Systematic Review of Empathic Conversational Agent Platform Designs and their Evaluation in the Context of Mental Health.

Ruvini Mapa, Denny Meyer, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe

Submitted to: JMIR Mental Health
on: March 30, 2024

Table of Contents

Systematic Review of Empathic Conversational Agent Platform Designs and their Evaluation in the Context of Mental Health.

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

[1]School of Health Sciences Swinburne University of Technology Hawthorn AU
[2]School of Computing, Engineering and Mathematical Sciences La Trobe University Bundoora AU

Corresponding Author:
Ruvini Mapa BSc
School of Health Sciences
Swinburne University of Technology
John Street
Hawthorn
AU

Abstract

Background: The demand for mental health 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 for helping to fill this gap.

Objective: The objective of this review is to identify existing empathic CA design architectures within the mental healthcare sector and to assess their technical performance in terms of classification accuracy. In addition, the approaches used to evaluate empathic CAs within the mental healthcare sector in terms of their acceptability to users will be evaluated. Finally, this review aims to identify limitations and future directions for empathic CAs in mental healthcare.

Methods: A systematic literature search was conducted across six academic databases to identify journal articles and conference proceedings using search terms covering three topics: ‘conversational agents’, ‘mental health’, and ‘empathy’. Only studies discussing CA interventions for the mental healthcare 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 and exclusion criteria. The observed terms used to identify a CA varied from ‘chatbot’ (47%), ‘conversational agent’ (32%), ‘dialog system’ (11%), ‘virtual assistant’ (5%) to ‘conversational AI agent’ (5%). Transformer-based (37%) and hybrid (26%) engines were the most employed designs. A technical evaluation of CA performance was conducted for 17 of the 19 papers reviewed. While a variety of single-engine CAs exhibited good accuracy (F1 scores >95%), superior accuracy was achieved using hybrid engines that were able to provide a more nuanced response. However, human evaluations of CAs were less positive. Only five (26%) of the 19 studies referred to an explicit definition of empathy and only 84% of the selected studies used human evaluation to assess the effectiveness of the CA designs. A direct evaluation of empathy of the CA was involved in only five studies using questionnaire responses, response ratings and in-depth interview responses. The human evaluation of CAs was performed mostly by end-users (75%), while experts in Mental Health (MH) assessed CAs in the remaining studies (25%). A variety of measures were used to evaluate the level of empathy exhibited by CAs. For example, Patient Health Questionnaires were used to show an improvement in the mean mood from 5.79 to 7.38 on a 10-point scale for one particular CA. In three other studies, empathy ratings were recorded with empathic percentages of 56%, 75% and 79%, with average empathy scores and emotional relevance scores of 2.85 or 3.05 out of 5 respectively in other studies.

Conclusions: CAs with good technical and empathic performance are now available to users of mental healthcare services. Clinical Trial: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022348130

(JMIR Preprints 30/03/2024:58974)
DOI: https://doi.org/10.2196/preprints.58974

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Systematic Review of Empathic Conversational Agent Platform Designs and their Evaluation in the Context of Mental Health.

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 & Mathematical Sciences, La Trobe University, Australia.

Corresponding Author:

Ruvini Sanjeewa,
Swinburne University of Technology,
Hawthorn, 3122
Australia
Phone 0422587030
Email: rsanjeewa@swin.edu.au

Abstract

Background:

The demand for mental health 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 for helping to fill this gap.

Objectives:

The objective of this review is to identify existing empathic CA design architectures within the mental healthcare 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 mental healthcare sector in terms of their acceptability to users will be considered. Finally, this review aims to identify limitations and future directions for empathic CAs in mental healthcare.

Methods:

A systematic literature search was conducted across six academic databases to identify journal articles and conference proceedings using search terms covering three topics: 'conversational agents', 'mental health', and 'empathy'. Only studies discussing CA interventions for the mental healthcare 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 and exclusion criteria. The majority (63%) of these empathic CA designs in mental healthcare were machine learning (ML) based, with 26% hybrid engines and 11% rule-based systems. Among the ML-based CAs, 47% employed neural networks, with transformer-based architectures being well represented (37%). The remaining 16% of ML models were unspecified. Technical assessments of these CAs focused on response accuracies and their ability to recognise, predict, and classify user emotions. While single-engine CAs demonstrated good accuracy, the hybrid engines achieved higher accuracy and provided more nuanced responses. Human evaluations were conducted in 16 (84%) of the studies, with only five papers (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 one paper (5%) 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. Additionally, 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 mental healthcare services, showing promise for new applications such as helpline services.

Keywords: conversational agents; chatbots; virtual assistants; empathy; emotionally aware; mental health; mental wellbeing.

Introduction

An escalation in Mental Health (MH) diagnoses in the community, inadequate facilities, and a mental healthcare workforce that does not meet demand are placing extraordinary pressures on an already strained system [1]. This service gap creates a significant opportunity for mental healthcare 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 mental healthcare domain, providing additional functionalities and support to address this gap [2]. Examples of CAs that employ ML include Woebot - providing cognitive behavioural therapy [3], Wysa - providing MH support by checking depressive symptoms [4], Saarthi - trained to provide personalised and empathic support to patients via therapeutic techniques [5] and ERIN - 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 also contributes to end-user mistrust [7].

Empathy in patient care has been defined by the World Health Organisation (WHO) 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]. Counsellor 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 bi-directional interaction between CA and end-user [13].

Assessment of the effectiveness of CA platforms has received little attention in the mental healthcare sector [14]. For the impact of these systems to be fully realised, these platforms need to meet the requirements of end-users, which suggests a key role for lived experience and co-production. The validity and reliability of these new digital technologies also need to be reviewed by mental healthcare 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 wellbeing 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 bi-directional interaction described above.

While there are existing reviews exploring the efficacy of CAs designed for mental healthcare [10, 17, 18], to our knowledge this is the first review to specifically examine how these CAs are designed and evaluated to deliver empathy. A comprehensive systematic review and meta-analysis of AI-based CAs for promoting MH was conducted by Li [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 targeted CA interventions for treating MH problems, with a specific focus on user experience outcomes as measures of efficacy [18]. Another such study explored evidence of effectiveness with regards to improving symptoms of MH conditions [19]. A critical finding of this review was that empathic response and personalisation 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].

This review aims to assess the types of CA designs found in the mental healthcare sector that are specifically tailored to convey empathy. It also aims to describe the methods used to evaluate these empathic designs from a technical and implementation perspective. This review therefore considers how empathy has been engineered and the limitations identified with its use by a CA from a human perspective. There are three objectives:

- To identify existing empathic CA design architectures within the mental healthcare sector and to assess their technical performance in detecting and responding to user emotions appropriately.

- To describe the approaches used to evaluate empathic CAs within the mental healthcare sector in terms of their acceptability to users.
- To identify limitations and future directions for empathic CAs in mental healthcare.

Method

Database search

A systematic literature search was conducted across six 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 1/1/2010 to 30/9/2023. 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 three topics: ‘conversational agents’, ‘mental health’, and ‘empathy’. Possible keywords were broadened using synonyms for each topic, pilot searching of existing literature, and discussion amongst research team members. Boolean operators combined different keywords and their synonyms to establish the final search strategy. Wildcards were included (e.g., empath* = empathic). MESH terms were used where appropriate. An example of the search syntax is available in Multimedia Appendix 1.

Eligibility Criteria

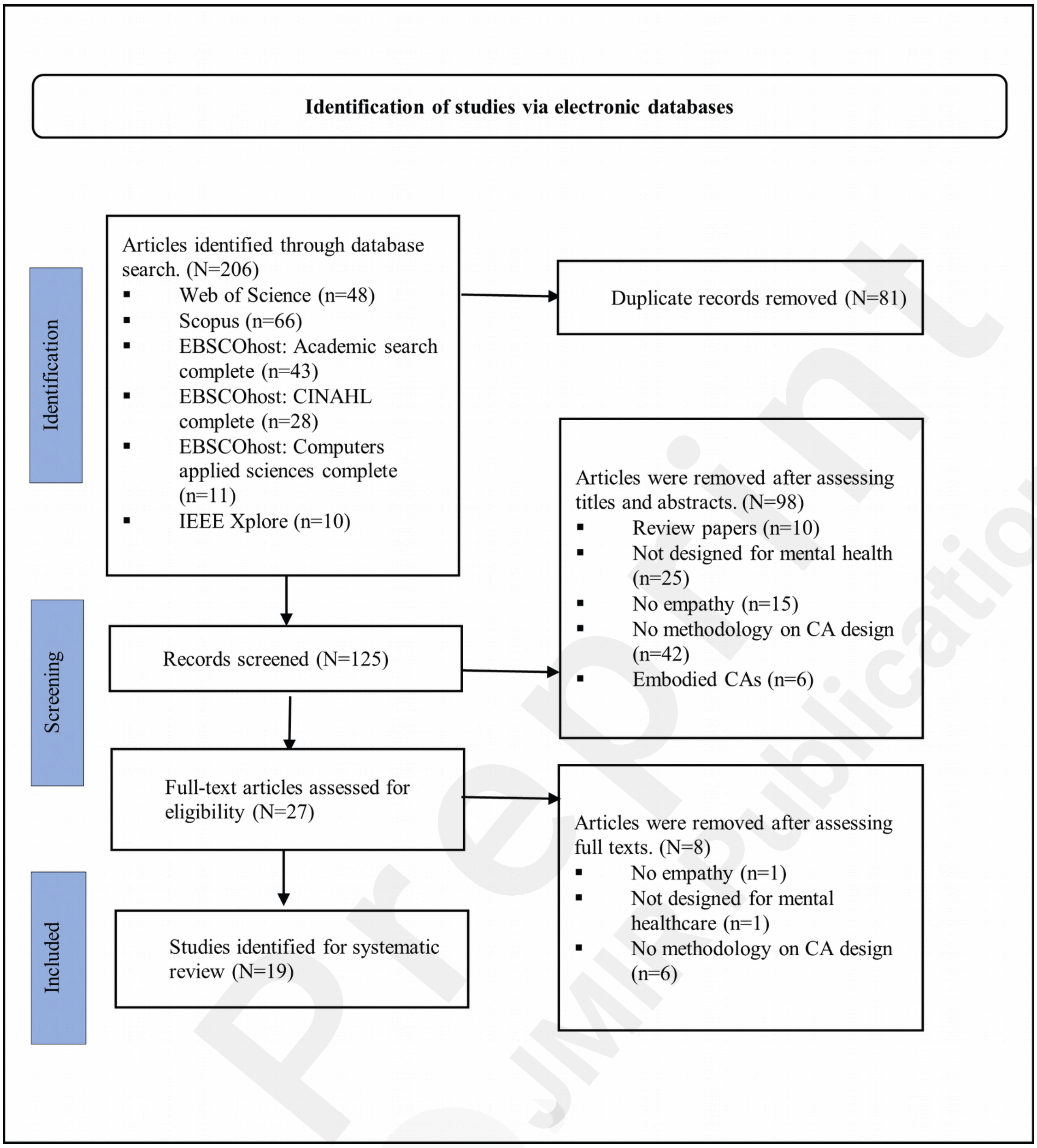
Publications discussing CA interventions for the mental healthcare domain were eligible for the review. There were no restrictions on research design (e.g. observational designs, narrative review). This review considered both textual and vocal modes of interaction with the CA. Publications were included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

Screening

Eligible references were exported to EndNote 20 software [21], 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 two other authors, DM and RI, independently. Any disagreements on full-text screening were discussed and agreed upon before proceeding. Figure 1 illustrates the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart describing the screening process. PRISMA checklist reported in Multimedia Appendix 2.

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

Figure 1: PRISMA Procedure Applied


Identification of studies via electronic databases
Identification
Screening
Included
Articles identified through database search. (N=206)
▪ Web of Science (n=48)
▪ Scopus (n=66)
▪ EBSCOhost: Academic search complete (n=43)
▪ EBSCOhost: CINAHL complete (n=28)
▪ EBSCOhost: Computers applied sciences complete (n=11)
▪ IEEE Xplore (n=10)
Duplicate records removed (N=81)
Records screened (N=125)
Articles were removed after assessing titles and abstracts. (N=98)
▪ Review papers (n=10)
▪ Not designed for mental health (n=25)
▪ No empathy (n=15)
▪ No methodology on CA design (n=42)
▪ Embodied CAs (n=6)
Full-text articles assessed for eligibility (N=27)
Articles were removed after assessing full texts. (N=8)
▪ No empathy (n=1)
▪ Not designed for mental healthcare (n=1)
▪ No methodology on CA design (n=6)
Studies identified for systematic review (N=19)

Quality assessment

The Joanna Briggs Institute (JBI) 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 [22]. This appraisal tool is specifically designed for assessment of the variety of study designs encountered in this systematic review. Decisional criteria were answered with either ‘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 only included a description of the design.

Risk-of-Bias

Risk of bias was assessed using the Revised Cochrane Risk-Of-Bias tool for randomised trials (ROB-2). This included risks of bias due to randomisation, deviations from the intended intervention, missing data, measurement of outcomes, and selection of results. The Risk-Of-Bias In Non-randomised Studies of Interventions (ROBINS-I) tool was used to evaluate the non-randomised studies.

Results

Nineteen studies met all inclusion criteria. The study characteristics are summarised in Table 1.

Table 1: Study Characteristics

Study	CA	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. [23]	Replika	14 Chinese female users (Aged 19-26)	Explore types of mediated empathy that occur in human-AI interactions	In-depth interviews/ Survey results: User ratings of empathy	Text and Voice	Transformer architecture
Brocki et al. [24]	Serena	Trained on 'Pushshift' Reddit Dataset, tested on psycho-therapy transcripts	Help improve outcomes of Counselling by lowering barriers to access.	Survey results: User ratings of engagement and helpfulness.	Text	Transformer architecture
Persons et al. [6]	ERIN	15 undergraduate students	Help users with finding resources about sensitive issues.	Survey results: User ratings for experience.	Text	Rule-based architecture
Trappey et al. [25]	Virtual reality empathy-centric counselling CA.	120 University Students	Provide complementary support for troubled students.	Survey results: User ratings of stress levels, life impact, and psychological sensitivity.	Voice and Text	Transformer architecture
.Ghandeharioun et al. [26]	EMMA	39 participants	Delivery of just-in-time MH interventions.	Survey results: User ratings of preference.	Text	Hybrid architecture.

				Behavioural metrics: to measure user engagement.		
Meng and Dai [27]	AI CA	278 participants from Midwestern University.	Check if 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. [28]	Empathic CA with an attention mechanism.	Trained with the Facebook AI Empathic Dialogue dataset.	Support users express their feelings and anxious thoughts.	None	Text	Neural network architecture
Adikari et al. [29]	Empathic CA.	Dataset from Cancer Chat Canada.	Provide empathic patient-centred mental healthcare.	Behavioural metrics for user engagement.	Text	Hybrid architecture
Inkster et al. [4]	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 architecture.
Beredo and Ong [30]	Vhope	Senior high school and college students. (Aged 17 to 20)	Help the students maintain their well-being.	Response ratings provided by experts.	Text	Hybrid architecture.
Rathnayaka et al. [31]	Bunji	Australian mobile users on Google	Remote health monitoring.	Survey results for symptom and mood assessment.	Text	Unspecified-ML architecture.

		Play Store.				
Morris et al. [32]	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. [33]	A behavioural change CA.	39 participants (7 were females and 32 were males)	Conducts experience sampling.	Survey results: user ratings of likability and CA intelligence.	Text	Rule-based architecture
Saha et al. [34]	Empathic CA.	Dataset: Conversations between the depressed support seekers.	Generate empathic and motivational responses.	Response ratings by users for fluency, adaptability, and motivation.	Text	Transformer architecture.
Agnihotri et al. [35]	Topic-driven and Affective CA.	Dataset: ‘ScenariOSA’, with affective state labels.	Tackle the emotional and contextual relevance for mental wellbeing.	Response ratings for emotional relevance.	Text	Transformer architecture.
Rani et al. [5]	Saarthi	None	None	None	Text	Unspecified-ML architecture.
Alazraki et al. [36]	An Empathic AI Coach.	23 participants through crowd-working websites.	Achieve a high level of engagement during virtual therapy sessions.	Survey results: User ratings of empathy, expert ratings of fluency.	Text	Hybrid architecture.
Gundavarapu et al. [37]	A CA companion.	Dataset: created using	Provide emotional support,	None	Text	Neural network architecture.

		sources like Wikipedia.	without judgement.			
Mishra et al. [38]	Counselling CA.	A novel dataset conversational dataset.	Provide MH and legal counselling.	Survey results: User ratings of empathy.	Text	Transformer architecture.

Six studies were conducted in the USA (32%) and six in India (32%). Single papers from Australia, Canada, China, The Philippines, Poland, Switzerland, and the UK were also included (5% each). The year of publication is summarised in Multimedia Appendix 3 indicating a sharp rise in the number of publications since 2022. Most studies, 14 out of 19 (74%), described both design and human evaluations. The types of study designs included, nine (48%) cross-sectional studies out of 19, five (26%) Randomised Controlled Trials (RCTs), four (21%) quasi-experimental designs and one (5%) qualitative study. Only five of 19 (26%) studies referred to an explicit definition of empathy, as summarised in Table 2.

Table 2: Definitions of empathy

Study	**Definition of empathy**
Jiang et al. [23]	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. [25]	Roger’s definition of empathy: Level 1 – Responding to an individual’s explicitly expressed meaning and feelings with a simple repetition of basic understanding. Level 2 – Respond 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, and then capturing the core of the emotion and responding to the patient’s desire with affirmations. Level 4 – When the person is suppressing their feelings or not feeling them 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. [31]	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. [34]	Empathy or empathic interactions is the ability to feel the emotions and experiences of others [39].
Alazraki et al. [36]	Definition of empathy by Barrett-Lennard [40]: First phase - where the listener sympathises and resonates with what is

	being expressed by the speaker. Second phase - in which 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' (47%), to 'conversational agent' (32%), 'dialog system' (11%), 'virtual assistant' (5%) or 'conversational AI agent' (5%). The mode of interaction chosen by most of the CA designs, 17 out of 19 (89%), was text (e.g. live chat, symptom checker and text-based counselling), with voice interactions being used in interactive avatar and counselling roles in two (11%).

Below we consider the technical designs utilised 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 Conversational Agents

The types of CA architectures (or engines) considered by the authors included a mix of recent technologies as summarised in Figure 2, with ML-based architectures used in 12 out of 19 cases (63%). The transformer-based engine, that learns meaning from context, was employed in seven studies (37%), sometimes in the form of a large language model (LLM). A minority of the papers, three out of 19 (16%), did not specify the type of engine employed 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 are shown in Multimedia Appendix 4. Figures in Multimedia Appendix 4 also illustrate how a single engine and a hybrid engine work with user input to provide an empathic response.

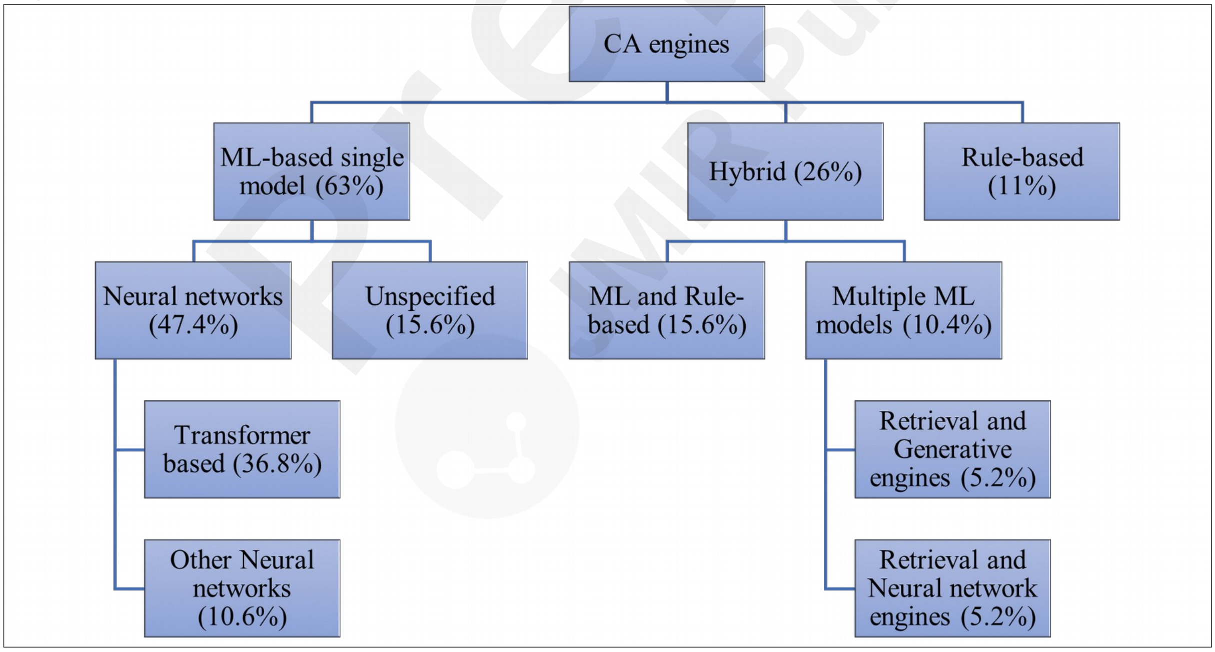

CA engines
ML-based single model (63%)
Hybrid (26%)
Rule-based (11%)
Neural networks (47.4%)
Unspecified (15.6%)
ML and Rule-based (15.6%)
Multiple ML models (10.4%)
Transformer based (36.8%)
Other Neural networks (10.6%)
Retrieval and Generative engines (5.2%)
Retrieval and Neural network engines (5.2%)

Figure 2: Types of CA Architectures

Transformer-based engines included Bidirectional Encoder Representations Transformer (BERT), SBERT, RoBERT, Generative Pre-trained Transformer (GPT-2) and sequence-2-sequence models. Other neural network architecture-based CA designs (11%) were incorporated in two papers [28, 37]. Five publications (26%) considered hybrid models. Two of these hybrid models applied a ML model to capture user emotion and then applied a rule-based algorithm to generate appropriate responses in dialogue management [26-29]. For example, EMMA gathered mobile sensor data to infer user mood, and then assigned users to appropriate wellness interventions [26]. Once assigned, the CA then responded with emotionally expressive responses selected at random from a set of pre-scripted phrases using a rule-based approach [29]. In another example, VHope, a virtual therapist, employed a hybrid model containing a retrieval model that deciphered user input combined with a generative model to elicit empathic responses [30].

The three papers (16%) utilising unspecified architectures commenced with Natural Language Processing (NLP) before employing different ML approaches. In one example, continuous emotional support via remote mental healthcare monitoring and personalised assistance was provided [31]. MH monitoring was performed via scheduling activities that were meaningful to each user, sending out reminders as encouragement, and forwarding satisfaction surveys to receive feedback.

Two publications (11%) 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 pre-defined scripts using a rule-based decision tree [33]. In the next section, we will discuss the technical performance of the CAs reviewed.

Summary Assessment of the Technical Performance of Conversational agents in terms of Classification Accuracy

The accuracy of these designs in detecting and responding to user emotions appropriately is summarised in Table 3. Technical evaluations of the CA designs usually involved comparisons with a “gold standard”, using data not previously used for training the CA.

Table 3: Measures used for evaluating the technical performance of CA designs.

Type of CA Assessment	**Assessment of user emotions/CA responses**	**Accuracy Measure**
Classification of sentiment and issues	User emotions	Mathews Correlation Coefficient = 0.857 [25]
Classification of Valence and Arousal	User emotions	Accuracy of Valence= 80.4% [26] Accuracy of Arousal = 50.4% [26]
Classification of recommended resources (for patients)	User emotions	F1 score = 0.87 [29]

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% [35]
		Precision = 0.954 [35]
		Recall = 0.947 [35]
		F-1 score = 0.95 [35]

Classification for empathy function	User emotions	Accuracy = 80.18% [36] F1 score = 80.66% [36] Weighted Accuracy (W-ACC) = 0.977 [38] Macro F1 score = 0.972 [38]
Prediction of Valence and Arousal	User emotions	Accuracy of Valence = 82.2% [26] Accuracy of Arousal = 65.7% [26]
Accuracy of the	CA responses	Bilingual Evaluation Understudy (BLEU)

response generation		score = 0.126 [28]
		BLEU-1 score (Focused on a single word) = 0.161 [34]
		Perplexity score = 50.90 [34]
		Recall Oriented Understudy for Gisting Evaluation – Longest Common Sequence (ROUGE-L) score = 0.124 [34]
		embedding-based metrics:
		Average = 0.733 [34]
		Extrema = 0.377 [34]
		Greedy = 0.478 [34]

Emotion prediction	User emotions	Accuracy - Correctly predict the next emotion as positive or negative = 79% [29] Correct emotion out of all emotions was predicted =63% [29]
Performance of the language model	CA responses	Perplexity score = 9.977 [30] Perplexity score = 1.91 [25] Response length = 18.71 [25]
Emotion recognition	User emotions	Accuracy = 94.96% [36] F1 score = 95.10% [36]

A technical evaluation of empathic CA performance was conducted in 17 of 19 papers reviewed, however, only 10 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 centred around the ability of the CA to discern user emotions correctly and to respond appropriately. Four papers focused on the CA responses during the technical assessments while the rest of the studies considered user emotions. A variety of measures were employed for each such assessment, highlighting the diversity in evaluation methodologies across studies. These metrics are categorised in detail under the type of CA performance in Multimedia Appendix 5.

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 RoBERTa transformer model, which was built integrating three classifiers for politeness, counselling strategy and empathic feedback, achieved good results overall. This empathy classifier achieved excellent performance with an accuracy score of 0.977 and 0.972 for W-ACC and F1 score respectively [38]. In a second study, a topic-driven classification model utilised a pre-trained GPT-2 transformer model for generating controlled responses, and the model accomplished relatively high scores of accuracy (95%), precision (0.954), recall (0.947) and an F1 score of 0.95 [35].

However, high accuracy and a more nuanced response generation was consistently apparent in all the CAs utilising hybrid architectures [26, 29, 30, 32, 36], 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 Conversational Agents

Most of the reviewed studies, 16 out of 19 (84%), 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. These human evaluations often described more complex evaluation designs as explained below.

The human evaluation was performed by CA users in most cases (75%), while experts in the field of

MH contributed to the process of assessing the CA in the remaining studies (25%). Table 4 summarises the empathy measures used in these papers.

Alazraki and colleagues conducted a cross-sectional study with 23 volunteers and two clinicians who engaged with a web-based chatbot platform using four pre-scripted conversations of different CA personas [36]. An anonymous online questionnaire collected participant feedback regarding the level of empathy displayed by the chatbot, engagement levels, and ability 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 [30] asked three 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. Based on expert feedback, 67% of the CA responses were relevant, 78% seemed human, and 70% were empathic.

In a 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 non-expressive chatbot, with two weeks of monitoring in each case [26]. The participants engaging with EMMA showed higher frequency of interactions and responded quicker compared to the control group. The feedback of the users was also 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 wellbeing [23]. The author conducted 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 to human-AI interactions through variations in cognitive empathy, affective empathy and empathic responses.

Table 4: Measurement of Empathy in CAs

Author	The method of empathy measurement	How was empathy measured?	Who did the evaluation?	Evaluation results
Jiang et al. [23]	Self -reports ▪ In-depth interview responses ▪ Multiple response ratings	Using Robot's Perceived Empathy (RoPE) Scale-Binary responses. Questionnaire of Cognitive and Affective Empathy (QCAE).	Replika users provided the empathy ratings.	Perceived cognitive empathy was higher than perceived affective empathy.
Beredo and Ong. [30]	Self-reports ▪ Response ratings	Affect criterion/empathy was measured by using a binary scale of 0 (No) -1(Yes)	Evaluated by three experts who studied and practice psychology	Responses were rated 79% empathic.
Alazraki et al.	Self-reports ▪ Multiple	Multiple ratings to evaluate Perceived	Evaluated by users. Two separate clinicians	When interacting with the Kai

[36]	response ratings	level of empathy – each rated from strongly disagree to strongly agree on a 5-point Likert scale.	specialised in MH also evaluated the chatbot personas.	persona, 75.0% agreed that the bot was empathic. Interaction with other study personas achieved a 56% empathic rating.
Mishra et al. [38]	Self-reports ▪ Response ratings	A single 5-point Likert scale.	Six evaluators rated each dialogue interaction for empathy. Cross-validated for quality by government-run institutions.	Average empathy rating = 57.0%.
Agnihotri et al. [35]	Self-reports ▪ Response ratings	Emotional relevance is rated using a single 5-point Likert scale.	Evaluated by three human annotators- male non-native English speakers from a technical university with an average Age of 21.	When an empathic response generator is used: Emotional relevance = 61.4% When a topic classifier was added: Emotional relevance= 43.0%

Risk of Bias and Quality Assessment Results

Included RCTs showed a low risk of bias on the ROB-2 tool. Of the 14 non-randomised studies included in the review, all showed a moderate to high risk of bias. Five [29], [34-36], [38] were moderately biased, and one [30] was seriously biased according to the ROBINS-I tool. The JBI quality assessment results were generally low when only the design component of the studies was assessed, with 32% of the papers receiving a score of zero. However, an overall moderate quality was seen in publications when both the design and implementation stages were appraised. Multimedia Appendix 6 shows the quality assessment results.

Discussion

The study and utilisation of CA technology has been the subject of extensive research across many fields such as education, customer service and the healthcare sector. There are also reviews focusing on AI-based CAs, their effectiveness, and their impact in the realm of mental healthcare [17, 18, 41]. While these reviews offer significant insights into AI based CA designs in mental healthcare, the importance of empathy is not central. Although these reviews suggest the need for empathy within CA innovations in mental healthcare, 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.

Conversational agent designs

This review has found that most researchers used a 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 categorical information to managers and mental healthcare 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 more than one model appears to facilitate a more robust model output. This is supported by superior accuracies reported in the classification and prediction tasks achieved by hybrid architectures. The hybrid model of Adikari (2022) achieved the highest accuracy of 87% (F1 score = 0.87) in recommending a resource based on the concerns expressed by the patients [29]. However, the highest accuracy in emotion recognition (95% accuracy in identifying sadness, anger, fear and happiness) was obtained by Alazraki (2021)[36]. These combined features of high accuracy and improved user experience probably make theses the best performing CAs within the review.

While the use of such robust Large Language Models (LLMs) has significantly improved language-based CA technology, it is important to recognise that these models are not without disadvantages [42]. These models have been found to perpetuate biases with regards to gender, race and MH conditions present in the training data [43, 44]. 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 for user trust, the credibility of the empathic CA and user wellbeing. Such biases can be mitigated by ensuring that the training datasets represent diverse gender categories, races and cultural backgrounds, and that advanced technical approaches are used to detect and minimise any such biases in the training data [45-47].

Ethical and privacy concerns associated with these LLMs are critical [48, 49]. Making sure that ethical guidelines centred 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 [50]. These strategies are especially important for an empathic CA design dealing with users seeking mental healthcare. Any breaches of privacy and ethical guidelines pose a high risk to user mental wellbeing as well as user trust 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 [49].

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, optimisation techniques and hardware that supports the requirements of these AI models has facilitated a decrease in execution times [52].

Human evaluations of Conversational agents

Among the reviewed publications, human evaluation of chatbots was common. However, only 26% of the studies used an RCT design to assess the CA platform. Random assignment to 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 a 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 experience with CAs could be an important confounding factor. Based on 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 favourable 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 upon 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 studies were evaluated as satisfactory by end-users. However, there were only five papers that provided quantitative evaluations of CA empathy, and only five of the 19 papers reviewed provided a definition of empathy, a significant omission.

Since 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 has been described by WHO [8]. Ideally standardised 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 employed. 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 two papers in this review [4, 31] used the PHQ as a measure of depressive symptoms as their measure of MH outcomes, while two other papers considered stress levels in their evaluation [25, 27] .

Furthermore, the human evaluations were mostly conducted by study participants. Experts and professionals in the field of mental healthcare were rarely consulted. There is a need for greater consultation with focus groups and user groups to ensure that CA design best reflects the needs of all stakeholders [24]. Future research in this area should also consider an iterative design framework, incorporating co-design and co-evaluation of prototypes involving all stakeholders [24].

In summary, there were deficiencies in all the human evaluations included in this review. Only five papers in this review included a direct evaluation of CA empathy in the design, while the rest were more concerned with general user satisfaction. Only two of these studies used multiple rating scales to measure the level of empathy portrayed by a CA and only one of these (Alazraki et al.) considered evaluations by both users and clinicians. However, there were four 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 four studies where voice data was 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 also 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 mental healthcare facilities such as helpline services, patient triage, and emergency services [23, 25].

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], including noise cancellation technologies [63] and training datasets reflecting diverse human demographic features for algorithm development, are some of the challenges if CAs are to provide effective vocal interaction in real-time.

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 non-judgemental and sensitive interaction for users within a safe environment. Improved empathy portrayal by the CA would help to enhance user engagement and CA acceptability, helping to reduce the gap between the demand and supply of available crisis helpline service.

Summary

This review confirms that empathy is an important characteristic for CA implementation for mental healthcare. 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 which 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 the review presented a mix of methods which made it challenging to compare and analyse 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 varied and a lack of empathy definitions and standard measures for perceived empathy made study comparisons difficult.

The quality assessment of the studies emphasised the need for complete reporting of CA designs as well as rigorous evaluation. Deficiencies in these areas meant that the quality assessments for several papers were low. Evaluation guidelines were often missing which made it challenging to appraise the performance of these systems. Classification accuracy and/or 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 assessment used in mental healthcare. 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 utilised the latest technologies in CA architectures, including designs developed using ML-based transformer engines (e.g. 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. Secondly 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 mental healthcare would benefit by involving co-creation activities, explicit definitions of empathy, and effective evaluation of empathy using standardised empathy scales, also 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 mental healthcare 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 is funded by Swinburne University of Technology. 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 article.

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial Intelligence
BERT: bidirectional encoder representations Transformer
BLEU: Bilingual evaluation understudy
CA: Conversational Agent
GPT: generative pre-trained Transformer
JBI: Joanna Briggs Institute
LLM: Large language Models

MH: Mental Health
ML: Machine learning
NLP: Natural Language Processing
PHQ: Patient Health Questionnaire
QCAE: Questionnaire of cognitive and affective empathy
ROB: risk of bias
ROBINS-I: risk of bias in non-randomised studies of interventions
RoPE: the Robot's perceived empathy
ROUGE: Recall oriented understudy or gisting evaluation
PRISMA: preferred reporting items for systematic reviews and meta-analyses
RCT: randomised controlled trials
WHO: World Health Organisation
W-ACC: Weighted accuracy

Multimedia Appendix 1: Screening process and study characteristics

Multimedia Appendix 2: PRISMA checklist 2020

Multimedia Appendix 3: Evolution of Conversational Agent (Year-by-Year)

Multimedia Appendix 4: Detailed summary of Conversational Agent types

Multimedia Appendix 5: Results of Conversational agent evaluations

Multimedia Appendix 6: Risk of Bias and Quality Assessment

Multimedia Appendix 7: Dictionary of technical terms

References

1. Organization WH. Mental Health Atlas 2020.2020 [ISBN: 978-92-4-003670-3]
2. Schick A, Feine J, Morana S, Maedche A, Reininghaus U. Validity of chatbot use for mental health assessment: experimental study. JMIR mHealth and uHealth. 2022;10(10):e28082 [doi: 10.2196/28082]
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 mental health. 2017;4(2):e7785 [doi:10.2196/mental.7785]
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 and Uhealth. 2018 Nov;6(11) [PMID: WOS:000451080100001] [doi: 10.2196/12106]
5. Rani K, Vishnoi H, Mishra M, editors. A Mental Health Chatbot Delivering Cognitive Behavior Therapy and Remote Health Monitoring Using NLP And AI. 2023 International Conference on

Disruptive Technologies (ICDT); 2023 11-12 May 2023 [doi: 10.1109/ICDT57929.2023.10150665]
6. Persons B, Jain P, Chagnon C, Djamasbi S, editors. Designing the Empathetic Research IoT Network (ERIN) Chatbot for Mental Health Resources. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2021 [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. Telemedicine and e-Health. 2023 [doi:10.1089/tmj.2023.0313]
8. Hojat M, Hojat M. A definition and key features of empathy in patient care. Empathy in health professions education and patient care. 2016:71-81 [ISBN: 978-3-319-27625-0]
9. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Transactions on Interactive Intelligent Systems (TiiS). 2022;12(2):1-39 [doi: 10.1145/3485874]
10. He YH, Yang L, Qian CL, Li T, Su ZY, Zhang Q, et al. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. Journal of Medical Internet Research. 2023 Apr;25 [doi: 10.2196/43862]
11. Kraus M, Seldschopf P, Minker W, editors. Towards the development of a trustworthy chatbot for mental health applications. MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27; 2021: Springer [ISBN: 978-3-030-67835-7]
12. Kallivalappil N, D'souza K, Deshmukh A, Kadam C, Sharma N, editors. Empath. ai: a Context-Aware Chatbot for Emotional Detection and Support. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023: IEEE [doi: 10.1109/ICCCNT56998.2023.10306584]
13. Lin S, Lin L, Hou C, Chen B, Li J, Ni S, editors. Empathy-Based communication Framework for Chatbots: A Mental Health Chatbot Application and Evaluation. Proceedings of the 11th International Conference on Human-Agent Interaction; 2023 [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 Review of Medical Devices. 2021 2021/12/03;18(sup1):37-49 [doi: 10.1080/17434440.2021.2013200]
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. 2021;2(3):Article 25 [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. Frontiers in digital health. 2020;2:576361 [doi: 10.3389/fdgth.2020.576361]
17. Li H, Zhang R, Lee Y-C, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digital Medicine. 2023;6(1):236.[doi: 10.1038/s41746-023-00979-5]
18. Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. Jmir Mental Health. 2019 Oct;6(10)[doi: 10.2196/14166]
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. Journal of medical Internet research. 2021;23(1):e17828. [doi: 10.2196/17828]
20. Gyant. The History and Evolution of Conversational AI. 2021 URL:https://gyant.com/the-

history-and-evolution-of-conversational-ai/[accessed 2023-03-13]

21. Gotschall T. EndNote 20 desktop version. Journal of the Medical Library Association: JMLA. 2021;109(3):520 [doi: 10.5195/jmla.2021.1260]

22. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: Systematic reviews of etiology and risk. Joanna briggs institute reviewer's manual The Joanna Briggs Institute. 2017;5:217-69 [doi: 10.46658/jbimes-20-08]

23. Jiang QL, Zhang YD, Pian W. Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. Information Processing & Management. 2022 Nov;59(6) [PMID: WOS:000864696600001] [doi: 10.1016/j.ipm.2022.103074]

24. Brocki L, Dyer GC, Gładka A, Chung NC, editors. Deep Learning Mental Health Dialogue System. 2023 IEEE International Conference on Big Data and Smart Computing (BigComp); 2023 13-16 Feb. 2023 [doi: 10.1109/BigComp57234.2023.00097]

25. Trappey AJC, Lin APC, Hsu KYK, Trappey CV, Tu KLK. Development of an Empathy-Centric Counseling Chatbot System Capable of Sentimental Dialogue Analysis. Processes. 2022;10(5) [doi: 10.3390/pr10050930]

26. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K, editors. EMMA: An Emotion-Aware Wellbeing Chatbot. 2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019; 2019 [doi: 10.1109/ACII.2019.8925455]

27. Meng J, Dai YN. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? Journal of Computer-Mediated Communication. 2021;26(4):207-22 [doi: 10.1093/jcmc/zmab005]

28. Goel R, Vashisht S, Dhanda A, Susan S, Ieee, editors. An Empathetic Conversational Agent with Attentional Mechanism. 11th International Conference of Computer Communication and Informatics (ICCCI); 2021 Jan 27-29; Sri Shakthi Inst Engn & Technol, Coimbatore, INDIA; 2021 [doi: 10.1109/ICCCI50826.2021.9402337]

29. 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 Generation Computer Systems. 2022;126:318-29 [doi: 10.1016/j.future.2021.08.015]

30. Beredo JL, Ong EC, editors. A Hybrid Response Generation Model for an Empathetic Conversational Agent. 2022 International Conference on Asian Language Processing, IALP 2022; 2022 [doi: 10.1109/IALP57159.2022.9961311]

31. 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 (14248220). 2022;22(10):3653- [PMID: 157239582] [doi: 10.3390/s22103653]

32. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. Journal of Medical Internet Research. 2018;20(6) [doi: 10.2196/10148]

33. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K, Ieee, editors. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. 8th International Conference on Affective Computing and Intelligent Interaction (ACII); 2019 Sep 03-06; Cambridge, ENGLAND; 2019 [doi: 10.1109/ACII.2019.8925433]

34. Saha T, Gakhreja V, Das AS, Chakraborty S, Saha S, Acm, editors. Towards Motivational and Empathetic Response Generation in Online Mental Health Support. 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR); 2022 Jul 11-15; Madrid, SPAIN; 2022 [doi: 10.1145/3477495.3531912]

35. Agnihotri M, Pooja Rao SB, Jayagopi DB, Hebbar S, Rasipuram S, Maitra A, et al., editors. Towards generating topic-driven and affective responses to assist mental wellness. Lecture Notes in

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2021 [doi: 10.1007/978-3-030-68790-8_11]

36. Alazraki L, Ghachem A, Polydorou N, Khosmood F, Edalat A, Ieee Comp SOC, editors. An Empathetic AI Coach for Self-Attachment Therapy. 3rd IEEE International Conference on Cognitive Machine Intelligence (IEEE CogMI); 2021 Dec 13-15; Electr Network; 2021 [doi: 10.1109/CogMI52975.2021.00019]

37. Gundavarapu MR, Saaketh Koundinya G, Bollina Devi Sai T, Kidambi Sree G, editors. Empathic Chatbot: Emotional Astuteness for Mental Health Well-Being. Smart Innov Syst Technol; 2022 [doi: 10.1007/978-981-19-2719-5_65]

38. Mishra K, Priya P, Ekbal A, editors. Help Me Heal: A Reinforced Polite and Empathetic Mental Health and Legal Counseling Dialogue System for Crime Victims. Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023; 2023 [doi: 10.1609/aaai.v37i12.26685]

39. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: An updated meta-analysis. Psychotherapy. 2018;55(4):399 [doi: 10.1037/pst0000175]

40. Barrett-Lennard GT. The empathy cycle: Refinement of a nuclear concept. Journal of counseling psychology. 1981;28(2):91 [doi: 10.1037/0022-0167.28.2.91]

41. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie. 2019 Jul;64(7):456-64.[doi: 10.1177/0706743719828977]

42. Xu XH, Yao BS, Dong YZ, Gabriel SD, Yu H, Hendler J, et al. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. Proceedings of the Acm on Interactive Mobile Wearable and Ubiquitous Technologies-Imwut. 2024 Mar;8(1) [doi: 10.1145/3643540]

43. Salutari F, Ramos J, Rahmani HA, Linguaglossa L, Lipani A, editors. Quantifying the bias of transformer-based language models for african american english in masked language modeling. Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2023: Springer.[doi: 10.1007/978-3-031-33374-3_42]

44. 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 arXiv:211015733. 2021.[doi: 10.48550/arXiv.2110.15733]

45. Kamboj P, Kumar S, Goyal V, editors. Measuring and Mitigating Gender Bias in Contextualized Word Embeddings. 2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS); 2023: IEEE.[doi: 10.1109/ICBDS58040.2023.10346586]

46. Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. arXiv preprint arXiv:211008527. 2021.[doi: 10.48550/arXiv.2110.08527]

47. Gira M, Zhang R, Lee K, editors. Debiasing pre-trained language models via efficient fine-tuning. Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion; 2022.[doi: 10.18653/v1/2022.ltedi-1.8]

48. Kumar A, Singh S, Murty SV, Ragupathy S. The Ethics of Interaction: Mitigating Security Threats in LLMs. arXiv preprint arXiv:240112273. 2024.[doi: 10.48550/arXiv.2401.12273]

49. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. Cognitive Computation. 2024:1-23.[doi: 10.1007/s12559-024-10285-1]

50. 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. Telemedicine and e-Health. 2024;30(3):722-30.[doi: 10.1089/tmj.2023.0313]

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. Digital health. 2023;9:20552076231183542.[doi: 10.1177/20552076231183542]

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 arXiv:240302310. 2024.[doi: 10.48550/arXiv.2403.02310]

53. Santhanam S, Karduni A, Shaikh S, editors. Studying the effects of cognitive biases in evaluation of conversational agents. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020.[doi: 10.1145/3313831.3376318]

54. Gnewuch U, Morana S, Adam MTP, Maedche A. Opposing Effects of Response Time in Human-Chatbot Interaction The Moderating Role of Prior Experience. Business & Information Systems Engineering. 2022 Dec;64(6):773-91. [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. International Journal of Behavioral Nutrition and Physical Activity. 2020;17:1-17.[doi: 10.1186/s12966-020-00938-3]

56. Iyer R, Nedeljkovic M, Meyer D. Using Voice Biomarkers to Classify Suicide Risk in Adult Telehealth Callers: Retrospective Observational Study. JMIR Ment Health. 2022;9(8):e39807 [PMID: 35969444] [doi: 10.2196/39807]

57. Iyer R, Nedeljkovic M, Meyer D. Using vocal characteristics to classify psychological distress in adult helpline callers: retrospective observational study. JMIR formative research. 2022;6(12):e42249.[doi: 10.2196/42249]

58. Schaeffler F, Jannetts S, Beck JM, editors. Reliability of clinical voice parameters captured with smartphones–measurements of added noise and spectral tilt. Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH, Graz, Austria, 15-19 September 2019; 2019: ISCA.[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. Journal of personality and social psychology. 2016;111(5):686.[doi: 10.1037/pspi0000066]

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. Hormones and behavior. 2016;81:84-96.[doi: 10.1016/j.yhbeh.2016.03.001]

61. Israelsson A, Seiger A, Laukka P. Blended Emotions can be Accurately Recognized from Dynamic Facial and Vocal Expressions. Journal of Nonverbal Behavior. 2023;47(3):267-84.[doi: 10.1007/s10919-023-00426-9]

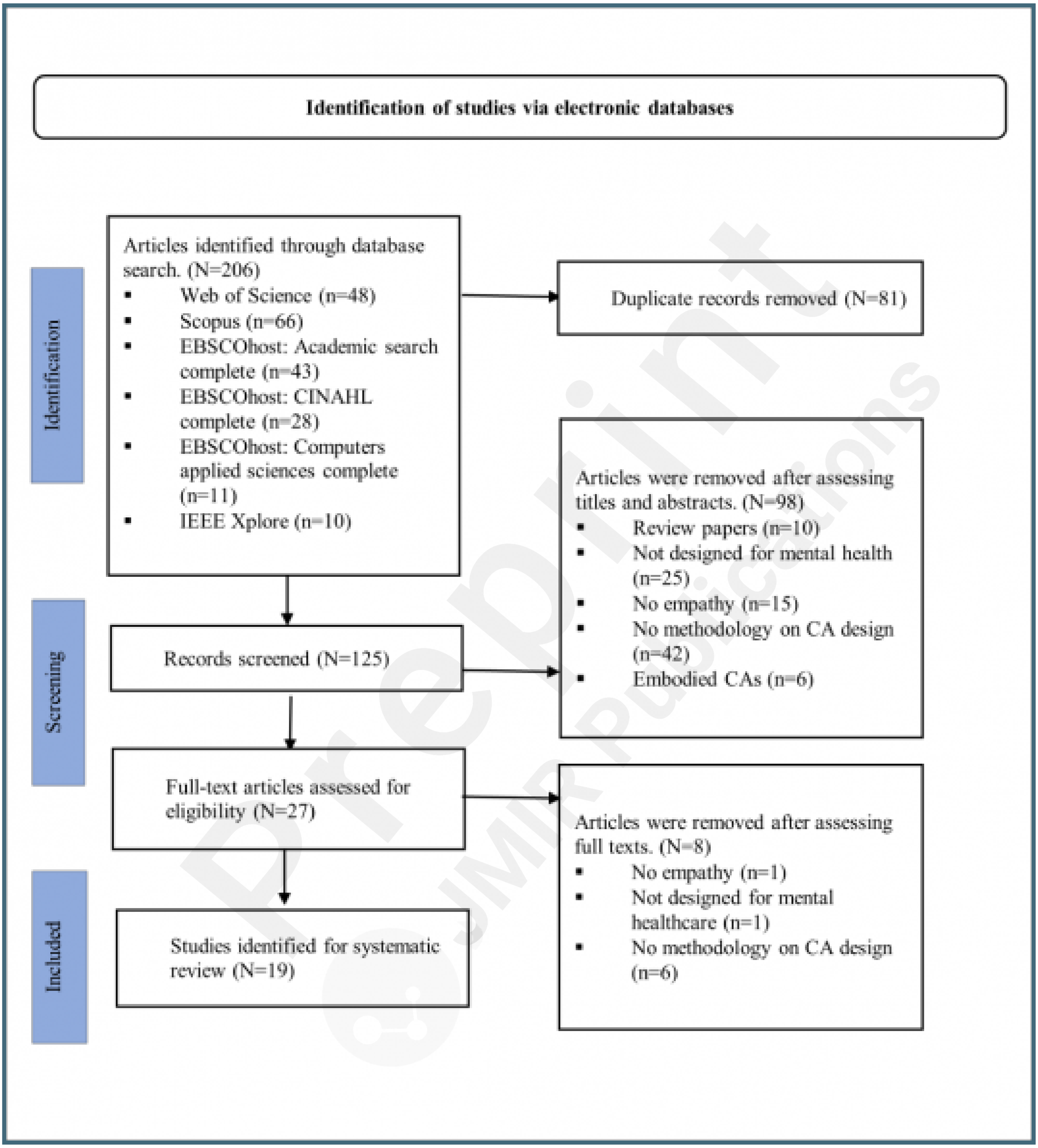
62. Busquet F, Efthymiou F, Hildebrand C. Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices. Behavior Research Methods. 2024;56(3):2114-34. [doi: 10.3758/s13428-023-02139-9]

63. Padmapriya J, Sasilatha T, Karthickmanoj R, Aagash G, Bharathi V, editors. Voice extraction from background noise using filter bank analysis for voice communication applications. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV); 2021: IEEE.[doi: 10.1109/ICICV50876.2021.9388453]

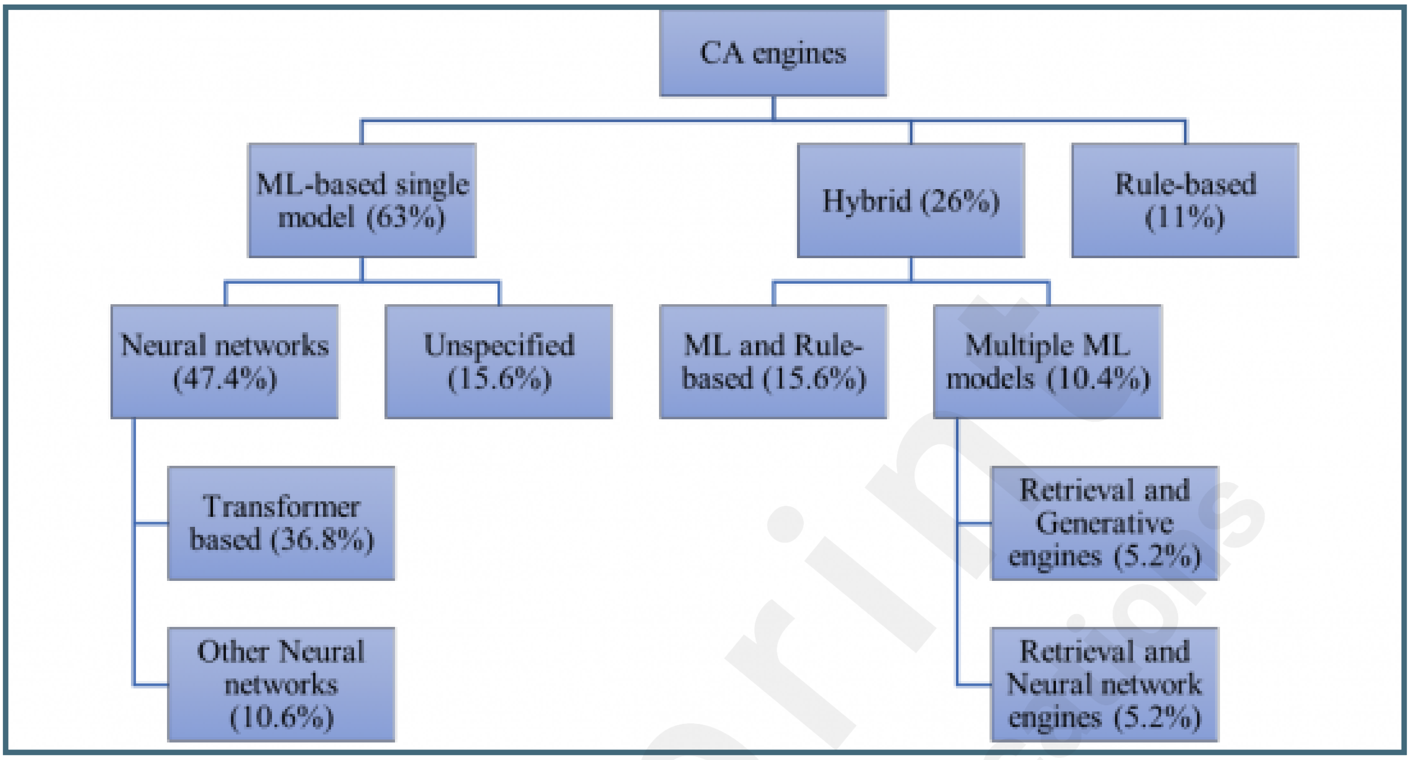
Supplementary Files

Figures

PRISMA Procedure Applied.


Identification of studies via electronic databases
Identification
Screening
Included
Articles identified through database search. (N=206)
▪ Web of Science (n=48)
▪ Scopus (n=66)
▪ EBSCOhost: Academic search complete (n=43)
▪ EBSCOhost: CINAHL complete (n=28)
▪ EBSCOhost: Computers applied sciences complete (n=11)
▪ IEEE Xplore (n=10)
Duplicate records removed (N=81)
Records screened (N=125)
Articles were removed after assessing titles and abstracts. (N=98)
▪ Review papers (n=10)
▪ Not designed for mental health (n=25)
▪ No empathy (n=15)
▪ No methodology on CA design (n=42)
▪ Embodied CAs (n=6)
Full-text articles assessed for eligibility (N=27)
Articles were removed after assessing full texts. (N=8)
▪ No empathy (n=1)
▪ Not designed for mental healthcare (n=1)
▪ No methodology on CA design (n=6)
Studies identified for systematic review (N=19)

Types of CA Architectures.


CA engines
ML-based single model (63%)
Hybrid (26%)
Rule-based (11%)
Neural networks (47.4%)
Unspecified (15.6%)
ML and Rule-based (15.6%)
Multiple ML models (10.4%)
Transformer based (36.8%)
Other Neural networks (10.6%)
Retrieval and Generative engines (5.2%)
Retrieval and Neural network engines (5.2%)

Multimedia Appendixes

Screening process and study characteristics.
URL: http://asset.jmir.pub/assets/476fcfa0909f825f711fc25a10cbcf40.docx

PRISMA checklist 2020.
URL: http://asset.jmir.pub/assets/096a95bea7b41b4c403a4a2bbddadc59.docx

Evolution of Conversational Agent (Year-by-Year).
URL: http://asset.jmir.pub/assets/2f40ac66168ed4a6c3753c31a6dd64c9.docx

Detailed summary of Conversational Agent types.
URL: http://asset.jmir.pub/assets/9646ce2e1ac619760a75dad7ee7692ae.docx

Results of Conversational agent evaluations.
URL: http://asset.jmir.pub/assets/ea0be7094c266abcf0971005e98493b6.docx

Risk of Bias and Quality Assessment.
URL: http://asset.jmir.pub/assets/fe527dc8f19f26f7d7d61f1a1fff65c4.docx

Dictionary of technical terms.
URL: http://asset.jmir.pub/assets/eb5e9ce62fc91265c3314ea2998fd3cb.docx