Mental Health Chatbots: Effects of Empathic Tone on Perceived Empathy and Trust

Mojgan Moshirpour[1,*], Mohammad Moshirpour[2] and Qiao Sun[3]

[1]Cumming School of Medicine, University of Calgary, Calgary, Canada
[2]Department of Informatics, University of California, Irvine, USA
[3]Schulich School of Engineering, University of Calgary, Calgary, Canada

Abstract

Empathy is frequently positioned as a desirable quality in mental-health chatbots, yet most evaluations conflate therapeutic content with the way that content is delivered. This position paper proposes a controlled study that isolates empathic tone—warm, validating, and contextually responsive phrasing—while holding coping strategies constant. We outline a scenario-based between-subjects experiment comparing an empathic versus neutral conversational style, measuring perceived empathy with the Perceived Empathy of Technology Scale (PETS) and trust with a validated human–computer trust scale. We test whether emotional validation mediates tone's effects on perceived empathy, and whether prior chatbot experience moderates tone's effects on trust. We argue that "empathic tone" is not merely a linguistic flourish but a sociomaterial performance shaped by interface cues, system claims, and users' expectations; therefore, tone should be treated as a first-class design variable in empathy-centric HCI. We close with workshop provocations on how to design and evaluate responsible empathic tone in AI systems deployed in emotionally sensitive contexts.

Keywords

empathic tone, mental health chatbots, perceived empathy, emotional validation, trust, sociomateriality, PETS

1. Introduction

Conversational agents are increasingly used to provide low-intensity mental-health support (e.g., psychoeducation, symptom tracking, CBT-inspired coping strategies), motivated in part by the shortage of mental-health professionals and barriers to accessing care [1, 2]. While controlled trials and reviews show potential benefits in some contexts, the evidence remains mixed and heterogeneous, with ongoing concerns about safety, fit, and the quality of the user experience [2, 3, 4]. In emotionally sensitive interactions, how a system responds can matter as much as what it says: users routinely interpret tone, pacing, and responsiveness as signals of "being understood," and these signals shape trust, disclosure, and sustained engagement [5, 6, 7]. Yet empathy is rarely treated as something a system does through language alone. It is co-produced: assembled from word choice, interface timing, system identity claims, and the expectations users bring into the interaction [11, 12]. This sociomaterial character of empathic communication is precisely what makes it difficult to study, and what most existing evaluations of mental-health chatbots fail to disentangle.

This submission responds to the (Re-)thinking Empathy's Materiality in HCI workshop by treating empathic tone not as a stylistic add-on but as a sociomaterial design variable: one whose effects are inseparable from the interfaces, roles, and contexts through which it is delivered [12]. We argue that isolating tone experimentally—holding therapeutic content constant while manipulating tonal delivery—is not in tension with a sociomaterial perspective but is required by it: only by separating tone from content can we begin to understand which aspects of the user experience are shaped by language itself versus by the broader material assemblage in which language is embedded.

EmpathiCH'26 Workshop Co-located with CHI'26 Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain
*Corresponding author.
mojganmoshirpour@ucalgary.ca (M. Moshirpour); mmoshirp@uci.edu (M. Moshirpour); qsun@ucalgary.ca (Q. Sun)

2. Background: Empathy and Tone in Mental-Health Chatbots

Empathy in interactive agents has been conceptualized as a multidimensional construct including cognitive empathy (understanding a user's mental state), affective empathy (emotional resonance), empathic concern (motivation to help), perspective-taking, and emotional validation [8, 9]. Contemporary mental-health chatbots implement subsets of these dimensions through emotion detection, reflective statements, and language models trained on supportive dialogue [2, 9, 10]. However, reviews highlight a persistent "authenticity gap": systems may recognize coarse emotion labels and generate supportive phrases, yet produce generic, mismatched, or context-insensitive responses that feel scripted or superficial [2, 5, 6].

We use the term *empathic tone* to refer to the warm, validating, and contextually appropriate manner of responding that signals care, e.g., reflective listening ("It sounds like you've been carrying this alone"), gentle curiosity ("What part felt hardest?"), and supportive follow-ups, rather than affectively restrained, solution-focused phrasing. Crucially, tone is not reducible to linguistic content: the same words land differently depending on the interface affordances surrounding them (e.g., typing indicators, memory, personalization), on system identity claims (e.g., 'I'm here for you'), and on the situational stakes (e.g., casual stress versus active crisis disclosure). Tone is therefore a distributed performance, assembled across text, UI cues, and context, rather than a property of any single message [7, 11]. Because the material assemblage shapes how tone is interpreted, studies that conflate tone with content cannot isolate which elements produce feelings of being understood. That gap is what our proposed experiment addresses.

3. Research Objective and Questions

Our objective is to isolate and test the effect of empathic tone on user perceptions while holding therapeutic content constant. Rather than only asking whether tone 'works,' we ask *how* it works, for whom, and which micro-features shape judgments of authenticity.

RQ1 (Mechanism): To what extent does empathic tone enhance perceived empathy, and is this relationship mediated by perceived emotional validation?

RQ2 (Boundary conditions): Does the effect of empathic tone on trust vary by users' prior experience with mental-health chatbots?

RQ3 (Features): What linguistic and conversational features differentiate perceived authentic empathy from perceived scripted empathy in chatbot interactions?

4. Methods: A Controlled Tone Manipulation

We propose a scenario-based, between-subjects study in which participants interact with a mental-health support chatbot under one of two conditions: (1) *empathic tone* (validation, reflective listening, supportive follow-up questions) or (2) *neutral tone* (objective, solution-focused, affectively restrained language). Both conditions deliver identical coping strategies and include a standardized limitations footer indicating the chatbot is not a professional service.

Empathic tone will be operationalized using established frameworks for supportive communication [9], developed as scripted chatbot turns that include: (a) reflective listening (e.g., "It sounds like you've been carrying a lot lately"), (b) validation of the user's emotional experience (e.g., "That makes a lot of sense given what you're going through"), and (c) open, curious follow-up questions (e.g., "What part of that felt hardest?"). Neutral-condition turns will convey the same coping content using direct, solution-focused phrasing without affective acknowledgement. All scripts will be developed and reviewed by the research team to ensure tonal differentiation is consistent across turns.

To preserve internal validity, conversations will be scripted or tightly templated so that content is constant (e.g., the same grounding exercise, cognitive reframing prompt, or behavioral activation suggestion) while phrasing differs only in tonal features (warmth, validation, perspective-taking, and

follow-up). A manipulation check will be included immediately after the interaction: participants will rate the degree to which the chatbot felt warm, caring, and understanding on a set of bipolar items (e.g., warm–cold, caring–indifferent), allowing us to verify that the empathic condition was perceived as meaningfully more empathic than the neutral condition before proceeding with the primary analyses. Participants will be recruited from an online pool with eligibility criteria designed to reduce risk (e.g., excluding individuals reporting active crisis or suicidal ideation at screening, using validated screeners such as the PHQ-2). All participants will receive crisis resource information regardless of condition, and a standardized safety footer will appear in every chatbot interaction. A target sample of approximately 50 participants (25 per condition) is planned, informed by power analysis for mediation (PROCESS; $\alpha = .05$, power = .80) using effect sizes from prior tone manipulation studies. The study will be reviewed under institutional ethics procedures prior to data collection.

4.1. Measures

Perceived empathy is the primary outcome, measured using the Perceived Empathy of Technology Scale (PETS), a validated 10-item instrument [13]. Perceived emotional validation is measured using three items adapted from social support research (e.g., "The chatbot understood how I was feeling," "I felt my emotions were acknowledged"), rated on a 5-point Likert scale, to test mediation in RQ1. Trust is measured using Gulati et al.'s human–computer trust scale [14], complemented by established trust-in-automation factors (reliability expectations, perceived safeguards) [18]. Prior chatbot experience is assessed using a granular scale capturing frequency of use (never, once or twice, occasionally, regularly), type of chatbot used (general-purpose, mental-health specific, other), and self-reported comfort with chatbot interactions, to enable nuanced moderation testing in RQ2. Open-ended questions for RQ3 ask participants to describe moments that felt empathic or inauthentic and explain what specific features led to those interpretations. Asking participants directly about perceived features complements the scripted dialogue by capturing users' own attributions and language, which may differ from researcher-coded dialogue features [15]; together these data sources capture both how users experience and describe empathy in their own words, and how those descriptions map onto the features built into the scripts.

4.2. Analysis Plan

RQ1 will be tested using mediation analysis (PROCESS) [16] examining whether empathic tone increases perceived empathy indirectly through emotional validation. RQ2 will be tested with moderation analysis; significant interactions will be probed using simple slopes [16, 17]. For RQ3, reflexive thematic analysis will identify conversational features associated with perceived authenticity versus scriptedness.

5. Why Tone Matters for Empathy's Materiality

A sociomaterial account of empathy [12] holds that technology, users, and context are constitutively entangled: what tone means and does cannot be read off from linguistic analysis alone. This creates a productive tension for experimental research. Our controlled design deliberately holds the material assemblage constant—same interface, same scenario, same coping content—so that observed differences in perceived empathy and trust are most plausibly attributed to tonal delivery, though we acknowledge that uncontrolled individual differences place limits on strong causal inference. The results will be partial by design: they tell us what tone contributes when the assemblage is fixed, not how tone, interface, and context interact in naturalistic use. Both insights matter. The controlled result establishes a causal baseline; the open-ended RQ3 data surface how participants interpret tone in relation to the broader system.

Three specific entanglements make this framing actionable for design. First, tonal cues are read in relation to interface signals: a warm phrase may register as sincere or patronizing depending on whether the system demonstrates memory, appropriate pacing, or consistent persona. Second, users actively negotiate the chatbot's relational role—tool, peer, coach, quasi-therapist—and empathic tone

is one of the strongest role-attribution cues available. Those attributions shape trust calibration and risk: users who read warmth as evidence of clinical competence may over-rely on a system that cannot safely bear that weight [7, 11, 14]. Third, tone is culturally and contextually situated: what registers as validating in one setting may feel presumptuous or intrusive in another, and meta-analytic evidence from therapist empathy research suggests that cultural fit moderates empathy's effects on outcomes [19]. This is a genuine limitation of our controlled design, one our thematic analysis can begin to surface. In short, our experiment is not a reductionist move that ignores materiality—it is a deliberate epistemic step that uses controlled isolation to generate the kind of causal evidence that sociomaterial critique then needs to interrogate and extend.

6. Workshop Provocations

We offer the following provocations to stimulate discussion and taxonomy building:

P1: Empathic tone is a sociomaterial performance, not a feature. Its effects depend on how language interacts with interface cues, system claims, and user expectations.

P2: Emotional validation may be the key mechanism linking tone to perceived empathy, but validation can also backfire if it implies clinical certainty or 'false understanding.'

P3: Prior experience with chatbots may shift what users treat as authentic. New users might be more influenced by warmth; experienced users may discount scripted empathy and prioritize transparency.

P4: Responsible empathic tone requires repair moves (e.g., acknowledging misunderstanding, clarifying questions) and visible uncertainty, not only warm phrasing.

7. Expected Contribution

This proposed study will contribute experimental evidence on whether empathic tone enhances perceived empathy and trust. By testing emotional validation as a mediator, it will clarify a psychological pathway from interactional style to perceived support. By examining prior chatbot experience, it will identify boundary conditions that can guide when empathic tone is beneficial versus when transparency or reliability dominates. Open-ended responses will generate feature-level design recommendations sensitive to the materiality of interaction.

8. Conclusion

Mental health chatbots are widespread, but what makes them feel supportive versus scripted is still unclear. We argue that empathic tone should be treated as a first-class design variable: not a surface flourish, but an interactional cue that shapes perceived understanding and care. The proposed study examines tone's contribution to perceived empathy and trust by holding therapeutic content and interface context constant while comparing empathic versus neutral delivery. We test emotional validation as a mediator of tone's effects on perceived empathy, examine prior chatbot experience as a boundary condition for trust, and use reflexive thematic analysis to identify the linguistic and conversational features users associate with authentic versus scripted empathy. Together, these results can provide actionable guidance for empathy-centric design and help build evaluation frameworks that combine causal clarity with attention to how empathy is interpreted in real interaction contexts.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to assist with drafting, editing, and refining the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. K. Fitzpatrick, A. Darcy, and 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(2):e19, 2017. https://doi.org/10.2196/mental.7785

[2] A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 22(7):e16021, 2020. https://doi.org/10.2196/16021

[3] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018. https://doi.org/10.1093/jamia/ocy072

[4] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464, 2019. https://doi.org/10.1177/0706743719828977

[5] C. Q. Jiang, Y. Zhang, and W. Pian. Chatbot as an emergency exit: Mediated empathy for resilience via human–AI interaction during the COVID-19 pandemic. *Information Processing & Management*, 59(5):103074, 2022. https://doi.org/10.1016/j.ipm.2022.103074

[6] M. Kang, E. Aizawa, S. Kim, and A. G. Dunn. "This app said I had severe depression, and now I don't know what to do": The unintentional harms of mental health applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 2024. https://doi.org/10.1145/3613904.3642178

[7] B. Cuff, S. Brown, L. Taylor, and D. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016. https://doi.org/10.1177/1754073914558466

[8] A. Brännström, J. Wester, and J. C. Nieves. A formal understanding of computational empathy in interactive agents. *Cognitive Systems Research*, 82:101203, 2023. https://doi.org/10.1016/j.cogsys.2023.101203

[9] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.425

[10] D. C. Ong, A. Goldenberg, and M. Inzlicht. AI-generated empathy: Opportunities, limits, and future directions. *PsyArXiv*, 2025. https://doi.org/10.31234/osf.io/8n5jw

[11] P. Wright and J. McCarthy. Empathy and experience in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, pages 637–646. ACM, 2008. https://doi.org/10.1145/1357054.1357156

[12] W. J. Orlikowski and S. V. Scott. Sociomateriality: Challenging the separation of technology, work and organization. *Academy of Management Annals*, 2(1):433–474, 2008. https://doi.org/10.1080/19416520802211644

[13] D. Schmidmaier, J. Rupp, D. Cvetanova, and S. Mayer. Perceived Empathy of Technology Scale (PETS): Measuring empathy of systems toward the user. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 2024. https://doi.org/10.1145/3613904.3642035

[14] S. Gulati, S. Sousa, and D. Lamas. Design, development and evaluation of a human–computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019. https://doi.org/10.1080/0144929X.2019.1656779

[15] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. https://doi.org/10.1191/1478088706qp063oa

[16] A. F. Hayes. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (2nd ed.). Guilford Press, 2018.

[17] R. M. Baron and D. A. Kenny. The moderator–mediator variable distinction in social psychological research. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986. https://doi.org/10.

1037/0022-3514.51.6.1173

[18] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. https://doi.org/10.1518/hfes.46.1.50.30392

[19] R. Elliott, A. C. Bohart, J. C. Watson, and D. Murphy. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399–410, 2018. https://doi.org/10.1037/pst0000175