-patient communication, and customer service. Empathy simulation in LLMs primarily manifests as cognitive rather than emotional empathy, encompassing emotion recognition, empathic response generation, and contextual adaptation. Current evaluation methods for LLMs' empathy simulation include human evaluation, automated evaluation, and task-driven evaluation, each with distinct advantages and limitations suited to different scenarios. Compared to human empathy, LLMs demonstrate strong performance in empathy generation tasks but still face fundamental limitations in emotional understanding. To further enhance LLMs' empathy simulation capabilities, strategies such as data augmentation, model framework and architecture optimization, reinforcement learning, and prompt engineering can be employed. Meanwhile, ethical considerations and potential risks associated with model deployment require careful attention.

Keywords: large language models, empathy simulation, empathy evaluation, ethical issues

1. Introduction

In recent years, large language models (LLMs) have continuously evolved, demonstrating intelligence comparable to humans (Wang, Ma, et al., 2024). Currently, LLMs have been widely applied to language generation tasks such as question answering, text summarization, and logical reasoning (Laskar et al., 2023; Ziyu et al., 2023), and their role-playing capabilities enable them to simulate human participants in experiments (Dillion et al., 2023). To better adapt LLMs to human needs in practical applications, research on their ability to simulate human empathy has become a focal point, particularly in medical communication, psychological counseling, and customer service, where empathy simulation demonstrates significant application value.

Empathy refers to the psychological capacity to recognize others' emotions, understand their affective states, and produce adaptive responses. In human-computer interaction (Liu-Thompkins et al., 2022; Svikhnushina & Pu, 2022), accurately understanding the thoughts and emotions embedded in others' narratives and providing appropriate responses is crucial. Existing research has found that empathy typically includes cognitive empathy and emotional empathy. The former refers to the ability to understand others' emotions, closely related to theory of mind, while the latter refers to the capacity to resonate with others' emotions and produce emotional reactions. These two types may correspond to different patterns of brain activity (Ma et al., 2024). Recent review studies on LLMs' empathy simulation characteristics indicate that LLMs' empathy simulation primarily manifests as cognitive empathy simulation rather than

emotional empathy simulation. This is because their empathic responses rely on statistical learning of linguistic patterns from training data rather than genuine emotional experiences or intrinsic motivations, fundamentally differing from human empathy, which is based on subjective experiences, emotional resonance, and moral motivations. At its core, LLMs cannot truly experience human emotions. However, in tasks involving emotion recognition and empathic response generation, LLMs sometimes outperform humans in certain contexts (Sorin et al., 2024). Nevertheless, for effective empathy simulation, LLMs must accurately display appropriate emotional reactions according to specific contexts to achieve emotional alignment with humans.

Compared to logical reasoning, text summarization, or information provision, empathy simulation poses greater challenges for LLMs, as it represents a complex and multi-layered response that requires not only proficient language mastery but also deep understanding of human psychology, emotions, and social contexts (Huang et al., 2025). Research on LLMs' empathy simulation capabilities is vital for enhancing human-computer interaction experiences and improving user trust and satisfaction. For instance, in healthcare, LLMs can generate empathic responses that improve patients' experiences in medical interactions and enhance satisfaction (Luo et al., 2024). In psychological counseling, LLMs can serve as mental health support tools, providing immediate emotional support and advice to help users cope with emotional distress. By integrating multimodal data, LLMs can more accurately identify user emotions, simulate counselors' empathy and proactive guidance capabilities, and assist users in self-reflection and emotional management (Ren et al., 2024).

2.1 Human Evaluation

Human evaluation employs two common approaches. The first relies on human annotators or target users to provide Likert-scale subjective ratings of LLMs' empathy simulation capabilities, for example, "Please rate the level of empathy in this response based on what you would provide in this situation" (Welivita & Pu, 2024). Likert scales typically range from 1 (not empathetic at all) to 5 (highly empathetic) (e.g., Ayers et al., 2023; Lee, Suh, et al., 2024) or 1 to 3 (e.g., Welivita & Pu, 2024). Evaluation dimensions typically include: 1) emotional understanding—assessing whether the model correctly identifies users' emotions, such as recognizing sadness when a user says "I feel very sad today" ; 2) emotional response—evaluating whether generated text demonstrates appropriate empathy, such as comfort, support, or advice; 3) contextual adaptation—checking whether model responses fit the conversational context and avoid irrelevant or inappropriate replies; and 4) language naturalness—assessing the fluency, appropriateness of word choice, and conformity to human communication habits. Additionally, professionals in psychological counseling and medical services can be invited to conduct expert evaluations of model performance (e.g., Ayers et al., 2023).

The second approach involves human evaluators judging which of two responses

in the same scenario (one human-generated, one LLM-generated) is more empathetic. For example, when comparing the empathy of LLM versus human physician responses, evaluators read patient questions and corresponding answers to determine which is more empathetic (Luo et al., 2024). This method is less widely used than Likert scoring, possibly because Likert scales enable fine-grained quantitative assessment of different models' empathy levels (e.g., 1–5 points), whereas simply comparing human and model responses provides limited information and is difficult to apply when horizontally comparing multiple models.

2.2 Automated Evaluation

Automated evaluation methods utilize algorithms or existing machine learning techniques to conduct rapid, objective quantitative analysis of LLMs' empathy simulation capabilities. For example, in emotional similarity assessment, pre-trained language models (such as BERT[2], RoBERTa[3]) can be employed for text emotion classification and consistency comparison. Cosine similarity can be used to calculate the similarity between user input and generated text in emotional vector space (e.g., Reimers & Gurevych, 2019). Automated evaluation methods (such as emotion classification and cosine similarity calculation) provide efficient and quantifiable means for assessing LLMs' empathy simulation capabilities. These methods are suitable for large-scale data analysis, helping researchers and developers rapidly optimize models to enhance their performance in empathic dialogue.

2.3 Task-Driven Evaluation

Task-driven evaluation assesses empathy simulation capabilities by having models perform specific emotion recognition and empathy tasks. For example, Emotion Recognition in Conversations (ERC) tasks evaluate LLMs' empathy simulation by measuring their accuracy in classifying emotions in dialogues (Zhao et al., 2023). Building upon this, the Recognizing Emotion Cause in Conversations (RECCON, Poria et al., 2021) task further expands the depth of emotion recognition by focusing on the sources of emotions in conversations, including two subtasks: Causal Span Extraction (CSE), which aims to identify the causal spans of speakers' non-neutral utterances (i.e., emotional causes), and Causal Emotion Entailment (CEE), which, given a speaker' s non-neutral utterance, aims to predict which specific utterances in the dialogue history caused the non-neutral emotion. The RECCON task requires models to analyze causal relationships in dialogues and identify which utterances triggered what emotional changes. Recognizing emotion causes in conversations helps comprehensively understand users' emotional states and enhances the interpretability and performance of emotion-based models (Zhao et al., 2023).

In addition to commonly used tasks in machine learning, numerous tasks from psychology can also measure LLMs'empathy simulation capabilities. Hagendorff

et al. (2023) noted that treating LLMs as participants in psychological experiments facilitates using psychological methods to test LLMs' complex and novel behavioral patterns, revealing capabilities that traditional natural language processing methods cannot detect. For instance, empathy accuracy measurement has developed relatively complete paradigms that can assess LLMs'empathy simulation capabilities by having them complete tasks. Ickes et al. (1990) designed a naturalistic task using videos as empathy materials. This paradigm requires participants to watch videos of dyadic conversations and infer one party' s emotions and thoughts, measuring empathic accuracy—the degree to which participants accurately infer others' emotions and thoughts. When applied to LLMs, this paradigm can assess models' cognitive empathy simulation capabilities by having them analyze dialogue or video descriptions to infer protagonists' emotions and thoughts, which are then compared with actual conditions. Additionally, researchers have used psychological scales such as the Interpersonal Reactivity Index (IRI) and Basic Empathy Scale (BES) to measure LLMs' empathy simulation capabilities (Pan et al., 2024).

In summary, evaluation of LLMs' empathy simulation capabilities focuses on the cognitive empathy domain, employing both human subjective evaluation and objective metrics using algorithms and tasks. Human evaluation can account for contextual factors and optimize models through human feedback; automated evaluation enables rapid quantitative analysis; and task-driven evaluation measures LLMs'empathy capabilities in real task scenarios, closer to application contexts. Different evaluation methods have distinct characteristics but may also present contradictions and complementarity (see Table 1). Future work should consider developing a unified standardized measurement framework to systematically assess LLMs' empathy simulation capabilities. For example, Huang et al. (2024) attempted to propose a unified emotional intelligence testing framework, EmotionBench . However, such unified frameworks for measuring LLMs' empathy simulation remain to be developed. Furthermore, using psychological standardized scales and empathy measurement paradigms to evaluate LLMs'empathy simulation capabilities has become a new research direction. For instance, using psychological scales can further improve evaluation accuracy, while using psychological paradigms (such as empathy accuracy experiments) can explore LLMs' complex and novel response patterns.

3. Current State of LLMs' Empathy Simulation

In the past two years, increasing research has supported that LLMs' ability to generate empathic responses is comparable to or exceeds that of humans. For example, Ayers et al. (2023) found that human participants rated GPT-3' s responses in the medical domain as more empathetic than those from physicians on online medical platforms. Luo et al. (2024) suggested that LLM-powered chatbots could potentially surpass human physicians in providing empathetic communication. Welivita and Pu (2024) compared LLM-generated empathic responses with original responses from the EmpatheticDialogues dataset, finding

that human raters perceived LLMs' responses as more empathetic. Lee, Suh et al. (2024) had multiple LLMs generate empathic responses to posts describing common life experiences (such as workplace situations, parenting, intimate relationships, and other anxiety- and anger-inducing scenarios), consistently finding that LLM-generated responses scored higher in empathy than human-written responses.

Furthermore, linguistic analyses indicate that these models exhibit unique and predictable "styles" in their use of punctuation, emoticons, and specific vocabulary (Lee, Suh et al., 2024). Loh and Raamkumar (2023) designed an experiment in psychology to evaluate LLMs' ability to generate empathic responses in simulated mental health counseling dialogues, finding that in most cases, LLMs' responses were significantly more empathetic than human responses. These findings highlight LLMs' potential to provide empathic responses and support humans in scenarios requiring empathy.

However, other studies have identified suboptimal performance in certain subtasks of empathy simulation. For example, Zhao et al. (2023) distinguished LLMs' emotional dialogue capabilities into understanding (including emotion recognition, emotion cause identification, and contextual adaptation) and generation (including empathic reactions and empathic support). Their research evaluated ChatGPT' s performance in emotional dialogue understanding and generation through a series of downstream tasks. Results showed that ChatGPT still has room for improvement in recognizing human emotions. For instance, in doctor-patient dialogues, when patients described headache symptoms, dataset annotations categorized the emotion as neutral, while ChatGPT interpreted it as sadness. This inconsistency in standards may not stem from ChatGPT' s capabilities but rather from a lack of sample prompts.

Regarding empathic reactions, ChatGPT often first restates the user' s emotions before expanding on information, a pattern that may become tiresome for users. Once a user' s predicament is confirmed, ChatGPT is too eager to provide corresponding advice and coping strategies while neglecting to soothe and care for the user' s emotions. Additionally, Pan et al. (2024) used psychological scales (Basic Empathy Scale and Interpersonal Reactivity Index) to compare scores of 1,200 participants simulated by GPT-4 with 1,200 real human participants, finding that GPT-4-simulated participants scored significantly lower than humans on both scales.

While previous studies evaluating LLMs' empathy simulation capabilities in specific contexts found they exhibited certain levels of empathy recognition and reaction, standardized questionnaire assessments suggest that LLMs have not yet reached human levels of empathy. In summary, LLMs' ability to generate empathic responses in specific contexts is already comparable to or exceeds human performance, enabling them to provide emotional support in affective scenarios. However, there remains room for improvement in subtasks of empathy simulation and standardized empathy scale tests, indicating that their empathy simulation capabilities require further enhancement. With the rapid

development of LLMs, current research findings may not fully represent the state of LLMs' empathy simulation capabilities. For instance, no studies have yet measured the empathy capabilities of the DeepSeek model .

More importantly, research on LLMs' empathy simulation is challenging long-standing human understanding of empathy' s essence. Empathy has traditionally been considered a uniquely human trait, with definitions typically based on human-to-human interactions. The complexity of empathy, influenced by personality, culture, and context, has led to ambiguity in its definition. LLMs' empathy simulation is fundamentally different from human empathy because algorithms do not genuinely experience emotions. Therefore, human-centric definitions of empathy may not apply to LLMs. However, if LLMs' generated responses align with human expectations or preferences, can these observable responses be considered empathy? When people cannot distinguish between human and LLM-generated responses, or even prefer LLM-generated answers, this may indicate that LLMs' simulated empathy is sufficient (Huang et al., 2025).

4. Enhancement Strategies for LLMs' Empathy Simulation

LLMs' ability to infer human emotions and thoughts still has considerable room for improvement. Further considering cultural and contextual influences, developing LLMs aligned with human emotions requires data augmentation, architecture optimization, reinforcement learning, and thoughtful prompt design.

4.1 Data Augmentation

Constructing empathic dialogue datasets containing richer emotional information enables fine-tuning of large language models. The most commonly used emotional dialogue dataset is the EmpatheticDialogues (ED) dataset proposed by Rashkin et al. (2018), which contains 25,000 rounds of emotional conversations. To further improve LLMs' empathy simulation capabilities, many researchers have augmented the ED dataset and used the enhanced data for LLM fine-tuning to improve empathic dialogue generation. For example, Cao et al. (2024) used large models to filter segments containing emotional information from the ED dataset and completed emotional information annotation, constructing the TOOL-ED dataset. They found that fine-tuning LLMs with this dataset improved their ability to generate empathic dialogue.

Increasing dataset size, improving data quality, and adding conversation rounds can also enhance empathic dialogue datasets and thus improve LLMs' empathy simulation capabilities. For instance, Synth-Empathy is a data generation and filtering method based on LLMs for automatically generating high-quality empathic dialogue data (Liang et al., 2024). It optimizes the data generation pipeline to efficiently and cost-effectively create large-scale empathic datasets. Through quality and diversity selection mechanisms, it automatically retains high-quality data and removes low-quality data, ensuring generated dialogue data meets high standards in empathy, relevance, and diversity. Chen et

al. (2023) constructed a multi-turn empathic dialogue dataset containing over 2 million samples, training models to use various emotional support expressions such as questioning, comforting, acknowledging, listening, and trust when facing different situations. Experiments demonstrated that fine-tuning LLMs with multi-turn dialogues closer to counselors' expressions can significantly enhance empathy simulation capabilities.

4.2 Model Architecture and Framework Optimization

Improving and adjusting model structure, components, and information flow can enhance LLMs' performance in generating empathic dialogue. Cai et al. (2024) designed a new model, Pecer , which can dynamically capture emotional and personality information from historical dialogue and integrate this information into empathic response generation, thereby achieving better empathic effects. Liu et al. (2024) combined grammatical dependency trees and emotional vocabulary modules (National Research Council Emotion Lexicon, NRC) to extract emotional words from historical dialogue and construct emotional dependency trees, thereby better capturing emotional information and conducting more targeted empathic response generation for speakers.

In addition to optimizing LLMs' underlying architecture, constructing usage frameworks for LLMs can also improve empathy simulation capabilities. While LLMs excel at generating empathic responses, they lack deep understanding of subtle differences between emotions and cognitive affect, whereas small-scale empathetic models (SEMs[1]) exhibit the opposite characteristics. The Hybrid Empathetic Framework (HEF, Yang et al., 2024) treats SEMs as flexible plugins that can enhance LLMs' subtle emotional and cognitive understanding, alleviating LLMs' difficulties in fine-grained emotion detection and improving their ability to identify emotion causes. The Empathizing Before Generation (EBG) framework (Zhu et al., 2025) enables LLMs to analyze chain of thought (COT[11]) before generating responses, improving emotion inference capabilities and accuracy in emotion label identification.

4.3 Reinforcement Learning

Introducing human or LLM feedback on generated responses optimizes models' performance in emotional expression and naturalness. This method guides models to generate responses more aligned with human expectations through reward mechanisms. For example, Wang, Xu et al. (2024) introduced a feedback framework called Muffin, which evaluates the effectiveness of generated responses through multifaceted AI feedback. Specifically, the framework optimizes the model by contrasting "helpful" and "unhelpful" responses, enabling it to better distinguish high-quality responses and reduce the probability of generating low-quality replies. However, the evaluation process for introducing human feedback remains to be standardized. For instance, in Sabour et al.' s (2022) study, only three people were recruited to compare 100 groups of empathic dialogue content

generated by different systems to determine which system was better. This evaluation method makes it difficult to guarantee result stability and effectiveness in solving practical problems. Sharma et al. (2023) used randomized controlled trials to compare empathy in human versus human+AI responses. However, the evaluation criteria used in this study only assessed empathy displayed in dialogue content from a third-party perspective, rather than empathy genuinely felt by the recipients. Future work should consider introducing standardized recipient feedback to guide models in generating empathy responses more aligned with human expectations.

4.4 Prompt Design

Adjusting input prompts or contextual information can guide LLMs to generate responses meeting specific emotional needs. Welivita and Pu (2024) found that including detailed definitions of empathy in prompts helps LLMs generate empathic responses. Lee, Lee et al. (2024) integrated counseling psychology theoretical models into prompts, proposing Chain-of-Empathy Prompting (CoE[12]) to optimize LLMs' empathic responses. For example, when optimizing empathic responses with cognitive behavioral therapy, prompts guide LLMs to analyze users'expressed emotions and potential cognitive traps in their statements before providing responses. LLMs prompted in this manner produce more empathic responses.

In summary, LLMs' empathy simulation capabilities have attracted increasing attention, and optimizing their performance in empathic dialogue has become a core research direction in affective computing. Researchers have improved LLMs' empathy simulation capabilities through multidimensional strategies including data augmentation, model architecture optimization, reinforcement learning, and refined input prompts (see Table 2). While these strategies have all enhanced LLMs' empathy simulation capabilities, they emphasize different aspects. Future work should optimize strategy combinations based on practical application needs to generate empathy responses that better meet human expectations.

Currently, most optimization strategies rely on the EmpatheticDialogues dataset as a comparison baseline, limiting evaluation diversity and real-world applicability. Future research should combine human evaluation and psychological experimental paradigms (such as controlled experiments and emotional scale measurements) to validate the empathy capabilities of optimized models in diverse real-world application scenarios (such as psychological counseling and customer service). Additionally, most enhancement strategies are based on LLMs' pre-training and fine-tuning, relying on their language generation and logical reasoning capabilities. Therefore, continuous improvement of LLMs' foundational capabilities remains crucial for empathy optimization and should not be overlooked.

5.1 Lack of Authenticity in Empathy Simulation

LLMs rely on natural language processing and statistical learning to generate text, and their "empathy" performance primarily stems from pattern learning (Hou et al., 2024), meaning their mechanism is essentially based on analyzing and simulating massive statistical data rather than genuine emotional experiences or intrinsic understanding (Shao et al., 2024; Sorin et al., 2024). While LLMs can identify emotional tendencies in text and generate corresponding empathic responses, and can select appropriate linguistic styles through contextual understanding to make expressions more contextually appropriate, LLMs' empathic responses often exhibit excessive stacking of emotional vocabulary and stereotypical response patterns (Naik et al., 2024; Sorin et al., 2024). This makes it difficult for them to demonstrate genuine, nuanced deep understanding of emotions, leading to questions about the authenticity of LLMs' empathic expression. A recent study compared the effectiveness of LLM-based chatbot-assisted venting versus traditional diary platform venting in reducing negative emotions and increasing perceived social support. Results showed that chatbot-assisted venting effectively reduced high- and medium-arousal negative emotions such as anger, frustration, and fear. However, participants' perceived social support did not significantly increase under chatbot-assisted venting conditions, and they still perceived loneliness, suggesting that participants may not have perceived the chatbot' s effective assistance as social support (Hu et al., 2025).

5.2 Difficulty Handling Complex Emotions and Contexts

Human emotional expression and experience possess high complexity, and LLMs lacking emotional understanding struggle to respond appropriately. For example, individuals may experience mixed emotional states such as bittersweet feelings and may employ non-literal expressions such as irony to convey emotions. However, when facing such complex emotions and expressions, LLMs tend to respond by selecting from emotion words in user inputs, demonstrating low emotional intelligence and struggling to achieve effective empathic responses (Naik et al., 2024). Moreover, LLMs show significant limitations when handling tasks involving social interaction and emotional understanding. For example, in assessments of social intelligence and "Theory of Mind," models such as GPT-3 performed poorly in understanding participants' intentions and reactions in social interactions, with accuracy rates of only 55% to 60%, far below human levels (Sap et al., 2022).

Cultural contextual differences further exacerbate challenges for LLMs in providing culturally sensitive empathic content. Empathy expression varies across cultural backgrounds, and LLMs struggle to adapt flexibly to these differences. Due to training data biases, most current LLMs exhibit an "English-centric" tendency in generated text, even when responding to prompts in other languages, still reflecting Western emotional expression norms. This indicates that current multilingual LLMs have not yet successfully learned appropriate emotional expression nuances across different cultures (Havaldar et al., 2023).

5.3 Ethical and Usage Risks

First, empathic simulation content generated by LLMs carries potential content risks. For instance, given inherent limitations in training data and algorithms, LLMs inevitably have the potential to generate harmful content such as aggressive or discriminatory information (Cuadra et al., 2024; Patil et al., 2024; Shen et al., 2024). Such harmful information not only significantly degrades user experience but may also negatively impact individual mental health.

Second, LLMs' empathy simulation capabilities may be misused, leading to information manipulation or users' psychological dependence on models, potentially violating public order and good customs. For example, chatbots can significantly reduce users' vigilance by simulating human behaviors such as friendly tones, emotional expressions, and open-ended question design. This design strategy leads users to perceive robots as entities with human-like intentions. Even when users rationally know that chatbots are non-human technical products, they may still develop emotional dependence on robots and share private information during interactions (Holmes et al., 2022). When users gradually become accustomed to seeking emotional support and empathic feedback from LLMs, this may paradoxically intensify individuals' loneliness experiences (Koranteng et al., 2023). Additionally, while emotional communication with LLMs may increase user experience pleasure in the short term, long-term use may weaken users' willingness and ability to engage in emotional interactions with real people, leading to decreased sensitivity to genuine emotional interactions and adversely affecting real-world interpersonal relationships and social adaptability.

Finally, LLMs' training data typically comes from broad sources such as the internet, books, and social media, inevitably containing sociocultural biases and stereotypes. These biases may be inadvertently amplified and disseminated when LLMs generate empathic responses, triggering ethical risks. For example, LLMs generate responses through statistical probabilities, tending to reproduce high-frequency linguistic patterns in training data. If certain groups are stereotyped in the data (e.g., "Asians are good at math"), LLMs may inadvertently repeat these stereotypes in empathic responses. For instance, when a user mentions academic pressure experienced by an Asian student, LLMs might generate responses like "Your math grades must be excellent, but the pressure is also high," reinforcing stereotypes. Similarly, for female users' career dilemmas, LLMs might generate responses like "As a woman, you might be better at coordination, so perhaps consider administrative positions," implying gender bias that appears "reasonable" due to its "empathic" tone. When biases are expressed and disseminated by LLMs in seemingly well-intentioned empathic ways, their potential harm may be concealed, subtly influencing users' cognition and behavior, solidifying existing biases and impressions, and even exacerbating social inequality.

6. Conclusion

As LLMs continue to develop, their empathy simulation capabilities have become a hot topic in artificial intelligence research, particularly demonstrating potential application value in healthcare and psychological counseling. Existing research primarily explores how to enhance LLMs' human-like empathic responses during interactions by evaluating multiple dimensions including emotional understanding, emotional response, contextual adaptation, and language naturalness.

Currently, evaluation methods for LLMs' empathy simulation capabilities are becoming increasingly sophisticated, with researchers measuring empathy simulation levels from different perspectives, including human subjective evaluation, algorithmic automated evaluation, and task-driven evaluation methods. Diverse evaluation approaches have assessed LLMs' empathic response capabilities multidimensionally, yet shortcomings remain. Future efforts should focus on constructing unified frameworks for systematically measuring LLMs' empathy simulation capabilities and consider incorporating psychological empathy measurement paradigms to explore more complex and novel empathic response patterns in LLMs.

Based on existing evaluation methods, most studies support that LLMs' ability to generate empathic responses is comparable to humans, though their capacity to accurately infer human emotions and thoughts remains to be improved. Current enhancement strategies include data augmentation, model architecture optimization, reinforcement learning from human feedback, and refined prompt design, which have improved LLMs' empathy simulation capabilities at different levels. Future work should select appropriate strategies based on practical needs to adapt to more complex emotional scenarios and application requirements.

However, despite progress in empathic expression, related research has highlighted potential ethical impacts, raising concerns about technology misuse, user privacy protection, and social trust. To ensure proper resolution of ethical and privacy issues, designers must balance enhancing chatbot functionality with protecting user data security and mental health, thereby preventing technology misuse and strengthening user trust in artificial intelligence.

References

Hou, H. C., Ni, S. G., Lin, S. Y., & Wang, P. S. (2024). When AI learns empathy: Themes, scenarios, and optimization of empathy computation from a psychological perspective. *Advances in Psychological Science*, *32*(5), 845–858.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ···& Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596.

https://doi.org/10.1001/jamainternmed.2023.1838

Cai, M., Wang, D., Feng, S., & Zhang, Y. (2024). PECER: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 10631-10635). IEEE. http://10.1109/ICASSP48485.2024.10446914

Cao, H., Zhang, Y., Feng, S., Yang, X., Wang, D., & Zhang, Y. (2025). TOOL-ED: Enhancing empathetic response generation with the tool calling capability of LLM. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5305-5320). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.355/

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1170–1183). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.83/

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? Notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, (pp. 1–18). https://doi.org/10.1145/3613904.3642336

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C., Lampinen, A., Wang, J. X., ···& Schulz, E. (2023). Machine psychology. *arXiv preprint arXiv:2303.13988*. https://doi.org/10.48550/arXiv.2303.13988

Havaldar, S., Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023). Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 202–214). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*, 504-526. https://doi.org/10.1007/s40593-021-00239-1

Hu, M., Chua, X. C. W., Diong, S. F., Kasturiratna, K. S., Majeed, N. M., & Hartanto, A. (2025). AI as your ally: The effects of AI-assisted venting on negative affect and perceived social support. *Applied Psychology: Health and Well-Being*, *17*(1), e12621. https://doi.org/10.1111/aphw.12621

Huang, J. T., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., & Lyu, M. R. (2025). Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans. *Advances in Neural Information Processing Systems*, *37*, 97053–97087. https://doi.org/10.48550/arXiv.2308.03656

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*(4), https://doi.org/10.1037/0022-3514.59.4.730

Koranteng, E., Rao, A., Flores, E., Lev, M., Landman, A., Dreyer, K., & Succi, M. (2023). Empathy and equity: Key considerations for large language model adoption in health care. *JMIR Medical Education*, *9*, e51199. https://doi.org/10.2196/51199

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 431-469). Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.29/

Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2024). Enhancing empathic reasoning of large language models based on psychotherapy models for AI-assisted social support. *Korean Journal of Cognitive Science*, *35*(1), 23–48. https://doi.org/10.19066/cogsci.2024.35.1.002

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*. https://doi.org/10.48550/arXiv.2403.18148

Liang, H., Sun, L., Wei, J., Huang, X., Sun, L., Yu, B., ···& Zhang, W. (2024). Synth-empathy: Towards high-quality synthetic empathy data. *arXiv e-prints*, arXiv-2407. https://doi.org/10.48550/arXiv.2407.21669

Liu, Y., Han, D., Wu, G., & Qiao, B. (2024). KnowDT: Empathetic dialogue generation with knowledge-enhanced dependency tree. *Applied Intelligence*, *54*(17), 8059–8072. https://doi.org/10.1007/s10489-024-05611-x

Liu-Thompkins, Y., Okazaki, S., & Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, *50*(6), 1198–1218. https://doi.org/10.1007/s11747-022-00892-5

Loh, S. B., & Sesagiri Raamkumar, A. (2023). Harnessing large language models' empathetic response generation capabilities for online mental health counselling support. *arXiv e-prints*, arXiv-2310. https://doi.org/10.48550/arXiv.2310.08017

Luo, M., Warren, C. J., Cheng, L., Abdul-Muhsin, H. M., & Banerjee, I. (2024). Assessing empathy in large language models with real-world physician-patient interactions. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 6510-6519). IEEE. https://doi.org/10.1109/BigData62323.2024.10825307

Ma, J., Chen, B., Wang, K., Hu, Y., Wang, X., Zhan, H., & Wu, W. (2024). Emotional contagion and cognitive empathy regulate the effect of depressive symptoms on empathy-related brain functional connectivity in patients with chronic back pain. *Journal of Affective Disorders*, *362*, 459–467. https://doi.org/10.1016/j.jad.2024.07.026

Naik, N., Jenkins, P., Prajapat, S., & Grace, P. (Eds.). (2024). *Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI, July 3-4, 2024, London, UK: The C3AI 2024* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-74443-3

Patil, D. D., Dhotre, D. R., Gawande, G. S., Mate, D. S., Shelke, M. V., & Bhoye, T. S. (2024). Transformative trends in generative ai: Harnessing large language models for natural language understanding and generation. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(4s), 309-319. https://ijisae.org/index.php/IJISAE/article/view/3794

Pan, S., Fan, C., Zhao, B., Luo, S., & Jin, Y. (2024). Can large language models exhibit cognitive and affective empathy as humans? *OSF Preprints*. https://doi.org/10.31219/osf.io/w5rsu

Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., ···& Mihalcea, R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation*, *13*(5), https://doi.org/10.1007/s12559-021-09925-7

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*. https://doi.org/10.48550/arXiv.1811.00207

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1908.10084

Ren, C., Zhang, Y., He, D., & Qin, J. (2024). WundtGPT: Shaping large language models to be an empathetic, proactive psychologist. *arXiv preprint arXiv:2406.15474*. https://aclanthology.org/D19-1410/

Sabour, S., Zheng, C., & Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11229-11237). https://doi.org/10.1609/aaai.v36i10.21373

Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? On the limits of social intelligence in LLMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3762-3780). Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.248/

Shao, M., Basit, A., Karri, R., & Shafique, M. (2024). Survey of different large language model architectures: Trends, benchmarks, challenges. *IEEE Access*, https://doi.org/10.1109/ACCESS.2024.3482107

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human−AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57. https://doi.org/10.1038/s42256-022-00593-2

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). "Do anything now" : Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1671-1685). Association for Computational Linguistics. https://doi.org/10.1145/3658644.3670388

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: A systematic review. *Journal of Medical Internet Research*, e52597. https://doi.org/10.2196/52597

Svikhnushina, E., & Pu, P. (2022). PEACE: A model of key social and emotional qualities of conversational chatbots. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1-29. https://doi.org/10.1145/3531064

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ···& Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-

Wang, J., Xu, C., Leong, C. T., Li, W., & Li, J. (2024). Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. *arXiv preprint arXiv:2401.05928*. https://doi.org/10.48550/arXiv.2401.05928

Welivita, A., & Pu, P. (2024). Are large language models more empathetic than humans?. *arXiv preprint arXiv:2406.05063*. https://doi.org/10.48550/arXiv.2406.05063

Yang, Z., Ren, Z., Yufeng, W., Peng, S., Sun, H., Zhu, X., & Liao, X. (2024). Enhancing empathetic response generation by augmenting LLMs with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*. https://doi.org/10.48550/arXiv.2402.11801

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., & Qin, B. (2023). Is ChatGPT equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*. https://doi.org/10.48550/arXiv.2304.09582

Zhu, J., Jiang, Z., Zhou, B., Su, J., Zhang, J., & Li, Z. (2024). Empathizing before generation: A double-layered framework for emotional support LLM. In *Lecture Notes in Computer Science* (pp. 490–503). Springer Nature Switzerland. https://doi.org/10.1007/978-981-97-8490-5_35

Zhuang, Z., Chen, Q., Ma, L., Li, M., Han, Y., Qian, Y., ···& Liu, T. (2023). Through the lens of core competency: Survey on evaluation of large language

models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)* (pp. 88–109). Chinese Information Processing Society of China. https://aclanthology.org/2023.ccl-2.8/

Footnotes

[1] Youth Mental Health and Crisis Intelligence Intervention Key Laboratory of Philosophy and Social Science of Anhui Province Open Fund Major Project (SYS2024XXX) funding.

[2] BERT is a natural language processing (NLP) model based on the Transformer architecture proposed by Google AI in 2018. It is a pre-trained language model that can be trained on large-scale text data through unsupervised learning and then fine-tuned for specific tasks (such as question answering, sentiment analysis, etc.).

[3] RoBERTa is an NLP model based on the BERT model proposed by Facebook AI in 2019. It adopts a Transformer architecture similar to BERT but optimizes the training process by using more training data, longer training time, and removing some BERT limitations (such as the Next Sentence Prediction task), thereby improving model performance.

Cosine similarity is a metric commonly used in NLP and information retrieval to measure directional similarity between two vectors in multidimensional space. In emotional similarity assessment, cosine similarity is used to calculate the similarity between user input text and generated text in emotional vector space, reflecting their proximity in emotional expression.

EmotionBench is a framework or dataset for evaluating and benchmarking models' (particularly LLMs') emotional understanding and generation capabilities. It aims to systematically measure model performance in processing, recognizing, or generating emotion-related content, typically including empathy, emotional accuracy, and contextual appropriateness. EmotionBench may involve various tasks such as emotion classification, emotion generation, and empathic response in dialogue, used to compare different models or optimize models' emotional intelligence.

The DeepSeek model is a series of open-source large language models developed by the Chinese AI company DeepSeek (Hangzhou Deep Seek Artificial Intelligence Basic Technology Research Co., Ltd.), aiming to provide high-performance AI solutions at low cost and high efficiency, suitable for general dialogue, programming, reasoning, and multimodal tasks. These models are released in open-source form (typically using MIT licenses) and are widely used in research, development, and commercial scenarios, challenging the market dominance of closed-source models such as OpenAI' s GPT-4.

EmpatheticDialogues is an open-source dataset containing approximately 25,000 dialogues, focusing on empathic dialogue research and covering 32

emotion categories. It is widely used for training and evaluating LLMs' empathy capabilities. Its dialogue data supports emotion analysis, model optimization, and mental health applications, though language and cultural limitations should be noted.

Pecer stands for Prompt-enhanced Empathetic Conversation and Evaluation Model. This model aims to generate high-quality empathic dialogue data by combining prompt engineering and reinforcement learning (RL) technology, while being able to evaluate and optimize empathy in dialogue. Pecer is particularly suitable for generating dialogue that meets human emotional expectations and is widely used in emotional support, mental health assistance, and customer service scenarios.

NRC Emotion Lexicon (National Research Council Emotion Lexicon) is an emotion analysis resource developed by the National Research Council Canada, widely used in NLP and affective computing. It is a lexicon containing a large number of words and their emotional annotations, used to identify emotional information in text (such as joy, sadness, anger, etc.). The NRC Emotion Lexicon helps researchers and developers analyze text' s emotional tendencies, emotional intensity, and related emotional features by associating words with specific emotion categories and intensities.

[1] Refers to small-scale neural network models specifically designed for emotion understanding and empathy tasks. Compared to LLMs, SEMs have smaller parameter scales but perform better in fine-grained emotion detection and cognitive nuances (such as emotion cause inference). SEMs are typically trained on specific emotion datasets (such as EmotionBench) and focus on capturing emotional details in dialogue.

[11] A prompting strategy that guides LLMs to reason step-by-step before generating final output, decomposing complex problems into multiple logical steps. COT improves model accuracy in emotion analysis, logical reasoning, and problem-solving by explicitly expressing reasoning processes (such as intermediate steps of emotion inference). In EBG, COT is used to analyze emotion causes and context.

[12] CoE (Chain-of-Empathy Prompting) is a prompt engineering method that incorporates counseling psychology theoretical models into prompt design to optimize LLMs' empathic output. Specifically, CoE guides LLMs to analyze users' emotional states, potential cognitive biases, or psychological needs through a series of structured reasoning steps (similar to chain of thought, COT) before generating responses, thereby producing more empathic, contextually appropriate, and psychologically supportive responses.

Note: Figure translations are in progress. See original paper for figures.