# Prioritizing Therapeutic Alignment Over Simulated Artificial Intelligence Empathy

Soraya S Anvari[1,*,†], Rina R Wehbe[1]

[1]Department of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

**Abstract**

The global mental health system faces a crisis with a shortage of professional resources. One consequence of this gap is the emergence of Large Language Models (LLMs) as a potential alternative for accessible psychological support which is largely driven by the models' perceived emotional intelligence. While studies show LLMs can surpass human benchmarks in generating empathetic responses, with GPT-4 receiving 31% higher ratings, their success is often limited to linguistic style. In a clinical context, empathy is a strategic tool for setting boundaries and managing crises; therefore, our paper argues that prioritizing maximum empathy in LLM design risks producing a persuasive yet shallow simulation of care. We highlight how the tendency of models to agree with users can validate harmful delusions and how a polite tone may cause a failure to detect crisis language. Furthermore, we address systemic biases that result in lower empathy scores for Black and LGBTQ2S++ users. We conclude that although LLM adoption is increasingly common due to their accessibility, current designs lack the clinical grounding required for safe intervention.

**Keywords**

Large Language Models, Mental Health, Empathy-Centric Design, AI Safety, HCI for Health

## 1. Introduction

The global mental health system faces a shortage of trained professionals as the demand for psychological support continues to increase. As of 2019, an estimated 970 million individuals lived with a mental disorder, yet the deficit in trained practitioners remains a primary barrier to care [1]. In this context, LLM-based systems have been positioned as scalable alternatives or supplements to human support. For many users, these models serve as a "five-minute therapist," offering affordability, accessibility, and anonymity that traditional systems might not provide [2, 3, 4, 5]. Mobile apps, in particular, have incorporated LLMs to fulfill various purposes in mental health care, including roles as therapeutic tools, personal coaches, and mindfulness guides designed to support user well-being [6, 7, 8, 4, 9].

Previous research illustrates that LLMs have the potential to outperform humans in linguistic empathy tasks [10, 11, 12], but the ability to simulate an empathetic tone does not necessarily translate to therapeutic efficacy [13]. Although therapeutic empathy is a strategic approach for facilitating behavioral change and crisis de-escalation, LLMs are mainly designed to be agreeable and user-friendly [14, 15]. This results in algorithmic over-compliance, where the algorithm automatically validates the user's beliefs, even if they are maladaptive or delusional, to simulate empathy [14].

Concerns about safety are further complicated by inequities in model performance. LLMs have been shown to generate empathy scores that are 2-13% lower for Black users compared to other groups [3]. LGBTQ2S++ users, who often turn to digital tools due to stigma or limited access to care, report that models frequently miss culturally specific stressors and contextual nuances [16]. Technical solutions such as fine-tuning are commonly proposed, yet they may not sufficiently address deeper representational biases embedded in training data [16, 17].

Currently, most studies are centered on the technical viability or short-term user satisfaction of LLM-driven tools, without considering the underlying structural dangers of empathetic imitation. Our

*EmpathiCH'26 Workshop Co-located with CHI'26 Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain*

soraya.anvari@dal.ca (S. S. Anvari); rina.wehbe@dal.ca (R. R. Wehbe)

https://hci4good.cs.dal.ca/Team/1 (S. S. Anvari); https://rinawehbe.ca/ (R. R. Wehbe)

0009-0006-9420-446X (S. S. Anvari); 0000-0003-3677-5185 (R. R. Wehbe)

position paper aims to fill this gap by analyzing the dangers of maximal empathy in Artificial Intelligence (AI) development. The politeness bias may hide red flag communications and symptoms of a crisis, leading to a failure of clinical escalation. We will examine how the current design of LLMs lacks the human stakes and clinical foundations necessary for the safe treatment of mental health challenges [18, 15].

## 2. The Risk of Constant Agreement

The reason for the strong performance of LLMs on empathy tasks is, in many cases, their capacity to replicate language patterns rather than a real understanding of human emotions. Although models such as GPT-4 demonstrate a substantial improvement in empathetic responses compared to human performance [10], this achievement is limited to cognitive empathy, which is the capacity to recognize and label emotions, as opposed to affective empathy, which is necessary for a real therapeutic relationship [11].

Since these algorithms are trained using reinforcement learning from human feedback to be useful and polite, they tend to focus more on user agreement than on medical facts [14]. In the field of professional therapy, empathy is more than just being "nice". While the term is often operationally defined in social AI as being positive and trustworthy, clinical empathy is a strategic approach used to confront a patient's distorted thinking and help them change their behavior [13, 19].

When an AI is too consistent in its agreement with a user, it leads to collusion that can be dangerous. For instance, if a user holds a maladaptive or delusional belief, a collusive model might agree with the statement in order to uphold its supportive character [14, 20]. In the absence of a licensed professional, who is accountable for the well-being of the patient, the AI's empathy is a superficial imitation that prefers conversational agreement over actual safety.

## 3. The Safety Blind Spot

When designers prioritize a kind and agreeable tone, they inadvertently create a "safety blind spot." Because LLMs are trained to be helpful and non-confrontational, they develop a politeness bias that can be life-threatening in a crisis. An example is the NEDA/Tessa chatbot incident [1], where a chatbot meant to help people with eating disorders began providing harmful weight-loss tips. Tessa was originally a rule-based program, but after generative AI features were introduced, the bot bypassed its clinical safety rules and endorsed dangerous behaviors in an apparent attempt to remain supportive of the user's stated goals.

Beyond these conversational failures, there is a deeper technical risk called persona drift [21]. Research shows that a model's persona can drift over the course of a conversation, gradually departing from its intended character without warning. For example, a model could shift from a supportive clinical tone toward boundary-violating behaviors as a dialogue extends [21, 22]. This unpredictability means that a bot's empathy is not a stable feature; it is a temporary mask. In a high-stakes mental health scenario, this instability poses a catastrophic risk, as a bot might fail to trigger an emergency referral. Without a foundation in established clinical rules, the attempt to be nice replaces the obligation to be safe [18, 3, 15].

## 4. Designing for Therapeutic Safety

To address these risks, the Human Computer Interaction (HCI) community should shift its focus from maximizing perceived empathy to ensuring therapeutic alignment. A possible component of this shift is the implementation of **Retrieval-Augmented Generation (RAG)**. By grounding LLM responses in validated clinical manuals, such as the DSM-5 [23], ICD-11 [24], or established Cognitive Behavioral

[1] https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-

Therapy (CBT) protocols, designers can prevent the therapeutic hallucinations that occur when a model prioritizes agreeableness over evidence-based practice [18].

Furthermore, we propose the adoption of **human-in-the-loop** strategies for high-stakes interactions [25, 26]. In these systems, AI is used strictly for low-intensity support (e.g., psychoeducation) while maintaining a pathway to a human clinician if risk is detected. By utilizing evaluative frameworks like READI (Readiness Evaluation for AI-Mental Health Deployment) [27], researchers can ensure that safety and accountability are the primary metrics for success. Although our focus is on clinical safety, these issues can be considered in other areas, such as education or human-robot interaction. In these areas, empathy can be leveraged as a means to engage users, possibly concealing issues of a similar nature in terms of pedagogical or assistive accuracy.

Designers should consider that a therapeutic relationship is often built not through constant validation, but through the balanced application of empathy to challenge and validate a patient [14]. Moreover, it is important to address the systemic issues of bias in the deployment of empathy to marginalized communities [3, 16]. Instead of focusing on technical improvements such as fine-tuning, which may be constrained by the data used, we propose a socio-technical approach that takes into account the cultural and social context of the user [16, 17].

Additionally, the move towards therapeutic alignment also requires a re-evaluation of user consent. Traditional static consent models are insufficient for the dynamic nature of LLM interactions. We propose the concept of **contextual consent**, where the boundaries of data use are renegotiated as the therapeutic depth of the conversation changes [28, 29, 30]. Implementing contextual consent ensures that users remain aware of how their sensitive disclosures are processed and stored after the initial agreement, preventing the privacy risks associated with misplaced trust in a simulated empathetic persona [15]. The following table outlines the proposed transition from empathy-centric to safety-centric design principles:

**Table 1**
Transitioning from Empathy-Centric to Safety-Centric Design Principles

| Design Aspect | Current Approach | Proposed Direction |
|---|---|---|
| **Core Objective** | User Comfort & Retention | Clinical Integrity & Safety |
| **Interaction Style** | Constant Agreement | Constructive Confrontation |
| **Risk Management** | Keyword Filters | Evidence-Based Grounding (RAG) |
| **Human Oversight** | Fully Automated Support | Human-in-the-Loop |
| **Consent Model** | Static Initial Agreement | Contextual Consent |
| **Evaluation Metric** | User-Rated Empathy | Accuracy of Clinical Referrals |

Future studies could investigate the idea of empathy-neutral or clinical-first interfaces. These interfaces would focus on crisis identification and safety before the need to simulate human compassion, ensuring that the technology serves as an aid and not a danger to replace human professionals [14]. By setting up proper safety guidelines and basing AI on clinical principles, the HCI community can create technology that helps mental health literacy without endangering the well-being of users [3, 31].

## 5. Conclusion

Our position paper has explored the paradox of synthetic empathy in digital mental health. Although LLMs have shown an ability to outperform humans in linguistic empathy tasks, this achievement is frequently the result of computational politeness rather than genuine therapeutic insight. We conclude that current LLMs lack the human accountability required for high-stakes therapy. For the HCI community, this suggests a need to re-evaluate the goal of maximizing perceived empathy in high-stakes contexts. Instead, we propose a design ethos centered on therapeutic alignment that values clinical safety over the simulation of compassion. We invite the community to consider: if empathy becomes a material property of infrastructure rather than a human quality, what new forms

of accountability and governance must emerge alongside it?

## Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] World Health Organization, World mental health report: Transforming mental health for all, World Health Organization, Geneva, 2022. URL: https://www.who.int/publications/i/item/9789240049338.

[2] I. Song, S. R. Pendse, N. Kumar, M. De Choudhury, The typing cure: Experiences with large language model chatbots for mental health support, Proc. ACM Hum.-Comput. Interact. 9 (2025). URL: https://doi.org/10.1145/3757430. doi:10.1145/3757430.

[3] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, M. Ghassemi, Can AI relate: Testing large language model response for mental health support, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2206–2221. URL: https://aclanthology.org/2024.findings-emnlp.120/. doi:10.18653/v1/2024.findings-emnlp.120.

[4] A. H. Y. Chan, M. L. L. Honey, User perceptions of mobile digital apps for mental health: Acceptability and usability—an integrative review, Journal of Psychiatric and Mental Health Nursing 29 (2022) 147–168. URL: https://doi.org/10.1111/jpm.12744. doi:10.1111/jpm.12744.

[5] K. K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial, JMIR mental health 4 (2017) e19. doi:10.2196/mental.7785.

[6] S. S. Anvari, S. Ppali, D. Hernandez, R. R. Wehbe, When your therapist is an algorithm: Understanding the role of ai in mental health mobile applications, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26, Association for Computing Machinery, New York, NY, USA, 2026. URL: https://doi.org/10.1145/3772318.3791326. doi:10.1145/3772318.3791326.

[7] A. Baumel, F. Muench, S. Edan, J. M. Kane, Objective user engagement with mental health apps: Systematic search and panel-based usage analysis, J Med Internet Res 21 (2019) e14567. URL: http://www.jmir.org/2019/9/e14567/. doi:10.2196/14567.

[8] O. Oyebode, F. Alqahtani, R. Orji, Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews, IEEE Access 8 (2020) 111141–111158. doi:10.1109/ACCESS.2020.3002176.

[9] K. Stawarz, A. L. Cox, A. Blandford, Beyond Self-Tracking and Reminders: Designing Smartphone Apps That Support Habit Formation, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 2653–2662. URL: https://doi.org/10.1145/2702123.2702230. doi:10.1145/2702123.2702230.

[10] A. Welivita, P. Pu, Are large language models more empathetic than humans?, 2024. doi:10.48550/arXiv.2406.05063. arXiv:2406.05063.

[11] V. Sorin, D. Brin, Y. Barash, E. Konen, A. Charney, G. Nadkarni, E. Klang, Large language models and empathy: Systematic review, J Med Internet Res 26 (2024) e52597. URL: https://www.jmir.org/2024/1/e52597. doi:10.2196/52597.

[12] K. Muthukumar, Empathy AI in healthcare, Frontiers in Psychology 16 (2025) 1680552. URL: https://doi.org/10.3389/fpsyg.2025.1680552. doi:10.3389/fpsyg.2025.1680552.

[13] C. R. Rogers, The necessary and sufficient conditions of therapeutic personality change., Psychotherapy: Theory, Research, Practice, Training 44 (2007) 240–248. doi:10.1037/0033-3204.44.3.240.

[14] J. Moore, D. Grabb, W. Agnew, K. Klyman, S. Chancellor, D. C. Ong, N. Haber, Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers., in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 599–627. URL: https://doi.org/10.1145/3715275.3732039. doi:10.1145/3715275.3732039.

[15] S. S. Anvari, R. R. Wehbe, Therapeutic ai and the hidden risks of over-disclosure: An embedded ai-literacy framework for mental health privacy, 2025. URL: https://arxiv.org/abs/2510.10805. arXiv:2510.10805.

[16] Z. Ma, Y. Mei, Y. Long, Z. Su, K. Z. Gajos, Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642482. doi:10.1145/3613904.3642482.

[17] A. Malik, N. Sabri, M. M. Karnaze, M. ElSherief, Are LLMs empathetic to all? investigating the influence of multi-demographic personas on a model's empathy, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 24938–24959. URL: https://aclanthology.org/2025.findings-emnlp.1358/. doi:10.18653/v1/2025.findings-emnlp.1358.

[18] H. Hu, Y. Zhou, J. Si, Q. Wang, H. Zhang, F. Ren, F. Ma, L. Cui, Q. Tian, Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling, 2025. doi:https://doi.org/10.48550/arXiv.2505.15715.

[19] W. Kidder, J. D'Cruz, K. R. Varshney, Empathy and the right to be an exception: What llms can and cannot do, 2024. doi:10.48550/arXiv.2401.14523. arXiv:2401.14523.

[20] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, M. R. Lyu, Apathetic or empathetic? evaluating llms' emotional alignments with humans, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Curran Associates Inc., Red Hook, NY, USA, 2024.

[21] C. Lu, J. Gallagher, J. Michala, K. Fish, J. Lindsey, The assistant axis: Situating and stabilizing the default persona of language models (2026). URL: https://arxiv.org/abs/2601.10387.

[22] Y. Cheng, Z. Kang, K. Jiang, C. Sun, Q. Pan, The slow drift of support: Boundary failures in multi-turn mental health llm dialogues (2026). URL: https://arxiv.org/abs/2601.14269.

[23] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed., text rev. ed., American Psychiatric Association, Washington, DC, 2022. doi:10.1176/appi.books.9780890425787.

[24] World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 11th ed. ed., World Health Organization, 2019. URL: https://icd.who.int/.

[25] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Generation Computer Systems 135 (2022) 364–381. URL: https://www.sciencedirect.com/science/article/pii/S0167739X22001790. doi:https://doi.org/10.1016/j.future.2022.05.014.

[26] L. Chan, Y.-C. Liao, G. B. Mo, J. J. Dudley, C.-L. Cheng, P. O. Kristensson, A. Oulasvirta, Investigating positive and negative qualities of human-in-the-loop optimization for designing interaction techniques, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3501850. doi:10.1145/3491102.3501850.

[27] E. C. Stade, J. C. Eichstaedt, J. P. Kim, S. W. Stirman, Readiness evaluation for artificial intelligence-mental health deployment and implementation (readi): A review and proposed framework, Technology, Mind, and Behavior 6 (2025). URL: https://doi.org/10.1037/tmb0000163.

doi:10.1037/tmb0000163.
[28] M. Dunn, Contextualising consent, Journal of Medical Ethics 42 (2016) 67–68. URL: https://jme.bmj.com/content/42/2/67. doi:10.1136/medethics-2016-103381.
[29] J. Kleinig, The ethics of consent, Canadian Journal of Philosophy 12 (1982) 91–118. doi:10.1080/00455091.1982.10715825.
[30] O. Corrigan, Empty ethics: the problem with informed consent, Sociology of Health & Illness 25 (2003) 768–792. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1467-9566.2003.00369.x. doi:https://doi.org/10.1046/j.1467-9566.2003.00369.x.
[31] S. S. Anvari, J. Hammer, R. R. Wehbe, "more than just a game, it's an app that builds awareness around mental health": Mental health stigma reduction using games for change, Proc. ACM Hum.-Comput. Interact. 8 (2024). URL: https://doi.org/10.1145/3677090. doi:10.1145/3677090.