Review

Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review

Ruvini Sanjeewa[1], BSc; Ravi Iyer[1], PhD; Pragalathan Apputhurai[1], PhD; Nilmini Wickramasinghe[2], PhD; Denny Meyer[1], PhD

[1]School of Health Sciences, Swinburne University of Technology, Hawthorn, Australia

[2]School of Computing, Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia

Corresponding Author:
Ruvini Sanjeewa, BSc
School of Health Sciences
Swinburne University of Technology
PO Box 218
John Street
Hawthorn, 3122
Australia
Phone: 61 422587030
Email: rsanjeewa@swin.edu.au

Abstract

Background: The demand for mental health (MH) services in the community continues to exceed supply. At the same time, technological developments make the use of artificial intelligence–empowered conversational agents (CAs) a real possibility to help fill this gap.

Objective: The objective of this review was to identify existing empathic CA design architectures within the MH care sector and to assess their technical performance in detecting and responding to user emotions in terms of classification accuracy. In addition, the approaches used to evaluate empathic CAs within the MH care sector in terms of their acceptability to users were considered. Finally, this review aimed to identify limitations and future directions for empathic CAs in MH care.

Methods: A systematic literature search was conducted across 6 academic databases to identify journal articles and conference proceedings using search terms covering 3 topics: "conversational agents," "mental health," and "empathy." Only studies discussing CA interventions for the MH care domain were eligible for this review, with both textual and vocal characteristics considered as possible data inputs. Quality was assessed using appropriate risk of bias and quality tools.

Results: A total of 19 articles met all inclusion criteria. Most (12/19, 63%) of these empathic CA designs in MH care were machine learning (ML) based, with 26% (5/19) hybrid engines and 11% (2/19) rule-based systems. Among the ML-based CAs, 47% (9/19) used neural networks, with transformer-based architectures being well represented (7/19, 37%). The remaining 16% (3/19) of the ML models were unspecified. Technical assessments of these CAs focused on response accuracies and their ability to recognize, predict, and classify user emotions. While single-engine CAs demonstrated good accuracy, the hybrid engines achieved higher accuracy and provided more nuanced responses. Of the 19 studies, human evaluations were conducted in 16 (84%), with only 5 (26%) focusing directly on the CA's empathic features. All these papers used self-reports for measuring empathy, including single or multiple (scale) ratings or qualitative feedback from in-depth interviews. Only 1 (5%) paper included evaluations by both CA users and experts, adding more value to the process.

Conclusions: The integration of CA design and its evaluation is crucial to produce empathic CAs. Future studies should focus on using a clear definition of empathy and standardized scales for empathy measurement, ideally including expert assessment. In addition, the diversity in measures used for technical assessment and evaluation poses a challenge for comparing CA performances, which future research should also address. However, CAs with good technical and empathic performance are already available to users of MH care services, showing promise for new applications, such as helpline services.

(JMIR Ment Health 2024;11:e58974) doi: 10.2196/58974

KEYWORDS

conversational agents; chatbots; virtual assistants; empathy; emotionally aware; mental health; mental well-being

Introduction

Background

An escalation in mental health (MH) diagnoses in the community, inadequate facilities, and a MH care workforce that does not meet demand are placing extraordinary pressures on an already strained system [1]. This service gap creates a significant opportunity for MH care interventions, enhanced using recent advances in modern technologies. Conversational agent (CA) platforms using artificial intelligence (AI) via machine learning (ML) techniques have emerged within the MH care domain, providing additional functionalities and support to address this gap [2]. Examples of CAs that use ML include Woebot, providing cognitive behavioral therapy [3]; Wysa, providing MH support by checking depressive symptoms [4]; Saarthi, trained to provide personalized and empathic support to patients via therapeutic techniques [5]; and Empathetic Research IoT Network, a chatbot that provides access to MH resources for students in need [6]. However, the lack of acceptance of CAs in the MH domain remains a barrier to the uptake of these innovations, and the lack of empathy often displayed by CAs contributes to end-user mistrust [7].

Empathy in patient care has been defined by the World Health Organization as an understanding of the patient's experiences, concerns, and perspectives, combined with a capacity to communicate this understanding and an intention to help [8]. Counselor empathy is an essential feature that enhances therapeutic outcomes for patients and can be measured via therapeutic alliance [9,10]. The same is true for CA-human interactions, where empathy exhibited by a CA system helps build rapport, encouraging users to more frequently engage with the CA system [11]. Contextual awareness, which allows CAs to respond to a user's current emotional situation when suggesting appropriate interventions, also facilitates empathic CA communication [12]. Both trustworthiness of the CA (as perceived by the user) and contextual awareness of the user's situation (as detected by the CA) are, therefore, important considerations when building an empathic CA. Empathy serves to enhance the bidirectional interaction between the CA and the end user [13].

Assessment of the effectiveness of CA platforms has received little attention in the MH care sector [14]. For the impact of these systems to be fully realized, these platforms need to meet the requirements of end users, which suggests a key role for lived experience and coproduction. The validity and reliability of these new digital technologies also need to be reviewed by MH care decision-makers and professionals to ensure successful integration in the sector [15]. In addition, evaluations need to assess the ability of such platforms to reduce symptoms of mental illness [16] while also enhancing user well-being and ensuring that patients feel understood [13]. However, any such evaluation needs to be conducted in the context of the role envisaged for the CA, considering the success of the bidirectional interaction described earlier.

While there are existing reviews exploring the efficacy of CAs designed for MH care [10,17,18], to our knowledge, this is the first review to specifically examine how these empathic CAs are designed and evaluated. A comprehensive systematic review and meta-analysis of AI-based CAs for promoting MH was conducted by Li et al [17], with a focus on the intervention and technical characteristics of effective CAs. The effectiveness of the CA designs was captured through user feedback. The meta-analysis explored the role of the CA, AI techniques, and delivery platforms that contributed to the success of these designs. In a similar review, Gaffney et al [18] targeted CA interventions for treating MH problems, with a specific focus on user experience outcomes as measures of efficacy. Another such study explored the evidence of effectiveness with regard to improving symptoms of MH conditions [19]. A critical finding of this review was that empathic response and personalization were significant facilitators of efficacy in these systems. However, the incorporation of this crucial empathy component within CAs has not been studied in any depth within the MH sector. Existing reviews have tended to focus on the inability of CAs to respond to unexpected user inputs rather than their ability to demonstrate empathy [19].

Objectives

This review aimed to assess the types of CA designs found in the MH care sector that are specifically tailored to convey empathy. It also aimed to describe the methods used to evaluate these empathic designs from a technical and implementation perspective. Therefore, this review considered how empathy has been engineered and the limitations identified with its use by a CA from a human perspective. There were three objectives: (1) to identify existing empathic CA design architectures within the MH care sector and to assess their technical performance in detecting and responding to user emotions appropriately; (2) to describe the approaches used to evaluate empathic CAs within the MH care sector in terms of their acceptability to users; and (3) to identify limitations and future directions for empathic CAs in MH care.

Methods

Database Search

A systematic literature search was conducted across 6 academic databases (Web of Science; Scopus; EBSCOhost: Academic Search Complete; CINAHL Complete; Computers and Applied Sciences Complete; and IEEE Xplore) for journal articles and conference proceedings from January 1, 2010, to September 30, 2023. The period of data capture dates from the time when AI-informed CA technology emerged as a distinct area of research [20], and conference proceedings were included to ensure that the most recent studies could be included.

The search terms covered 3 topics: "conversational agents," "mental health," and "empathy." Possible keywords were broadened using synonyms for each topic, pilot searching of existing literature, and discussion among research team members. Boolean operators combined different keywords and their synonyms to establish the final search strategy. Wildcards were included (eg, empath* = empathic). Medical Subject Heading terms were used where appropriate. An example of the search syntax is available in Multimedia Appendix 1 [4-6,21-36].

Eligibility Criteria

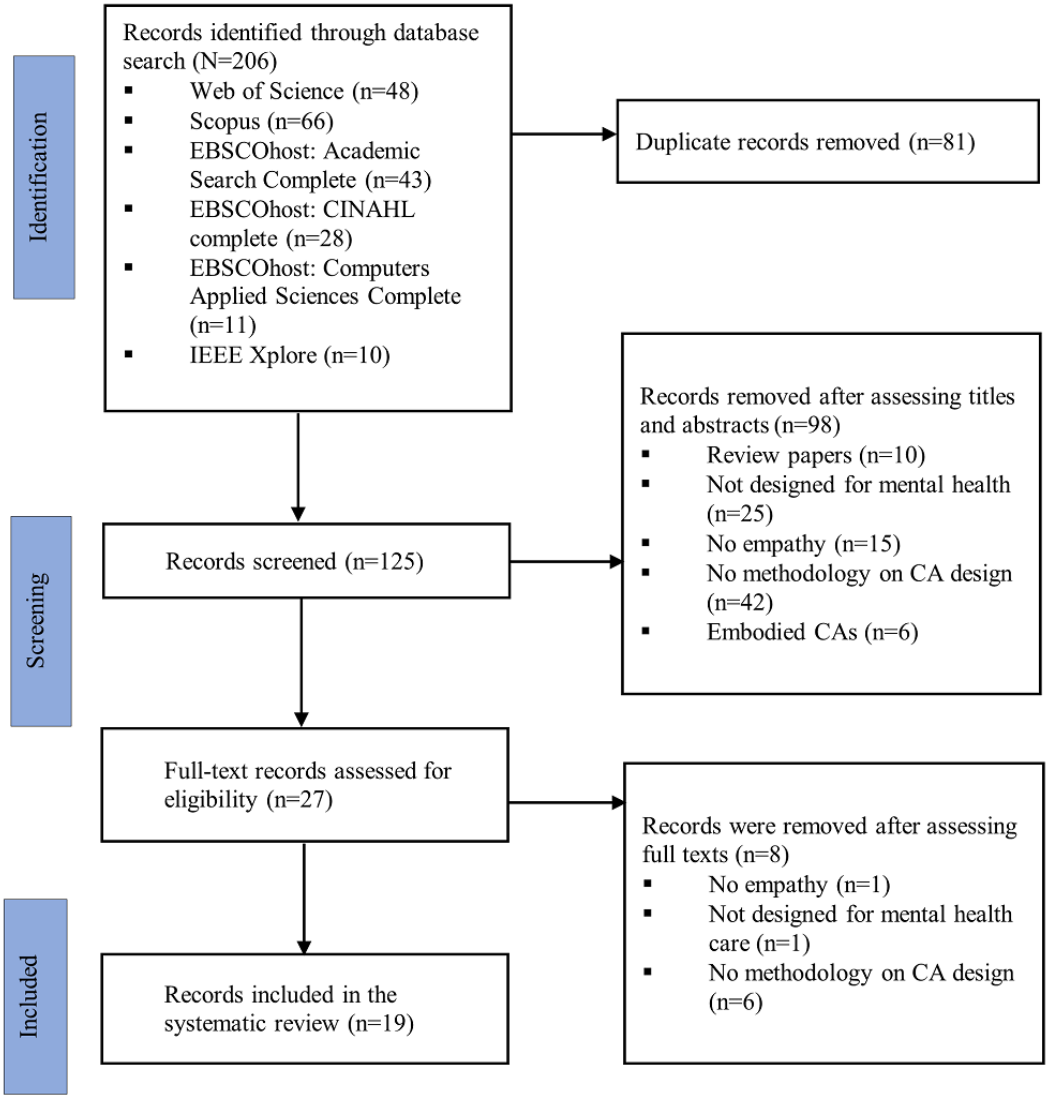
Publications discussing CA interventions for the MH care domain were eligible for the review. There were no restrictions on research design (eg, observational designs and narrative review). This review considered both textual and vocal modes of interaction with the CA. Publications were included if they referred to CA empathy or related terms (eg, emotional intelligence, emotional awareness, and compassion). Publications that did not feature a methodology section that detailed CA design, types of data sets, and participants were excluded. Systematic reviews, scoping reviews, and meta-analyses were excluded. Publications that used data inputs other than text and vocal cues (eg, facial recognition) were also excluded. Multimedia Appendix 1 provides the full-text screening checklist.

Screening

Eligible references were exported to the EndNote (version 20; Clarivate) software [37], where duplicates were removed. The first author (RS) conducted the title and abstract search, mapping against the eligibility criteria. A full-text screening was then performed by the first author and by 2 other authors, DM and RI, independently. Any disagreements on full-text screening were discussed, and an agreement was reached before proceeding. Figure 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart describing the screening process. PRISMA checklist is reported in Multimedia Appendix 2.

Data including details on the study designs, how empathy was evaluated, and the types of CA architectures used were extracted to obtain a summary of all findings (Multimedia Appendix 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) procedure applied. CA: conversational agent.



Quality Assessment

The Joanna Briggs Institute critical appraisal tool was used to assess the methodological quality of the papers shortlisted while also considering the extent to which each study addressed the possibility of bias in design, conduct, and analysis [38]. This appraisal tool was specifically designed for the assessment of the variety of study designs encountered in this systematic review. Decisional criteria were answered with *yes, no, unclear, or not applicable*. The proportion of *yes* responses relative to the total number of assessment questions was used for quality assessment purposes. Separate quality assessments were

conducted for publications that included a description of the implementation as well as the design of the CA platform and for publications that included only a description of the design.

Risk of Bias

Risk of bias was assessed using the revised Cochrane risk-of-bias tool for randomized trials This included risks of bias due to randomization, deviations from the intended intervention, missing data, the measurement of outcomes, and the selection of results. The risk of bias in nonrandomized studies of interventions tool was used to evaluate the nonrandomized studies.

Results

Overview

A total of 19 studies met all the inclusion criteria. The study characteristics are summarized in Table 1.

Table 1. Study characteristics.

Study	CA[a]	Training database	Aim of the study	Evaluation measures for detecting and responding to user emotions	Mode of exchange	Analysis model for generating empathic responses
Jiang et al [21], 2022	Replika	14 Chinese female users (aged 19-26 years)	Explore types of mediated empathy that occur in human-AI[b] interactions	In-depth interviews and survey results: user ratings of empathy	Text and voice	Transformer architecture
Brocki et al [22], 2023	Serena	Trained on "Pushshift" Reddit data set and tested on psychotherapy transcript	Help improve outcomes of counseling by lowering barriers to access	Survey results: user ratings of engagement and helpfulness	Text	Transformer architecture
Persons et al [6], 2021	ERIN[c]	15 undergraduate students	Help users with finding resources about sensitive issues	Survey results: user ratings for experience	Text	Rule-based architecture
Trappey et al [23], 2022	Virtual reality empathy-centric counseling CA	120 university students	Provide complementary support for students who were troubled	Survey results: user ratings of stress levels, life impact, and psychological sensitivity	Voice and text	Transformer architecture
Ghandeharioun et al [24], 2019	EMMA[d]	39 participants	Delivery of just-in-time MH[e] interventions	Survey results: user ratings of preference; behavioral metrics: user engagement	Text	Hybrid architecture
Meng and Dai [25], 2021	AI CA	278 participants from Midwestern University	Check whether the CA's emotional support was effective in reducing people's stress and worry	Survey results: user ratings of stress, worry, and perceived support	Text	Transformer architecture
Goel et al [26], 2021	Empathic CA with an attention mechanism	Trained with the Facebook AI Empathic Dialogue data set	Support users express their feelings and anxious thoughts	None	Text	Neural network architecture
Adikari et al [27], 2022	Empathic CA	Data set from Cancer Chat Canada	Provide empathic patient-centered MH care	Behavioral metrics for user engagement	Text	Hybrid architecture
Inkster et al [4], 2018	Wysa	129 users with self-reported symptoms of depression	Evaluation of the effectiveness and engagement levels of Wysa	Survey results for symptom assessment	Text	Unspecified ML[f] architecture
Beredo and Ong [28], 2022	Vhope	Senior high school and college students (aged 17-20 years)	Help the students maintain their well-being	Response ratings provided by experts	Text	Hybrid architecture
Rathnayaka et al [29], 2022	Bunji	Australian mobile users on Google Play Store	Remote health monitoring	Survey results for symptom and mood assessment	Text	Unspecified ML architecture
Morris et al [30], 2018	Koko	37,169 individuals who signed up for the Koko platform	A corpus-based approach to simulate expressed empathy	Response ratings provided by users	Text	Hybrid architecture
Ghandeharioun et al [31], 2019	A behavioral change CA	39 participants (n=7, 18% were female, and n=32, 82% were male)	Conduct experience sampling	Survey results: user ratings of likability and CA intelligence	Text	Rule-based architecture
Saha et al [32], 2022	Empathic CA	Data set: conversations between the support seekers who were depressed	Generate empathic and motivational responses	Response ratings by users for fluency, adaptability, and motivation	Text	Transformer architecture

Study	CA[a]	Training database	Aim of the study	Evaluation measures for detecting and responding to user emotions	Mode of exchange	Analysis model for generating empathic responses
Agnihotri et al [33], 2021	Topic-driven and affective CA	Data set: "ScenarioSA" with affective state labels	Tackle the emotional and contextual relevance for mental well-being	Response ratings for emotional relevance	Text	Transformer architecture
Rani et al [5], 2023	Saarthi	None	None	None	Text	Unspecified ML architecture
Alazraki et al [34], 2021	An empathic AI coach	23 participants recruited through crowd working websites	Achieve a high level of engagement during web-based therapy sessions	Survey results: user ratings of empathy and expert ratings of fluency	Text	Hybrid architecture
Gundavarapu et al [35], 2022	A CA companion	Data set: created using sources such as Wikipedia	Provide emotional support, without judgment	None	Text	Neural network architecture
Mishra et al [36], 2023	Counseling CA	A novel conversational data set	Provide MH and legal counseling	Survey results: user ratings of empathy	Text	Transformer architecture

[a]CA: conversational agent.

[b]AI: artificial intelligence.

[c]ERIN: Empathetic Research IoT Network.

[d]EMMA: Emotion-aware mHealth agent.

[e]MH: mental health.

[f]ML: machine learning.

Of the 19 studies, 6 (32%) were conducted in the United States and 6 (32%) in India. In addition, 1 (5%) study each from Australia, Canada, China, the Philippines, Poland, Switzerland, and the United Kingdom were also included. The year of publication is summarized in Multimedia Appendix 3, indicating a sharp rise in the number of publications since 2022. Most studies, 14 (74%) out of 19, described both design and human evaluations. The types of study designs among the 19 studies included are 9 (47%) cross-sectional studies, 5 (26%) randomized controlled trials (RCTs), 4 (21%) quasi-experimental designs, and 1 (5%) qualitative study. Only 5 (26%) of the 19 studies referred to an explicit definition of empathy, as summarized in Textbox 1.

Textbox 1. Definitions of empathy.

Studies and definition of empathy

- Jiang et al [21], 2022
  - Empathy processing is a situation-specific, cognitive-affective state or process with the projection of oneself into another's feelings, actions, and experiences.
- Trappey et al [23], 2022
  - Roger's [39] definition of empathy:
    - Level 1: responding to an individual's explicitly expressed meaning and feelings with a simple repetition of basic understanding.
    - Level 2: responding to the implicit, half-expressed, or implied feelings of the person with corresponding emotional words to acknowledge them and bring their true feelings to the surface.
    - Level 3: recognizing the individual's confusing and contradictory feelings that subconsciously obscure what they really care about, capturing the core of the emotion, and then responding to the patient's desire with affirmations.
    - Level 4: when the person is suppressing their feelings or not expressing their feelings in the conversation, guessing their intentions from what they are describing, capturing the core of the emotion, and responding to it directly or indirectly in a way that is acceptable to the person.
- Rathnayaka et al [29], 2022
  - Empathic engagement means, "making the impression of a credible and trustworthy conversation partner that can hear you out and offer a detached point of view on things."
- Saha et al [32], 2022
  - Empathy or empathic interactions refer to the ability to feel the emotions and experiences of others [40].
- Alazraki et al [34], 2021
  - Definition of empathy by Barrett-Lennard [41]:
    - First phase: where the listener sympathizes and resonates with what is being expressed by the speaker.
    - Second phase: where the listener compassionately responds to the speaker. Third phase: where the speaker assimilates the listener's response.

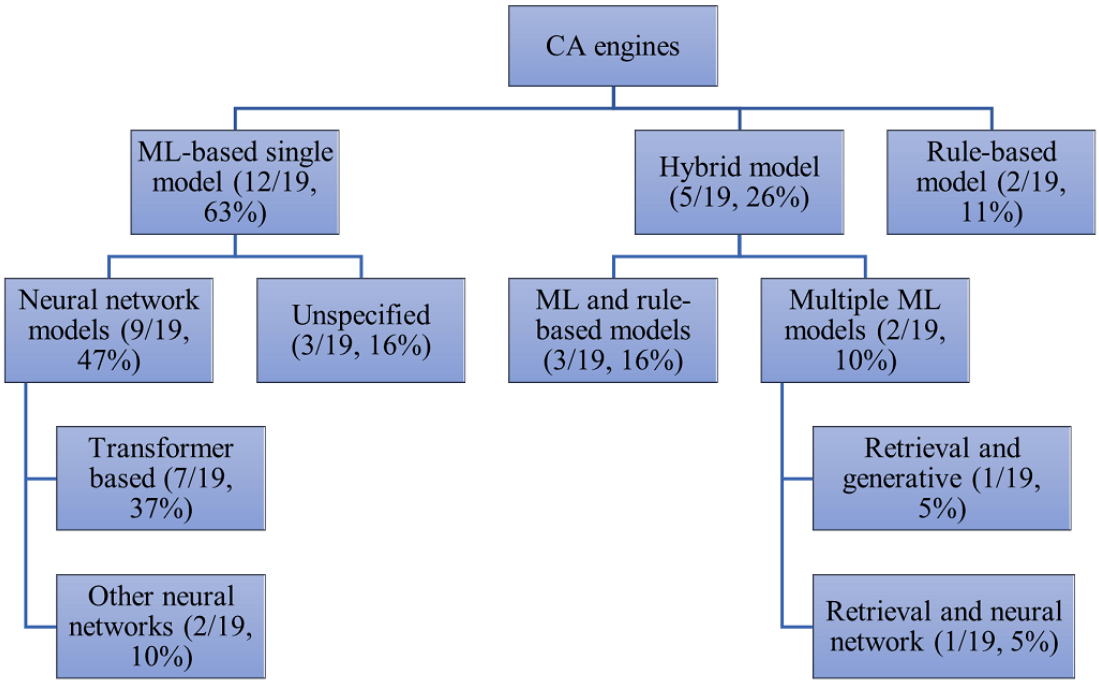
Keywords used to identify a CA varied across studies from "chatbot" (9/19, 47%) to "conversational agent" (6/19, 32%) to "dialog system" (2/19, 11%) to "virtual assistant" (1/19, 5%) to "conversational AI agent" (1/19, 5%). The mode of interaction chosen by most of the CA designs, 17 (89%) out of 19, was text (eg, live chat, symptom checker, and text-based counseling), with voice interactions being used in interactive avatar and counseling roles in 2 (11%) studies.

In the *Technical Design of the CAs* section, we consider the technical designs used for these CAs and their performance in detecting and responding to user emotions before discussing how human-user evaluations were conducted and the conclusions reached from these evaluations.

Technical Design of the CAs

The types of CA architectures (or engines) considered by the authors included a mix of recent technologies, as summarized in Figure 2, with ML-based architectures used in 12 (63%) out of 19 cases. The transformer-based engine, which learns meaning from context, was used in 7 of the 19 (37%) studies, sometimes in the form of a large language model (LLM). A minority of the papers, 3 (16%) out of 19, did not specify the type of engine used within the design. Hybrid or ensemble models use several models in parallel to improve the accuracy of the overall CA design. A more detailed breakdown of the CA engine types with explanations is shown in Multimedia Appendix 4. Figures S1 and S2 in Multimedia Appendix 4 also illustrate how a single engine and a hybrid engine work with user input to provide an empathic response.

Figure 2. Types of conversational agent (CA) architectures. ML: machine learning.



Transformer-based engines included Bidirectional Encoder Representations Transformer (BERT), Sentence-BERT, Robustly Optimized-BERT, Generative Pre-trained Transformer 2, and sequence-2-sequence models. Other neural network architecture–based CA designs were incorporated in 2 (11%) of the 19 papers [26,35].

Of the 19 publications, 5 (26%) considered hybrid models. Of these hybrid models, 2 applied a ML model to capture user emotion and then applied a rule-based algorithm to generate appropriate responses in dialogue management [24-27]. For example, EMMA gathered mobile sensor data to infer user mood and then assigned users to appropriate wellness interventions [24]. Once assigned, the CA then responded with emotionally expressive responses selected at random from a set of prescripted phrases using a rule-based approach [27]. In another example, VHope, an internet-based therapist, used a hybrid model containing a retrieval model that deciphered user input combined with a generative model to elicit empathic responses [28].

Among the 19 papers, the 3 (16%) papers using unspecified architectures commenced with natural language processing (NLP) before using various ML approaches. In one example, continuous emotional support via remote MH care monitoring and personalized assistance was provided [29]. MH monitoring was performed by scheduling activities that were meaningful to each user, sending out reminders as encouragement, and forwarding satisfaction surveys to receive feedback.

Overall, 2 (11%) of the 19 publications implemented CA design approaches based on rule-based NLP architectures. For example, a mobile phone–based CA measured the level of emotion in user input and then selected an appropriate empathic response from a set of predefined scripts using a rule-based decision tree [31]. In the *Summary of the Results of the Assessment of the Technical Design of CAs in Terms of Classification Accuracy* section, we will discuss the technical performance of the CAs reviewed.

Summary of the Results of the Assessment of the Technical Design of CAs in Terms of Classification Accuracy

The accuracy of the designs in detecting and responding to user emotions appropriately is summarized in Table 2. Technical evaluations of the CA designs usually involved comparisons with a "gold standard," using data not previously used for training the CA.

Table 2. Measures used for evaluating the technical performance of CA[a] designs.

Type of CA assessment	Assessment of user emotions or CA responses	Accuracy measure
Classification of sentiment and issues	User emotions	• Mathews correlation coefficient=0.857 [23]
Classification of valence and arousal	User emotions	• Accuracy of valence=80.4% [24] • Accuracy of arousal=50.4% [24]
Classification of recommended resources (for patients)	User emotions	• F1-score=0.87 [27]
Classification of objections during conversations	User emotions	• Accuracy=99.2% [4] • Specificity=99.7% [4] • Precision=74.7% [4] • Recall=62.1% [4]
Performance of the topic classifier	User emotions	• Accuracy=95% [33] • Precision=0.954 [33] • Recall=0.947 [33] • F1-score=0.95 [33]
Classification for empathy function	User emotions	• Accuracy=80.18% [34] • F1-score=80.66% [34] • W-ACC[b]=0.977 [36] • Macro F1-score=0.972 [36]
Prediction of valence and arousal	User emotions	• Accuracy of valence=82.2% [24] • Accuracy of arousal=65.7% [24]

Accuracy of the response generation	CA responses	• BLEU[c] score=0.126 [26]
		• BLEU-1 score (focused on a single word)=0.161 [32]
		• Perplexity score=50.90 [32]
		• ROUGE-L[d] score=0.124 [32]
		• Embedding-based metrics:
		  • Average=0.733 [32]
		  • Extrema=0.377 [32]
		  • Greedy=0.478 [32]

Emotion prediction	User emotions	• Accuracy
		  • Correctly predict the next emotion as positive or negative=79% [27]
		  • Proportion of correct emotion out of all emotions predicted=63% [27]

Performance of the language model	CA responses	• Perplexity score=9.977 [28] • Perplexity score=1.91 [23] • Response length=18.71 [23]
Emotion recognition	User emotions	• Accuracy=94.96% [34] • F1-score=95.10% [34]

[a]CA: conversational agent.

[b]W-ACC: weighted accuracy.

[c]BLEU: Bilingual Evaluation Understudy.

[d]ROUGE-L: Recall Oriented Understudy for Gisting Evaluation–Longest Common Sequence.

A technical evaluation of empathic CA performance was conducted in 17 (89%) of the 19 papers reviewed; however, only 10 (53%) papers reported these results. These studies conducted comprehensive assessments where technical performance was measured in terms of recognition, classification, prediction, and response generation abilities during interactions with end users. The assessments were centered around the ability of the CA to discern user emotions correctly and to respond appropriately. Of the 19 papers, 4 (21%) focused on the CA responses during the technical assessments, while the rest of the studies (n=15, 79%) considered user emotions. A variety of measures were used for

each such assessment, highlighting the diversity in evaluation methodologies across studies. These metrics are categorized in detail under the type of CA performance in Multimedia Appendix 5 [4-6,23-34,36].

In general, the performances of the CA designs were satisfactory. The highest classification accuracy for user emotions was reported by ML-based CAs. In one of these studies, a Robustly Optimized-BERT transformer model, which was built integrating 3 classifiers for politeness, counseling strategy, and empathic feedback, achieved good results overall. This empathy classifier achieved excellent performance with a weighted accuracy score of 0.977 and an F_1-score of 0.972 [36].

In a second study, a topic-driven classification model used a Generative Pre-trained Transformer 2 model for generating controlled responses, and the classification model accomplished relatively high scores of accuracy (95%), precision (0.954), and recall (0.947) and an F_1-score of 0.95 [33].

However, high accuracy and a more nuanced response generation were consistently apparent in all the CAs using hybrid architectures [24,27,28,30,34], suggesting that hybrid models lead to enhanced performance in tasks requiring complex understanding of user emotions and the generation of contextual responses.

Human Evaluation of CAs

Most of the reviewed studies, 16 (84%) out of 19, conducted a human evaluation of the implemented CA designs. Acceptability by end users was evaluated in terms of user experience, satisfaction, and levels of engagement. A detailed summary of the human evaluations of these designs is presented in Multimedia Appendix 5.

The human evaluation was performed by only CA users in most cases (13/16, 81%), while experts in the field of MH contributed to the process of assessing the CA in the remaining studies (3/16, 19%). Table 3 summarizes the empathy measures used in these papers.

Table 3. Measurement of empathy in CAs[a].

Study and year	The method of empathy measurement	How was empathy measured?	Who did the evaluation?	Evaluation results

Jiang et al [21], 2022	• Self-reports:
	  • In-depth interview responses
	  • Multiple response ratings
		Using the RoPE[b] scale (binary responses) and QCAE[c]	• Replika users provided the empathy ratings	• Perceived cognitive empathy was higher than perceived affective empathy

Beredo and Ong [28], 2022	• Self-reports: • Response ratings	Affect criterion or empathy was measured using a binary scale of 0 (no) to 1 (yes)	• Evaluated by 3 experts who studied and practiced psychology	• Responses were rated 79% empathic
Alazraki et al [34], 2021	• Self-reports: • Multiple response ratings	Multiple ratings to evaluate the perceived level of empathy, with ratings ranging from strongly disagree to strongly agree on a 5-point Likert scale	• Evaluated by users • 2 separate clinicians specialized in MH[d] also evaluated the chatbot personas	• When interacting with the Kai persona, 75% of users agreed that the bot was empathic • Interaction with other study personas achieved a 56% empathic rating
Mishra et al [36], 2023	• Self-reports: • Response ratings	A single 5-point Likert scale	• 6 evaluators rated each dialogue interaction for empathy • Empathy ratings by evaluators cross-validated for quality by government-run institutions	• Average empathy rating=57%
Agnihotri et al [33], 2021	• Self-reports: • Response ratings	Emotional relevance is rated using a single 5-point Likert scale	• Evaluated by 3 human annotators—male nonnative English speakers from a technical university with an average age of 21 years	When an empathic response generator was used, emotional relevance=61.4% • When a topic classifier was added, emotional relevance=43%

[a]CA: conversational agent.

[b]RoPE: Robot's Perceived Empathy.

[c]QCAE: Questionnaire of Cognitive and Affective Empathy.

[d]MH: mental health.

Alazraki et al [34] conducted a cross-sectional study with 23 volunteers and 2 clinicians who engaged with a web-based chatbot platform using 4 prescripted conversations of different CA personas. An anonymous web-based questionnaire collected participant feedback regarding the level of empathy displayed by the chatbot, engagement levels, and the ability of the chatbot to identify emotions in the participant. The survey results revealed that 75% of users agreed that the CA persona Kai was empathic, 63% found it engaging, and 75% rated it as useful. In contrast, Beredo and Ong [28] asked 3 psychologists to provide feedback on chatbot user logs. Empathy was measured using the affect criterion, a measure of the ability of the CA to read and respond to users with empathy, along with performance and humanlike characteristics. On the basis of expert feedback, 67% of the CA responses were relevant, 78% seemed human, and 70% were empathic.

In an RCT, a group of 39 participants were randomly allocated to a treatment group interacting with the emotion-aware chatbot EMMA, while a control group (n=39) was assigned to an emotionally nonexpressive chatbot, with 2 weeks of monitoring in each case [24]. The participants engaging with EMMA showed higher frequency of interactions and responded quicker than the control group. The feedback of the users was useful in understanding how empathy was perceived during the study.

The only qualitative experimental study involved an AI-based chatbot, Replika, designed to improve resilience and user well-being [21]. The author followed an ethnographic approach for their study of empathy, asking users to download the Replika application and write down reflective notes on their conversations with Replika. The results of this study expand the empathy theories within human conversations to human-AI interactions through variations in cognitive empathy, affective empathy, and empathic responses. A list of technical terms used in the paper is further explained in Multimedia Appendix 6.

Risk of Bias and Quality Assessment Results

The included RCTs showed a low risk of bias on the revised Cochrane risk-of-bias tool. Of the 14 nonrandomized studies included in the review, all showed a moderate to high risk of bias. A total of 5 (36%) studies [27,32-34,36] were moderately biased, and 1 (7%) study [28] was seriously biased according to the risk of bias in nonrandomized studies of interventions tool. The Joanna Briggs Institute quality assessment results were generally low when only the design component of the studies was assessed, with 32% (6/19) of the papers receiving a score of 0. However, an overall moderate quality was seen in publications when both the design and implementation stages were appraised. Multimedia Appendix 7 [4-6,21-36] shows the quality assessment results.

Discussion

Principal Findings

The study and use of CA technology have been the subject of extensive research across many fields, such as education, customer service, and health care. Moreover, there are reviews focusing on AI-based CAs, their effectiveness, and their impact in the realm of MH care [17,18,42]. While these reviews offer significant insights into AI-based CA designs in MH care, the importance of empathy is not central. Although these reviews suggest the need for empathy in CA innovations in MH care, they do not consider CA designs specifically aimed at generating and evaluating empathy. To address this gap, this review compares various empathic CA designs, their effectiveness in detecting and responding to user emotions, and their acceptability to users.

CA Designs

This review has found that most researchers used an ML-based transformer engine for designing empathic CAs, achieving excellent classification and prediction results. Surprisingly, several researchers used rule-based architectures and retrieval engines. While lacking the sophistication of transformer-based engines in terms of comprehension, rule-based approaches were able to efficiently identify keywords and themes, ensuring that consumer needs were addressed within a limited number of categories. Rule-based systems are comparatively easy to design and implement, allowing for a trade-off between classification accuracy and economic feasibility. However, rule-based systems tend to generate more predictable, inflexible, and repetitive responses compared to advanced LLM engines and, therefore, might be more suitable for providing simple information to managers and MH care workers, rather than responding to end users requiring more nuanced responses.

Hybrid architecture seems best suited to the detection of user emotion followed by the retrieval of a suitable response. Therefore, having >1 model appears to facilitate a more robust model output. This is supported by the superior accuracies achieved by hybrid architectures in the classification and prediction tasks. The hybrid model of Adikari et al [27] achieved the highest accuracy of 87% (F_1-score=0.87) in recommending a resource based on the concerns expressed by the patients. However, the highest accuracy in emotion recognition (95% accuracy in identifying sadness, anger, fear, and happiness) was obtained by Alazraki et al [34]. The combined features of high accuracy and improved user experience probably make these the best performing CAs within the review.

While the use of such robust LLMs has significantly improved language-based CA technology, it is important to recognize that these models are not without disadvantages [43]. These models have been found to perpetuate biases with regard to gender, race, and MH conditions present in the training data [44,45]. Such biases can strengthen gender stereotypes and reduce response accuracy when dealing with users from diverse cultural backgrounds, potentially causing harm to users. Such issues may have serious impacts on user trust, the credibility of the empathic CA, and user well-being. Such biases can be mitigated by ensuring that the training data sets represent diverse gender categories, races, and cultural backgrounds and that advanced technical approaches are used to detect and minimize any such biases in the training data [46-48].

Ethical and privacy concerns associated with these LLMs are critical [49,50]. Following ethical guidelines centered around transparency, accountability, and adherence are pivotal to user privacy, while measures to maintain data security through strict access controls and regular security checks also need to be in place. Privacy should be a core component of CA designs, with limitations placed on personal data collection whenever possible [7]. These strategies are especially important for an empathic CA design dealing with users seeking MH care. Any breaches of privacy and ethical guidelines pose a high risk to user mental well-being as well as users' trust in and acceptance of these new technologies [51]. The AI safety guidelines established by the European Union provide a key foundation for the creation of secure and ethical experiences for users [50].

Due to the complexity of LLMs and the many parameters involved, some models can have high latency in response time, which can cause potential challenges for a real-time CA dealing with vulnerable users waiting for a response. However, the use of parallel processing, optimization techniques, and hardware that supports the requirements of these AI models has facilitated a decrease in execution times [52].

Human Evaluations of CAs

Among the reviewed publications, human evaluation of chatbots was common. However, only 26% (5/19) of the studies used an RCT design to assess the CA platform. Random assignment to the treatment arm is known to reduce bias while improving the reliability of the experimental results. Any confounding factors are, therefore, likely to be controlled for in an RCT, making it important to overcome the practical difficulties these designs present in this context. RCTs provide the opportunity to observe user experiences with the CA designs over time. Ideally, future studies should consider RCT designs for their human evaluations, and ideally, the long-term effects of the CA can be examined over an extended timeline.

Previous experiences with CAs could be an important confounding factor. On the basis of these experiences, expectations of users regarding CA performance may affect actual engagement with the CA. Previous bad experiences may

make it less likely that a user will try to engage fully with a CA, resulting in a less favorable evaluation and satisfaction levels [53]. Another confounding factor could be the rate at which the user likes to communicate. If the CA cannot automatically adapt its speed of response to that preferred by the user, it is likely that this will also impact evaluation results [54].

The human evaluations of CAs in this review focused on their ability to portray empathy, satisfy user needs, provide useful and contextually informed responses, and facilitate user engagement. Most CAs were evaluated as satisfactory by end users. However, among the 19 papers reviewed, only 5 (26%) papers provided quantitative evaluations of CA empathy, and only 5 (26%) papers provided a definition of empathy.

Because empathy has been defined in numerous ways in the literature, it is important that in future studies users are given a framework that guides their perceptions of empathy. Future research on empathic CA designs would, therefore, benefit from a clear and well-established definition of empathy, such as that provided by the World Health Organization [8]. Ideally, standardized scales for perceived empathy should be used to enhance the reliability, comparability, and validity of survey results. In this review, other self-report measures were used as surrogates for empathy, with considerable variation in the types of scales used. However, self-report scales are subjective and prone to bias, with different meanings based on users' lived experiences [55]. Ideally, the impact of the CA on MH outcomes should also be assessed. Only 2 (11%) of the 19 papers in this review [4,29] used the Patient Health Questionnaire as their measure of MH outcomes, while 2 (11%) other papers considered stress levels in their evaluation [23,25].

Furthermore, the human evaluations were mostly conducted by study participants. Experts and professionals in the field of MH care were rarely consulted. There is a need for greater consultation with focus groups and user groups to ensure that the CA design best reflects the needs of all stakeholders [22]. Future research in this area should also consider an iterative design framework, incorporating the co-design and coevaluation of prototypes involving all stakeholders [22].

In summary, there were deficiencies in all the human evaluations included in this review. Only 5 (26%) of the 19 papers in this review included a direct evaluation of CA empathy in the design, while the rest (n=14, 74%) were more concerned with general user satisfaction. Only 2 (40%) of 5 these studies used multiple rating scales to measure the level of empathy portrayed by a CA, and only 1 (20%) of 5 these studies [34] considered evaluations by both users and clinicians. However, there were 4 studies that did consider the impact of the CA on MH outcomes.

Future Opportunities

A significant limitation of the CAs reviewed was the use of only textual input in all but 2 (11%) of the 19 studies where voice data were included, thus losing a valuable opportunity to leverage alternative and powerful forms of data input for evaluating empathy.. A range of vocal characteristics have been associated with the detection of suicide risk and psychological distress, which suggests that vocal characteristics might provide a natural extension for the detection of levels of empathy [56,57]. The omission of voice data is surprising given that empathy is communicated predominately through vocal cues. However, textual information is not without its advantages. As we have shown in this review, NLP approaches have been used to successfully detect and convey empathy by CAs. A novel approach would be to leverage both streams of information to identify vocal characteristics indicative of different levels of empathy in addition to textual cues. Characteristics of vocal and textual cues that are associated with empathy could be combined to create a CA design to attend to users of MH care facilities such as helpline services, patient triage, and emergency services [21,23].

Creating a CA design that accurately portrays empathy and adjusts the level of empathy to match the emotional status of patients is a significant challenge. Effective vocal interaction often faces hurdles due to technical issues in voice analysis, including the smooth processing and interpretation of data. These challenges are compounded by poor audio quality [58]; the presence of overlapping psychological states in users; and linguistic variability influenced by culture, age, gender, and accents [59-61]. The use of high-quality audio devices to capture user voice [62, 63] and the use of training data sets reflecting diverse human demographic features are two challenges in algorithm development aiming to provide effective vocal interaction in CAs in real time.

The integration of an empathic CA with voice analysis capabilities into crisis helpline services could benefit users and the service provider. Attending to callers during peak hours for the collection of demographic information, triage, and risk assessment of callers using their voice patterns are some of the possible roles that CAs could fulfill. The involvement of CAs in these capacities could help reduce caller wait times, streamline processes, and ensure 24-hour service availability while providing a nonjudgmental and sensitive interaction for users within a safe environment. Improved empathy portrayal by the CA would help enhance user engagement and CA acceptability, helping reduce the gap between the demand and supply of available crisis helpline services.

Summary

This review confirms that empathy is an important characteristic for CA implementation for MH care. It highlights the strengths of the ML-based architectures when it comes to CA design and provides evidence of both technical and human assessments of CA performance. The need for improvement in measures used for detecting the level of empathy exhibited by CAs is manifest. The importance of AI safety regarding ethical and privacy concerns is a neglected area and should be considered as a priority for future designs. The promise of empathic CA applications that use vocal inputs and outputs is another area warranting further research, with opportunities for crisis helpline services.

Limitations of the Review

The studies included in this review presented a mix of methods, which made it challenging to compare and analyze the results. This relates to the diversity in the CA designs included, along

with the different data formats obtained through human evaluations, such as survey results, response ratings, and interview feedback. The methods used to assess the accuracy of the technical designs were also varied, and a lack of empathy definitions and standard measures for perceived empathy made study comparisons difficult.

The quality rating of the studies emphasized the need for the complete reporting of CA designs as well as rigorous evaluation. Deficiencies in these areas meant that the quality ratings for several papers were low. Evaluation guidelines were often missing, which made it challenging to appraise the performance of these systems. Classification accuracy and the accuracy of the responses generated were assessed using a variety of methods, further complicating this comparison.

Conclusions

The objective of this systematic review was to identify the existing architectures of empathic CA designs and the types of CA design assessments used in MH care. A further aim was to determine how CA empathy is evaluated and to examine the limitations and future ideas for CAs in this specific context. More than half of the selected papers used the latest technologies in CA architectures, including designs developed using ML-based transformer engines (eg, LLMs). Evaluations of technical capabilities were conducted in most of the papers and demonstrated good levels of accuracy.

This review suggests that a hybrid design is ideally used for the design of an empathic CA, allowing an initial assessment of user emotion before any CA response is developed. This review indicates that human feedback is required to assess the extent to which the CA is successful in demonstrating empathy. It is recommended that well-validated scales be used for this purpose. Further research on the portrayal of empathy in CAs for MH care would benefit by involving cocreation activities, explicit definitions of empathy, and effective evaluation of empathy using standardized empathy scales, as well as by using vocal features associated with empathy in addition to textual cues.

Despite its limitations, this review demonstrates that it is possible to design AI-empowered CAs that evoke empathy within MH care applications, with many of these CAs being rated as satisfactory by human users. This suggests that such CAs could prove beneficial in a range of settings, such as crisis helpline services, gathering data on user characteristics and emotions, and in postvention follow-up, helping to bridge the gap between the existing supply and demand for MH services.

Acknowledgments

This study was funded by Swinburne University of Technology.

Authors' Contributions

RS, DM, and RI contributed to the study selection process. RS and DM conducted the quality assessment of the included studies. RS, DM, RI, PA, and NW were involved in the concept, design, revisions, and final approval of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Screening process and study characteristics.
[DOCX File , 42 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist 2020.
[DOCX File , 32 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Evolution of conversational agent (year by year).
[DOCX File , 56 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Detailed summary of conversational agent types.
[DOCX File , 49 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Results of conversational agent evaluations.
[DOCX File , 99 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Dictionary of technical terms.
[DOCX File , 16 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Risk of bias and quality assessment.
[DOCX File , 16 KB-Multimedia Appendix 7]

References

1. Mental health atlas 2020. World Health Organization. Oct 8, 2021. URL: https://www.who.int/publications/i/item/9789240036703 [accessed 2024-08-10]
2. Schick A, Feine J, Morana S, Maedche A, Reininghaus U. Validity of chatbot use for mental health assessment: experimental study. JMIR Mhealth Uhealth. Oct 31, 2022;10(10):e28082. [FREE Full text] [doi: 10.2196/28082] [Medline: 36315228]
3. Fitzpatrick KK, Darcy A, Vierhile M. 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 Ment Health. Jun 06, 2017;4(2):e19. [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
4. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth. Nov 23, 2018;6(11):e12106. [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
5. Rani K, Vishnoi H, Mishra M. A mental health chatbot delivering cognitive behavior therapy and remote health monitoring using NLP and AI. In: Proceedings of the International Conference on Disruptive Technologies. 2023. Presented at: ICDT 2023; May 11-12, 2023; Greater Noida, India. [doi: 10.1109/icdt57929.2023.10150665]
6. Persons B, Jain P, Chagnon C, Djamasbi S. Designing the empathetic research IoT network (ERIN) chatbot for mental health resources. In: Proceedings of the HCI in Business, Government and Organizations: 8th International Conference, HCIBGO 2021, Held as Part of the 23rd HCI International Conference, HCII 2021. 2021. Presented at: HCI 2021; July 24-29, 2021; Virtual Event. [doi: 10.1007/978-3-030-77750-0_41]
7. Li L, Peng W, Rheu MM. Factors predicting intentions of adoption and continued use of artificial intelligence chatbots for mental health: examining the role of UTAUT model, stigma, privacy concerns, and artificial intelligence hesitancy. Telemed J E Health. Mar 01, 2024;30(3):722-730. [doi: 10.1089/tmj.2023.0313] [Medline: 37756224]
8. Hojat M. A definition and key features of empathy in patient care. In: Hojat M, editor. Empathy in Health Professions Education and Patient Care. Cham, Switzerland. Springer; 2016.
9. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst. Jul 20, 2022;12(2):1-39. [doi: 10.1145/3485874]
10. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. Apr 28, 2023;25:e43862. [FREE Full text] [doi: 10.2196/43862] [Medline: 37115595]
11. Kraus M, Seldschopf P, Minker W. Towards the development of a trustworthy chatbot for mental health applications. In: Proceedings of the 27th International Conference on MultiMedia Modeling. 2021. Presented at: MMM 2021; June 22-24, 2021; Prague, Czech Republic. [doi: 10.1007/978-3-030-67835-7_30]
12. Kallivalappil N, D'souza K, Deshmukh A, Kadam C, Sharma N. Empath.ai: a context-aware chatbot for emotional detection and support. In: Proceedings of the 14th International Conference on Computing Communication and Networking Technologies. 2023. Presented at: ICCCNT 2023; July 6-8, 2023; Delhi, India. [doi: 10.1109/icccnt56998.2023.10306584]
13. Lin S, Lin L, Hou C, Chen B, Li J, Ni S. Empathy-based communication framework for chatbots: a mental health chatbot application and evaluation. In: Proceedings of the 11th International Conference on Human-Agent Interaction. 2023. Presented at: HAI '23; December 4-7, 2023; Gothenburg, Sweden. [doi: 10.1145/3623809.3623865]
14. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices. Dec 2021;18(sup1):37-49. [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
15. Sweeney C, Potts C, Ennis E, Bond R, Mulvenna MD, O'neill S, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare. Jul 15, 2021;2(3):1-15. [doi: 10.1145/3453175]
16. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health. Nov 30, 2020;2:576361. [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
17. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. Dec 19, 2023;6(1):236. [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]

18. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health. Oct 18, 2019;6(10):e14166. [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
19. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res. Jan 13, 2021;23(1):e17828. [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
20. Behrens S. The history and evolution of conversational AI. Fabric. Apr 17, 2021. URL: https://gyant.com/the-history-and-evolution-of-conversational-ai/ [accessed 2023-03-13]
21. Jiang Q, Zhang Y, Pian W. Chatbot as an emergency exist: mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. Inf Process Manag. Nov 2022;59(6):103074. [FREE Full text] [doi: 10.1016/j.ipm.2022.103074] [Medline: 36059428]
22. Brocki L, Dyer GC, Gładka A, Chung NC. Deep learning mental health dialogue system. In: Proceedings of the IEEE International Conference on Big Data and Smart Computing. 2023. Presented at: BigComp 2023; February 13-16, 2023; Jeju, Republic of Korea. [doi: 10.1109/bigcomp57234.2023.00097]
23. Trappey AJ, Lin AP, Hsu KY, Trappey CV, Tu KL. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. Processes. May 08, 2022;10(5):930. [doi: 10.3390/pr10050930]
24. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. EMMA: an emotion-aware wellbeing chatbot. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. 2019. Presented at: ACII 2019; September 3-6, 2019; Cambridge, UK. [doi: 10.1109/acii.2019.8925455]
25. Meng J, Dai Y. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput Mediat Commun. May 19, 2021;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
26. Goel R, Vashisht S, Dhanda A, Susan S. An empathetic conversational agent with attentional mechanism. In: Proceedings of the International Conference on Computer Communication and Informatics. 2021. Presented at: ICCCI 2021; January 27-29, 2021; Coimbatore, India. [doi: 10.1109/iccci50826.2021.9402337]
27. Adikari A, de Silva D, Moraliyage H, Alahakoon D, Wong J, Gancarz M, et al. Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare. Future Gener Comput Syst. Jan 2022;126:318-329. [doi: 10.1016/j.future.2021.08.015]
28. Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. In: Proceedings of the International Conference on Asian Language Processing. 2022. Presented at: IALP 2022; October 27-28, 2022; Singapore, Singapore. [doi: 10.1109/ialp57159.2022.9961311]
29. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel). May 11, 2022;22(10):3653. [FREE Full text] [doi: 10.3390/s22103653] [Medline: 35632061]
30. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res. Jun 26, 2018;20(6):e10148. [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]
31. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. Towards understanding emotional intelligence for behavior change chatbots. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. 2019. Presented at: ACII 2019; September 3-6, 2019; Cambridge, UK. [doi: 10.1109/acii.2019.8925433]
32. Saha T, Gakhreja V, Das AS, Chakraborty S, Saha S. Towards motivational and empathetic response generation in online mental health support. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022. Presented at: SIGIR '22; July 11-15, 2022; Madrid, Spain. [doi: 10.1145/3477495.3531912]
33. Agnihotri M, Rao SB, Jayagopi DB, Hebbar S, Rasipuram S, Maitra A, et al. Towards generating topic-driven and affective responses to assist mental wellness. In: Proceedings of the 25th International Conference on Pattern Recognition. 2021. Presented at: ICPR 2021; January 10-15, 2021; Milan, Italy. [doi: 10.1007/978-3-030-68790-8_11]
34. Alazraki L, Ghachem A, Polydorou N, Khosmood F, Edalat A. An empathetic AI coach for self-attachment therapy. In: Proceedings of the IEEE Third International Conference on Cognitive Machine Intelligence. 2021. Presented at: CogMI 2021; December 13-15, 2021; Atlanta, GA. [doi: 10.1109/cogmi52975.2021.00019]
35. Gundavarapu MR, Koundinya GS, Sai TB, Sree GK. Empathic chatbot: emotional astuteness for mental health well-being. In: Proceedings of the 7th International Conference on Computing in Engineering & Technology. 2022. Presented at: ICCET 2022; February 25-27, 2022; Virtual Event. [doi: 10.1007/978-981-19-2719-5_65]
36. Mishra K, Priya P, Ekbal A. Help me heal: a reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. 2023. Presented at: AAAI 2023; February 7-14, 2023; Washington, DC. [doi: 10.1609/aaai.v37i12.26685]
37. Gotschall T. EndNote 20 desktop version. J Med Libr Assoc. Jul 01, 2021;109(3):520-522. [FREE Full text] [doi: 10.5195/jmla.2021.1260] [Medline: 34629985]
38. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Systematic reviews of etiology and risk. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia. Joanna Briggs Institute; 2024.

39. Rogers CR. Empathic: an unappreciated way of being. Couns Psychol. Sep 04, 2016;5(2):2-10. [doi: 10.1177/001100007500500202]
40. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic). Dec 2018;55(4):399-410. [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
41. Barrett-Lennard GT. The empathy cycle: refinement of a nuclear concept. J Counsel Psychol. 1981;28(2):91-100. [doi: 10.1037//0022-0167.28.2.91]
42. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry. Jul 2019;64(7):456-464. [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
43. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 06, 2024;8(1):1-32. [doi: 10.1145/3643540]
44. Salutari F, Ramos J, Rahmani HA, Linguaglossa L, Lipani A. Quantifying the bias of transformer-based language models for African American English in masked language modeling. In: Proceedings of the 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2023. Presented at: PAKDD 2023; May 25-28, 2023; Osaka, Japan. [doi: 10.1007/978-3-031-33374-3_42]
45. Li B, Peng H, Sainju R, Yang J, Yang L, Liang Y, et al. Detecting gender bias in transformer-based models: a case study on BERT. arXiv. Preprint posted online on October 15, 2021. [doi: 10.48550/arXiv.2110.15733]
46. Kamboj P, Kumar S, Goyal V. Measuring and mitigating gender bias in contextualized word embeddings. In: Proceedings of the IEEE International Conference on Blockchain and Distributed Systems Security. 2023. Presented at: ICBDS 2023; October 6-8, 2023; New Raipur, India. [doi: 10.1109/icbds58040.2023.10346586]
47. Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. arXiv. Preprint posted online on October 16, 2021. [FREE Full text] [doi: 10.18653/v1/2022.acl-long.132]
48. Gira M, Zhang R, Lee K. Debiasing pre-trained language models via efficient fine-tuning. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. 2022. Presented at: LTEDI 2022; May 27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.ltedi-1.8]
49. Kumar A, Murthy SV, Singh S, Ragupathy S. The ethics of interaction: mitigating security threats in LLMs. arXiv. Preprint posted online on January 22, 2024
50. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L. ChatGPT needs SPADE (sustainability, PrivAcy, digital divide, and ethics) evaluation: a review. Cogn Comput. May 05, 2024. [doi: 10.1007/s12559-024-10285-1]
51. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: ethical issues with using chatbots in mental health. Digit Health. Jun 22, 2023;9:20552076231183542. [FREE Full text] [doi: 10.1177/20552076231183542] [Medline: 37377565]
52. Agrawal A, Kedia N, Panwar A, Mohan J, Kwatra N, Gulavani BS, et al. Taming throughput-latency tradeoff in LLM inference with Sarathi-serve. arXiv. Preprint posted online on March 4, 2024. [FREE Full text] [doi: 10.48550/arXiv.2403.02310]
53. Santhanam S, Karduni A, Shaikh S. Studying the effects of cognitive biases in evaluation of conversational agents. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020. Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376318]
54. Gnewuch U, Morana S, Adam MT, Maedche A. Opposing effects of response time in human–chatbot interaction. Bus Inf Syst Eng. May 30, 2022;64:773-791. [doi: 10.1007/s12599-022-00755-x]
55. Prince SA, Cardilli L, Reed JL, Saunders TJ, Kite C, Douillette K, et al. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. Int J Behav Nutr Phys Act. Mar 04, 2020;17(1):31. [FREE Full text] [doi: 10.1186/s12966-020-00938-3] [Medline: 32131845]
56. Iyer R, Nedeljkovic M, Meyer D. Using voice biomarkers to classify suicide risk in adult telehealth callers: retrospective observational study. JMIR Ment Health. Aug 15, 2022;9(8):e39807. [FREE Full text] [doi: 10.2196/39807] [Medline: 35969444]
57. Iyer R, Nedeljkovic M, Meyer D. Using vocal characteristics to classify psychological distress in adult helpline callers: retrospective observational study. JMIR Form Res. Dec 19, 2022;6(12):e42249. [FREE Full text] [doi: 10.2196/42249] [Medline: 36534456]
58. Schaeffler F, Jannetts S, Beck J. Reliability of clinical voice parameters captured with smartphones — measurements of added noise and spectral tilt. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. 2019. Presented at: INTERSPEECH 2019; September 15-19, 2019; Graz, Austria. [doi: 10.21437/interspeech.2019-2910]
59. Laukka P, Elfenbein HA, Thingujam NS, Rockstuhl T, Iraki FK, Chui W, et al. The expression and recognition of emotions in the voice across five nations: a lens model analysis based on acoustic features. J Pers Soc Psychol. Nov 2016;111(5):686-705. [doi: 10.1037/pspi0000066] [Medline: 27537275]

60. Markova D, Richer L, Pangelinan M, Schwartz DH, Leonard G, Perron M, et al. Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. Horm Behav. May 2016;81:84-96. [doi: 10.1016/j.yhbeh.2016.03.001] [Medline: 27062936]
61. Israelsson A, Seiger A, Laukka P. Blended emotions can be accurately recognized from dynamic facial and vocal expressions. J Nonverbal Behav. May 17, 2023;47(3):267-284. [doi: 10.1007/s10919-023-00426-9]
62. Busquet F, Efthymiou F, Hildebrand C. Voice analytics in the wild: validity and predictive accuracy of common audio-recording devices. Behav Res Methods. Mar 30, 2024;56(3):2114-2134. [FREE Full text] [doi: 10.3758/s13428-023-02139-9] [Medline: 37253958]
63. Padmapriya J, Sasilatha T, R K, Aagash G, Bharathi V. Voice extraction from background noise using filter bank analysis for voice communication applications. In: Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. 2021. Presented at: ICICV 2021; February 4-6, 2021; Tirunelveli, India. [doi: 10.1109/icicv50876.2021.9388453]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations Transformer
CA: conversational agent
LLM: large language model
MH: mental health
ML: machine learning
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by J Torous; submitted 30.03.24; peer-reviewed by Y Okoro, P Upadhyaya, BG Teferra, M Bagewadi Ellur; comments to author 06.05.24; revised version received 01.07.24; accepted 02.07.24; published 09.09.24

Please cite as:
Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D
Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review
JMIR Ment Health 2024;11:e58974
URL: https://mental.jmir.org/2024/1/e58974
doi: 10.2196/58974
PMID: