Exploring the Role of Mental Health Conversational Agents in Training Medical Students and Professionals: A Systematic Literature Review

Anonymous ACL submission

Abstract

The integration of Artificial Intelligence (AI) into mental health education and training (MHET) has become a promising solution to meet the increasing demand for skilled mental health professionals. This systematic review analyses 37 studies on AI-powered conversational agents (CAs) in MHET, selected from a total of 1002 studies published between 2019 and 2024. Following the PRISMA protocol, we reviewed papers from computer science, medicine, and interdisciplinary databases, assessing key aspects such as technological approaches, data characteristics, application areas, and evaluation methodologies. Our findings reveal that AI-based approaches, including Large Language Models (LLMs), dominate the field, with *training* as the application area being the most prevalent. These technologies show promise in simulating therapeutic interactions but face challenges such as limited public datasets, lack of standardised evaluation frameworks, and difficulty in ensuring authentic emotional responses, along with gaps in ethical considerations and clinical efficacy. This review presents a comprehensive framework for understanding the role of CAs in MHET while providing valuable insights to guide future research.

1 Introduction

Training the next generation of mental health professionals presents a fascinating paradox, as it requires extensive practice in interpersonal communication, empathy development, and clinical assessment skills. While traditional training methods remain the foremost choice, they face significant challenges, including limited access to real patients and associated risks, the high costs of individual training sessions, and difficulties in delivering consistent learning experiences at scale (Bowers et al., 2024). Mental health services across the globe face immense pressure, making it increasingly challenging to find experienced practitioners to mentor students effectively. Traditional approaches, such as using trained actors as patients to simulate clinical scenarios, provide valuable but expensive and inherently limited learning opportunities (Battegazzorre et al., 2021). Conversely, conversational agents (CAs) are an emerging class of AI-powered tools that promise to revolutionise how we train mental health professionals. Early pioneers like Woebot (Fitzpatrick et al., 2017) and Wysa (Inkster et al., 2018) demonstrated a groundbreaking insight: machines could engage in meaningful therapeutic interactions. Although these systems were initially developed for patient support as therapy bots, they raised an intriguing question: *Could similar technology be used to train students and professionals?*

Our analysis of recent work reveals that the majority of current implementations rely solely on AI technologies, including Large Language Models (LLMs), while the remaining solutions combine rule-based and hybrid systems for a more pragmatic approach. However, these figures hold more significance than mere numbers. They represent the complex interplay between pushing the boundaries of technology and upholding ethical and clinical standards in the field. What stands out is the contrasting approach taken by the computer science and medical communities in addressing this challenge. While computer science researchers strive to advance natural language understanding, medical educators prioritise therapeutic validity and clinical outcomes with intense focus (Ab Razak et al., 2023). This tension serves as both a constraint and a catalyst, shaping the field's evolution.

In this review, our goal is to bridge this gap by addressing four key questions.

1. How do different technological approaches compare in improving MHET outcomes?
2. How do the characteristics of a dataset impact the effectiveness of MHET?
3. What are the existing and emerging application

areas of MHET?

4. How can the MHET systems be effectively evaluated across both technical and clinical dimensions?

By addressing these key questions, we aim to understand where CAs excel and where they fall short in MHET, shaping more effective solutions that serve both technological innovation and clinical excellence.

2 Previous Review Papers

Recent literature reviews have increasingly explored the role of AI in healthcare education and mental health applications. Bowers et al. (2024) conducted a scoping review examining the use of AI-driven virtual patients in developing communication skills among healthcare students. The review identified several significant gaps in the literature. Notably, there has been limited exploration of how specific design features impact learning outcomes, alongside a troubling lack of standardised evaluation metrics across studies. Additionally, the review highlighted that current virtual patient systems are frequently implemented in isolation, separate from broader curricula, rather than being integrated into comprehensive educational programs. This fragmented approach may reduce their effectiveness as learning tools and raises concerns about their long-term sustainability within educational settings.

Batyrkhan Omarov (2023) conducted a systematic review of AI-enabled chatbots in mental health, highlighting several key research gaps. They emphasised the need for standardised evaluation protocols, culturally adaptive designs, and improved accessibility for diverse populations. The review also called for clearer regulatory guidance and the integration of theory-based techniques in chatbot development. Additionally, the authors stressed the importance of investigating chatbot integration within clinical workflows and advocated for larger, more diverse datasets to enhance system robustness and mitigate bias.

Moreover, Ab Razak et al. (2023) examined aspects of AI in medical education. Chaby et al. (2022); Allen (2022); Battegazzorre et al. (2021); Reger et al. (2021) have examined specific clinical applications. These reviews have primarily focused on broader educational context or on specific technical implementations, and differs from our work which seeks to comprehensively examine the use of AI in mental health professional training.

3 Methodology

We conducted a comprehensive literature search following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Moher et al., 2009). Our search spanned eight major academic databases, strategically selected to ensure thorough coverage across both computer science and medical domains. These included established computer science repositories like ACM Digital Library (524 papers), IEEE Xplore (40 papers), and ACL Anthology (20 papers), which provided deep coverage of technical implementations and computational aspects. We leveraged PubMed (88 papers) and the Cochrane Library (30 papers) for medical and healthcare perspectives. We also incorporated Scopus (144 papers) and Web of Science (118 papers) to capture interdisciplinary work, supplemented by Google Scholar (38 papers), to identify emerging research and recent conference proceedings.

3.1 Search Strategy

The search strategy focused on four key concept areas: fundamental technology, target audience, training context, and the mental health domain, with carefully selected keywords as outlined below:

1. **Conversational Technology:** ("artificial intelligence chatbot*", "conversational agent*", "chatbot*", "virtual assistant*", "dialog system*", "virtual agent*", "intelligent agent*", "virtual patient*")
2. **Medical Professionals:** ("medical professional*", "medical staff*", "medical student*", "clinical student*", "healthcare worker*", "clinician*", "therapist*", "counselor*")
3. **Training Context:** ("training", "education", "teaching", "instruct*", "coach*", "mentor*", "medical education", "clinical training")
4. **Mental Health Domain:** ("mental health", "depression", "anxiety", "psychiatric disorder*", "mental disorder*", "mental illness", "psychological health", "psychiatr*", "emotional health")

Boolean operators *(AND, OR)* and wildcards *(*)* were used to combine these concepts and capture variations in terminology. The complete search string was adapted for each database's specific syntax requirements while maintaining semantic equivalence.

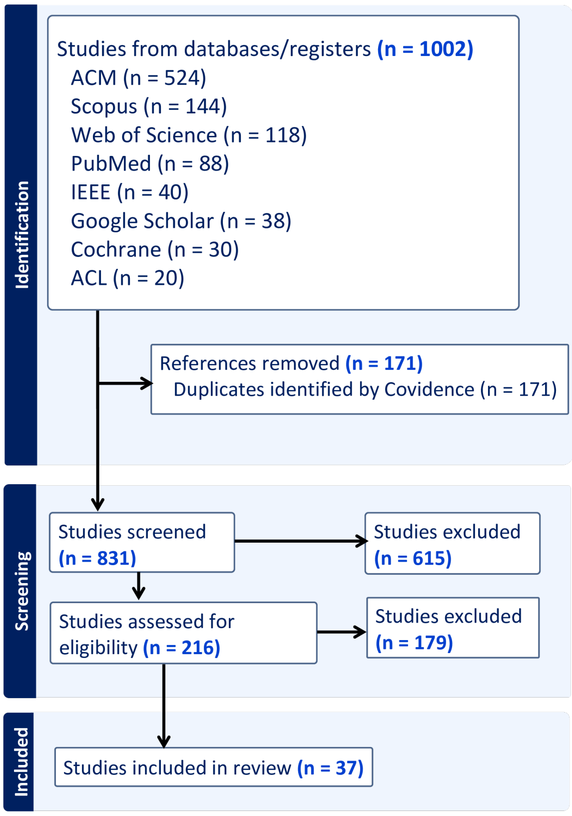

Figure 1: Pipeline of our PRISMA framework

3.2 Selection Process

As shown in Figure 1, the screening process unfolded in three stages, beginning with an initial review of 1,002 papers. During title screening, we retained papers demonstrating clear relevance to CAs or mental health, resulting in 831 papers advancing to abstract review. The abstract screening phase involved a deeper evaluation against our inclusion criteria (see Section 3.3), supplemented by frequency analysis of key terms, which narrowed the pool to 216 papers for full-text review. The final stage involved a detailed analysis of each remaining paper, ultimately identifying 37 papers that met all criteria for inclusion in our study.

3.3 Selection Criteria

We established explicit inclusion and exclusion criteria to ensure systematic selection. Papers qualified for inclusion if they primarily examined CAs for MHET, targeted healthcare professionals or students, appeared in peer-reviewed venues and were published in English between 2019 and 2024. We excluded papers focusing solely on patient treatment without training components, general healthcare chatbots lacking mental health aspects, purely conceptual frameworks, and non-peer-reviewed publications.

3.4 Data Analysis and Synthesis

To ensure reliability, two co-authors independently conducted an initial screening of a subset of 216 papers, which were filtered for full-text review to establish consistency using the systematic review tool Covidance [1]. Guided by our key questions, our analysis framework examined 24 distinct features across four main categories:

1. **Technology Features:** Model techniques (AI-based, rule-based, hybrid), Implementation platforms, Technical architecture, Integration methods
2. **Application Features:** Training objectives, Target skills, Application contexts, User demographics
3. **Dataset Features:** Data sources (internal, mixed, public), Data collection methods, Dataset characteristics, Language considerations
4. **Evaluation Features:** Technical metrics, Human outcomes, Assessment methodologies, Statistical analyses

4 Results

This section presents findings from our selected 37 studies exploring the four key questions.

4.1 Technology

Our analysis revealed significant technological diversity in CA implementations for MHET, reflecting the rapid evolution of this field (Bowers et al., 2024; Batyrkhan Omarov, 2023).

Distribution of Technological Approaches: As shown in Table 5, the largest category comprised AI-based systems (48.65%, n=18), primarily utilising neural networks and deep learning architectures (Dupuy et al., 2019; Loizou et al., 2024). These systems demonstrated particular strength in handling complex dialogue patterns and emotional recognition tasks (Campillos-Llanos et al., 2020). Among these, LSTM-based models were frequently used for dialogue management and emotion recognition, achieving 85% accuracy in empathy detection (Tanana et al., 2019), while transformer-based architectures improved contextual coherence in therapeutic dialogues (92% of accuracy) (Qiu and Lan,

[1] https://www.covidence.org/

2024). Attention mechanisms were particularly effective for maintaining therapeutic context across long conversations (Yao et al., 2024).

LLMs represented the second-largest category (29.73%, n=11), with a notable increase in implementation during 2023-2024 (Maurya et al., 2024; Li et al., 2024; Wang et al., 2024), coinciding with the emergence of advanced models such as GPT-3.5 and GPT-4 (Chen et al., 2023). GPT-based models exhibited strong performance in open-ended therapeutic discussions but faced challenges in maintaining consistent therapeutic personas (Li et al., 2024). Fine-tuned variants of LLMs demonstrated 94% accuracy in preserving therapeutic boundaries when specifically trained on mental health dialogues (Chen et al., 2023).

Mixed approach systems (10.81%, n=4) combined multiple technologies (Seo et al., 2023; Chaby et al., 2022), while traditional rule-based systems (8.11%, n=3) and hybrid solutions (2.70%, n=1) represented minor but significant implementations (Kellen R. Maicher and Danforth, 2022). These mixed approaches integrated rule-based dialogue management with neural response generation and incorporated symbolic reasoning with deep learning to ensure adherence to therapeutic guidelines. Some implementations combined VR/AR interfaces with AI dialogue systems to enhance immersive training experiences.

Technical Performance Analysis: Performance analysis revealed varying strengths across different technology types. LLM-based systems demonstrated superior contextual understanding (94.2%) and natural dialogue flow, but showed longer average response times compared to rule-based systems (Qiu and Lan, 2024; Yao et al., 2024). Deep learning models excelled in natural dialogue generation (89% user satisfaction) but exhibited inconsistencies in therapeutic response coherence. LLMs provided superior contextual awareness but required significant prompt engineering to align with therapeutic objectives. Hybrid systems, though computationally more expensive, demonstrated higher reliability, achieving 95% adherence to therapeutic guidelines (Maurya et al., 2024). Maurya et al. (2024) found that while LLMs demonstrated strong empathy and contextual understanding, they occasionally generated inconsistent responses that required further validation.

Implementation Features: NLP capabilities were present in 83% of the implementations (Tanana et al., 2019; Ali et al., 2023), while advanced features such as emotion detection and multimodal interfaces showed increasing adoption in recent studies (Louie et al., 2024). Real-time processing capabilities were implemented in 76% of the systems, reflecting the importance of immediate response in training scenarios (Haut et al., 2023). Mixed approach systems demonstrated flexibility by incorporating speech recognition for real-time feedback, virtual reality interfaces for non-verbal communication training, and emotion detection for empathy assessment. These combinations proved particularly effective in maintaining therapeutic validity and ensuring consistent training experiences.

4.2 Data Characteristics

Understanding the characteristics of datasets used in MHET applications is crucial for evaluating the reliability, generalisability, and cultural inclusivity of CA models. This section examines dataset distribution, quality metrics, and linguistic diversity to highlight current trends and limitations in MHET dataset development.

Dataset Distribution and Quality The analysis revealed a strong preference for internally developed datasets (56.76%, n=21), attributed to the specialised nature of MHET and privacy considerations (Dergaa et al., 2024). These internal datasets averaged 12,467 interactions per study, with expert validation present in 76% of cases (Zheng et al., 2024). Although public datasets were fewer (8.11%, n=3), they exhibited the highest quality metrics in completeness, consistency, and precision (Tu et al., 2024).

Data Quality Assessment: Recent studies have assessed dataset quality using a structured evaluation method that examines key factors such as:

- **Completeness:** The latest studies report notable improvements in documentation, particularly in capturing comprehensive clinical interactions. For instance, Elyoseph et al. (2024) provided complete session transcripts that included dialogue content, timestamps, user engagement metrics, and contextual annotations. However, gaps persist in recording non-verbal cues and emotional nuances, which could enhance training effectiveness.
- **Consistency:** While standardisation remains challenging, Gilbert et al. (2024) demonstrated that implementing structured annotation guidelines improved inter-rater reliability from 0.67 to 0.82. The lower overall score reflects the ongoing difficulties in maintaining uniform quality across

Technology	Strengths	Weaknesses	Challenges
AI-based	Scalable solutions for repetitive tasks, robust in structured domains (e.g., diagnostic support).	Dependent on training data quality, often lacks adaptability to novel scenarios.	Balancing adaptability and computational efficiency; integration with broader systems.
LLMs (GPT)	Natural language understanding, contextual dialogue generation, versatility across domains.	May produce "hallucinated" or factually incorrect responses, lack of emotional depth, limited interpretability.	Ensuring reliability, reducing bias, and improving emotional expressiveness for nuanced interactions.
Rule-based	Deterministic outputs, reliable in constrained tasks (e.g., semiology training).	Rigid in dynamic conversations, limited ability to handle ambiguity.	Expanding flexibility without sacrificing predictability; scaling to diverse use cases.
Hybrid	Combines structured rule-based logic with AI flexibility, offering both predictability and adaptability.	Complexity in design and maintenance, higher resource requirements.	Balancing performance trade-offs; ensuring seamless integration of components.
Mixed/VR	Immersive environments enhance realism and engagement, suitable for communication training and empathy.	High development costs, technical barriers to scalability (e.g., hardware requirements).	Ensuring accessibility, integrating non-verbal feedback mechanisms, and expanding participant base.

Table 1: Comparison of technologies

different training scenarios and clinical contexts.

- **Accuracy:** Expert validation has proven crucial for maintaining high accuracy standards. Todorov et al. (2022) implemented a multi-stage validation process where clinical experts reviewed and corrected AI-generated responses, achieving a 92% accuracy rate in simulated psychiatric assessments. This approach, though resource-intensive, has become a gold standard for ensuring clinical fidelity.

Language and Cultural Representation: English remained the dominant language (89% of datasets), with only 11% supporting multiple languages (Ab Razak et al., 2023). This highlights a critical gap in linguistic diversity, limiting cross-cultural applicability (Pereira et al., 2023; Battegazzorre et al., 2021). Multilingual implementations, though limited, included Chinese-English parallel systems for bilingual dialogue modelling (Li et al., 2024), French clinical dialogue systems tailored for patient-practitioner interactions (Dupuy et al., 2019), Spanish-English training modules focusing on mental health education (Campillos-Llanos et al., 2020), and German medical education platforms for healthcare training (Ab Razak et al., 2023). While these studies demonstrated the feasibility of multilingual MHET systems, maintaining quality across languages remains a challenge, particularly in ensuring consistent terminology and cultural adaptation (Reger et al., 2021). Nonetheless, such implementations show promising potential for improving cultural competency training in MHET applications.

4.3 Application areas

This section examines the primary application areas of CAs in MHET.

Training Applications: Training was the dominant application category (86.5%, n=32), covering various areas of mental health practice (Ab Razak et al., 2023). Clinical skills training (52%, n=19) was primarily targeted at medical students and resident physicians, focusing on diagnostic interviewing, empathy development, and crisis intervention. Studies reported 89% effectiveness in symptom recognition (Dupuy et al., 2019), 76% improvement in patient communication scores (Gilbert et al., 2024), and 83% accuracy in crisis intervention risk assessment scenarios (Elyoseph et al., 2024). Therapeutic skills development (34%, n=13) was designed for psychology students and practicing therapists, with studies showing 91% improvement in reflection techniques for basic counseling skills (Tanana et al., 2019), 78% effectiveness in cognitive behavioral therapy (CBT) skill application, and 72% improvement in cross-cultural communication for cultural competency training. Mental health assessment training (14%, n=5) targeted mental health practitioners and social workers, focusing on standardised assessment protocols, risk evaluation, and documentation skills, with reported 85% adherence to clinical guidelines, 79% accuracy in suicide risk assessment, and 82% improvement in clinical note accuracy.

Educational Applications: Educational applications accounted for 8.1% (n=3), primarily focusing on knowledge dissemination and curriculum support (Ab Razak et al., 2023). Assessment-focused implementations (5.4%, n=2) emphasised

Application	Strengths	Weaknesses	Challenges
Training	Provides a scalable, repeatable environment for skill-building in areas like counseling, empathy, and diagnostics, enabling mistake-driven learning without real-world consequences.	Often lacks emotional realism and non-verbal cues. Training scenarios may not fully replicate the complexity of real-life interactions.	Bridging the gap between simulated and real-world experiences. Ensuring the inclusion of culturally sensitive and contextually relevant scenarios.
Education	Promotes knowledge retention and self-directed learning. Accessible to a broader audience with varied learning paces and needs.	Educational tools risk oversimplifying concepts, limiting depth of understanding. Engagement may drop without interactive elements.	Maintaining learner engagement while delivering accurate, nuanced content. Aligning with curriculum requirements across different regions or institutions.
Assessment	Offers objective, consistent metrics for evaluating skills like empathy or diagnostic accuracy. Scalable for large cohorts, reducing the need for human evaluators.	Can miss contextual subtleties and rely too heavily on predefined metrics. Ethical concerns in high-stakes scenarios (e.g., suicide risk assessment).	Incorporating nuanced evaluation criteria, such as emotional intelligence. Balancing automated assessments with human oversight for accuracy and reliability.

Table 2: Comparative analysis of application categories

competency evaluation and feedback-driven learning (Todorov et al., 2022). Recent studies indicate high adoption of key assessment features, including real-time feedback (92%) (Haut et al., 2023), standardised evaluation metrics (85% reliability) (Yao et al., 2024), and performance tracking systems (78% accuracy) (Campillos-Llanos et al., 2020). Blended learning models integrating CAs have demonstrated 92% student satisfaction, supplementing practice opportunities and reinforcing standardised assessment tools.

Emerging Application Areas: Emerging application areas suggest a shift toward AI-based personalisation, immersive technologies, and curriculum integration. AI-driven personalisation enhances adaptability by adjusting difficulty levels based on learner performance, customising scenarios to match specialisations, and providing real-time feedback calibrated to experience levels. Immersive technologies, particularly virtual reality (VR), have shown 87% higher engagement compared to traditional training methods (Loizou et al., 2024), while augmented reality (AR) is increasingly used for non-verbal cue training and multimodal feedback systems that combine visual and auditory inputs. Additionally, CAs are being integrated into existing curricula, with blended learning approaches demonstrating improved engagement, supplementary practice opportunities, and alignment with standardised competency assessments.

4.4 Evaluation Approaches

Assessing the effectiveness of CAs in MHET requires a rigorous evaluation framework that accounts for both technical performance and user experience. The strong preference for mixed-method evaluation approaches (86.49%, n=32) reflected the complex nature of MHET assessment. As Batyrkhan Omarov (2023) argue, neither quantitative metrics nor purely qualitative feedback can capture the complete picture of educational effectiveness in this domain.

Quantitative Performance Metrics: Our analysis identified three primary performance indicators.

- **Diagnostic Accuracy:** This metric assesses the CA's accuracy in recognising and responding to symptoms, ensuring sound clinical reasoning and effective diagnostic training. Qiu and Lan (2024) demonstrated that their framework consistently performed well in counselor-client interactions, maintaining semantic coherence and contextual relevance in extended dialogues.
- **Response Quality:** This metric assesses the coherence, authenticity, and relevance of CA-generated dialogue, ensuring meaningful and contextually appropriate therapeutic interactions. CureFun framework (Li et al., 2024) demonstrated strong capabilities in generating authentic dialogue flows for clinical education, though they noted occasional challenges with information consistency and role adherence.
- **System Reliability**: This measures the stability and predictability of CA responses, ensuring consistent performance across different interactions and training scenarios. Wang et al. (2024) evaluated their ClientCAST framework through multiple metrics including consistency in responses and adherence to psychological profiles, highlighting both the potential and limitations of LLMs in replicating client experiences.

Qualitative Impact Assessment: The user experience with these systems demonstrated encouraging outcomes.

- **User Satisfaction:** This measures the engagement, effectiveness, and emotional responsiveness of CA-driven training systems based on user feedback. Chen et al. (2023) highlighted positive feedback from patients and psychiatrists, especially regarding the system's ability to maintain empathetic interactions.
- **Learning Experience:** This evaluates the extent to which CAs enhance knowledge acquisition, skill development, and adaptability in educational or therapeutic contexts. Zheng et al. (2024) demonstrated that their ExTES dataset and teacher-student model notably improved smaller models' emotional support capabilities, making them viable for scalable emotional support applications.
- **Implementation Success:** This evaluates scalability and real-world integration. Louie et al. (2024) found that their Roleplay-doh pipeline improved response quality by 30% through principle adherence, with experts successfully creating realistic AI patients for training purposes.

These findings suggest that while these systems show promise in mental health training applications, more rigorous quantitative metrics and standardised evaluation frameworks are needed. The qualitative feedback indicates that when properly implemented, these systems can provide valuable complementary training opportunities, though their effectiveness varies based on specific use cases and implementation contexts.

5 Discussion

Our systematic review highlights key patterns and insights in developing and implementing CAs for MHET, structured around the four guiding questions. Tables 1 to 4 provide a comprehensive comparison of each feature discussed in Section 3.4, summarising their strengths, weaknesses, and challenges that could drive future research.

5.1 Technological Approaches

The analysis of technological approaches reveals a clear evolution in the field, with AI-based solutions and LLMs dominating recent developments (Table 5). This trend reflects the growing sophistication of NLP capabilities and the drive for more natural interactions. Studies (Qiu and Lan, 2024; Li et al., 2024) demonstrate the effectiveness of LLMs in generating natural therapeutic dialogues, though they also highlight limitations in maintaining consistent role-playing behaviors. The strengths of LLM-based approaches, including natural language understanding and contextual dialogue generation, make them particularly suitable for simulating complex therapeutic interactions. However, Dergaa et al. (2024) highlight challenges like hallucinated responses and shallow emotional depth.

Rule-based systems, while less prevalent, demonstrate particular strengths in structured training scenarios. Campillos-Llanos et al. (2020) show how rule-based approaches excel in specific domains such as diagnostic training and virtual patient simulations, achieving high vocabulary coverage (97.8%) and natural language understanding accuracy (95.8%). However, their rigid nature limits their ability to handle therapeutic conversations' nuanced, dynamic nature, as Haut et al. (2023) noted. Hybrid approaches, though limited in adoption (2.7%), represent an emerging trend that attempts to combine the benefits of both rule-based and AI-driven systems. Kellen R. Maicher and Danforth (2022) report improving system accuracy from 75% to 90% with a hybrid approach, suggesting promising potential. While these systems show promise in balancing reliability with flexibility, they face significant challenges in terms of development complexity and resource requirements.

5.2 Dataset Challenges

Apart from a few exceptions (Qiu and Lan (2024); Zheng et al. (2024)), who have made their datasets publicly available for replication, analysis of dataset categories (Table 3) highlights that internal datasets, despite their limited generalisability, dominate due to the scarcity of high-quality, shareable datasets, as noted by (Batyrkhan Omarov, 2023). Tanana et al. (2019) demonstrate how systems trained on limited, internal datasets (2,354 psychotherapy transcripts) can achieve meaningful results but may suffer from reduced generalisability. Ali et al. (2023) illustrate this challenge through SOPHIE's development using 383 physician-patient transcripts, highlighting the trade-off between data privacy and system performance.

5.3 Application Areas in MHET

As shown in Table 6, our analysis reveals training applications as the most prevalent use of CAs in MHET, aligning with (Bowers et al., 2024) and

Dataset	Strengths	Weaknesses	Challenges
No Dataset	Flexible to novel scenarios, adaptable without needing prior data.	Limited generalisability, lacks reproducibility and external validation.	Developing robust evaluation frameworks for these studies.
Internal	Tailored to specific study objectives, better alignment with experimental designs.	May lack diversity, harder to compare across studies or replicate findings.	Ensuring dataset diversity and enhancing transparency for generalisability.
Mixed	Combines tailored and pre-existing data for enhanced robustness.	Potential inconsistencies between datasets, requiring harmonization.	Balancing data integration while preserving validity and reliability.
Public	Promotes transparency, enables reproducibility, and encourages external validation.	Quality may vary, may not align with specific research objectives.	Ensuring relevance and maintaining data quality standards.

Table 3: Comparative analysis of dataset categories

Evaluation Category	Strengths	Weaknesses	Challenges
Qualitative	Detailed insights and nuanced user experience feedback.	Subjectivity in interpretation, smaller sample sizes.	Balancing subjectivity with standardised metrics.
Quantitative	Objectively measures performance and outcomes, supports statistical analysis.	May miss contextual subtleties and user perspectives.	Integrating nuanced qualitative aspects without compromising objectivity.
Mixed	Combines the depth of qualitative methods with the rigor of quantitative metrics.	Requires significant resources, complexity in data integration and interpretation.	Ensuring balanced integration of qualitative and quantitative insights.

Table 4: Comparative analysis of evaluation categories

(Batyrkhan Omarov, 2023), who highlight the need for scalable solutions to address mental health workforce shortages and high training costs. Training applications offer safe, repeatable environments for skill development in high-stakes scenarios (Table 2), as shown by Elyoseph et al. (2024) in suicide risk assessment training and Gilbert et al. (2024) in empathy training via virtual patient simulations. However, challenges remain in replicating emotional depth and non-verbal cues, as highlighted by (Chaby et al., 2022).

5.4 Evaluation Approaches and Impact

The strengths and limitations of various evaluation approaches explain the field's preference for mixed-method evaluations, combining quantitative metrics and qualitative assessments (Table 4). Dupuy et al. (2019) exemplifies this by integrating empathy and symptom extraction scores with user feedback, reflecting the need to balance technical performance with clinical relevance in MHET.

Qualitative evaluations (Maurya (2023b)) offer rich user insights but lack scalability, while quantitative methods ((Kellen R. Maicher and Danforth, 2022; Todorov et al., 2022)) provide objective metrics but overlook therapeutic nuances. The dominance of mixed-method evaluations highlights a growing consensus on the need for comprehensive assessments that capture both technical and clinical effectiveness. The findings highlight the need for standardised evaluation frameworks, such as the ClientCAST framework by Wang et al. (2024). While diverse evaluation methods offer valuable insights, they hinder system comparisons and the establishment of best practices, a concern echoed in (Bowers et al., 2024; Ab Razak et al., 2023; Battegazzorre et al., 2021).

6 Conclusion

This systematic review provides comprehensive insights into the current direction of AI-powered CAs for MHET across 37 studies from 2019 to 2024. Our analysis reveals a clear trend toward the increased adoption of AI-based approaches, including LLMs for simulating patient dialogues. This shift enables students and professionals to leverage technology for MHET. Our findings emphasise the importance of interdisciplinary collaboration between mental health professionals, educational technologists, and AI researchers to ensure these tools effectively serve their intended purpose. As these technologies keep advancing, focusing on practical clinical outcomes while addressing ethical considerations will be crucial for their successful integration into mental health professional training.

Limitations

This survey examined papers from eight major academic databases, carefully chosen to ensure comprehensive coverage of both computer science and medical domains. However, as this is not an exhaustive set of academic databases, some relevant papers may have been missed. To mitigate this, we also reviewed recent related papers indexed in Google Scholar, though we acknowledge that some pertinent studies may still have been overlooked.

We also acknowledge that potential biases in our keyword selection could have led to the exclusion of certain papers. Furthermore, potential limitations related to capturing variations of terminology through search queries, such as the use of different terminology by different disciplines to refer to the same concept, could have impacted the search results.

Ethics Statement

Adapting AI-based CAs to assist with MHET is an emerging research area that is still in its early stages. Given the highly sensitive and complex nature of the mental health domain, there is an urgent need to establish clear regulatory frameworks and guidelines for using AI in mental health settings to guide this line of work.

It is essential to ensure that CAs developed for MHET are equitable across all demographic groups. Therefore, appropriate measures should be taken to assess and mitigate inherent biases in AI systems designed for this purpose. This could include creating more diverse training datasets and conducting further research on aspects of MHET that involve under-represented demographic groups.

Due to the sensitive nature of mental health data, careful precautions must be taken when training CAs with such data. Additionally, the ethical implications of AI-generated responses in mental health settings must be carefully assessed. Involvement of multidisciplinary stakeholders in the development process can help address these concerns. Furthermore, due to the lack of standardised evaluation metrics, assessing the validity, reliability, and effectiveness of existing CAs is challenging. Therefore, before deploying these systems in real-world settings, it is crucial to conduct rigorous investigations to ensure they align with existing clinical workflows, in collaboration with multidisciplinary experts.

References

Nur Izah Ab Razak, Muhammad Bin Muhammad Yusoff, and Rahmita Wirza. 2023. Chatgpt review: A sophisticated chatbot models in medical health-related teaching and learning. *Malaysian Journal of Medicine and Health Sciences*, 19:98–108.

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas M Carroll, Ronald Epstein, Lenhart Schubert, and Ehsan Hoque. 2023. Novel computational linguistic measures, dialogue system and the development of sophie: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing*, 14(1):223–235.

Summer Allen. 2022. Improving psychotherapy with ai: From the couch to the keyboard. *IEEE Pulse*, 13(5):2–8.

Edoardo Battegazzorre, Andrea Bottino, and Fabrizio Lamberti. 2021. Training medical communication skills with virtual patients: Literature review and directions for future research. In *Intelligent Technologies for Interactive Entertainment*, pages 207–226, Cham. Springer International Publishing.

Zhandos Zhumanov Batyrkhan Omarov, Sergazi Narynov. 2023. Artificial intelligence-enabled chatbots in mental health: A systematic review. *Computers, Materials & Continua*, 74(3):5105–5122.

Patrick Bowers, Kelley Graydon, Tracii Ryan, Jey Han Lau, and Dani Tomlin. 2024. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. *Australasian Journal of Educational Technology*, 40(3):39–57.

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Laurence Chaby, Amine Benamara, Maribel Pino, Elise Prigent, Brian Ravenet, Jean-Claude Martin, Hélène Vanderstichel, Raquel Becerril-Ortega, Anne-Sophie Rigaud, and Mohamed Chetouani. 2022. Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills: An overview of their use in psychiatric and geriatric care. *Frontiers in Virtual Reality*, 3.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *Preprint*, arXiv:2305.13614.

Doyanne Darnell, Patricia A Areán, Shannon Dorsey, David C Atkins, Michael J Tanana, Tad Hirsch, Sean D Mooney, Edwin D Boudreaux, and Katherine Anne Comtois. 2021. Harnessing innovative technologies to train nurses in suicide safety planning with hospitalized patients: Protocol for formative and pilot feasibility research. *JMIR Res Protoc*, 10(12):e33695.

Ismail Dergaa, Feten Fekih-Romdhane, Souheil Hallit, Alexandre Andrade Loch, Jordan M. Glenn, Mohamed Saifeddin Fessi, Mohamed Ben Aissa, Nizar Souissi, Noomen Guelmami, Sarya Swed, Abdelfatteh El Omri, Nicola Luigi Bragazzi, and Helmi Ben Saad. 2024. Chatgpt is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, 14.

Lucile Dupuy, Etienne De Sevin, Orlane Ballot, Hélène Cassoudesalle, Patrick Dehail, Bruno Aouizerate, Emmanuel Cuny, Jean-Arthur Micoulaud-Franchi, and Pierre Philip. 2019. A virtual patient to train semiology extraction and empathic communication skills for psychiatric interview. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 188–190, New York, NY, USA. Association for Computing Machinery.

Lucile Dupuy, Etienne de Sevin, Hélène Cassoudesalle, Orlane Ballot, P. Dehail, Bruno Aouizerate, Emmanuel Cuny, Jean-Arthur Micoulaud Franchi, and Pierre Philip. 2021. Guidelines for the design of a virtual patient for psychiatric interview training. *Journal on Multimodal User Interfaces*, 15.

Lucile Dupuy, Jean-Arthur Micoulaud-Franchi, Hélène Cassoudesalle, Orlane Ballot, Patrick Dehail, Bruno Aouizerate, Emmanuel Cuny, Etienne de Sevin, and Pierre Philip. 2020. Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. *Journal of Affective Disorders*, 263:1–8.

Zohar Elyoseph, Inbar Levkovitch, Yuval Haber, and Yossi Levi-Belz. 2024. Using genai to train mental health professionals in suicide risk assessment: Preliminary findings. *medRxiv*.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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*, 4(2):e19.

Alan Gilbert, Stephanie Carnell, Benjamin Lok, and Anna Miles. 2024. Using virtual patients to support empathy training in health care education: An exploratory study. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*, 19.

Kurtis Haut, Caleb Wohn, Benjamin Kane, Thomas Carroll, Cathrine Guigno, Varun Kumar, Ronald Epstein, Lenhart Schuber, and Ehsan Hoque. 2023. Validating a virtual human and automated feedback system for training doctor-patient communication skills. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Chester Holt-Quick and Jim Warren. 2021. Establishing a dialog agent policy using deep reinforcement learning in the psychotherapy domain. In *Proceedings of the 2021 Australasian Computer Science Week Multiconference*, ACSW '21, New York, NY, USA. Association for Computing Machinery.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*, 6(11):e12106.

Marisa Scholl Michael White Eric Fosler-Lussier William Schuler Prashant Serai Vishal Sunder Hannah Forrestal Lexi Mendella Mahsa Adib Camille Bratton Kevin Lee Kellen R. Maicher, Adam Stiff and Douglas R. Danforth. 2022. Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity. *Medical Teacher*, 45(3):279–285. PMID: 36346810.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *Preprint*, arXiv:2404.13066.

Michael Loizou, Sylvester Arnab, Petros Lameras, Thomas Hartley, Fernando Loizides, Praveen Kumar, and Dana Sumilo. 2024. Designing, implementing and testing an intervention of affective intelligent agents in nursing virtual reality teaching simulations—a qualitative study. *Frontiers in Digital Health*, 6.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *Preprint*, arXiv:2407.00870.

Rakesh K. Maurya. 2023a. A qualitative content analysis of chatgpt's client simulation role-play for practising counselling skills. *Counselling and Psychotherapy Research*, 24(2):614–630. Publisher Copyright: © 2023 British Association for Counselling and Psychotherapy.

Rakesh K. Maurya. 2023b. Using ai based chatbot chatgpt for practicing counseling skills through role-play. *Journal of Creativity in Mental Health*, 19(4):513–528.

Rakesh K. Maurya, Steven Montesinos, Mikhail Bogomaz, and Amanda C. DeDiego. 2024. Assessing the use of chatgpt as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*, n/a(n/a).

David Moher, Alessandro Liberati, James M Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of Internal Medicine*, 151:264–269.

Magalie Ochs, Daniel Mestre, Grégoire de Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. 2019. Training doctors' social skills to break bad news:

evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*, 13(1):41–51.

Daniela S.M. Pereira, Filipe Falcão, Lilian Costa, Brian S. Lunn, José Miguel Pêgo, and Patrício Costa. 2023. Here's to the future: Conversational agents in higher education- a scoping review. *International Journal of Educational Research*, 122:102233.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

Greg M Reger, Aaron M Norr, Michael A Gramlich, and Jennifer M Buchman. 2021. Virtual standardized patients for mental health education. *Current psychiatry reports*, 23(9):57.

Jinsil Hwaryoung Seo, Rohan Chaudhury, Ja-Hun Oh, Caleb Kicklighter, Tomas Arguello, Elizabeth Wells-Beede, and Cynthia Weston. 2023. Development of virtual reality sbirt skill training with conversational ai in nursing education. In *Artificial Intelligence in Education*, pages 701–707, Cham. Springer Nature Switzerland.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *J Med Internet Res*, 21(7):e12529.

Milen Todorov, Gergana Avramova-Todorova, Krasimira Dimitrova, and Valentin Irmov. 2022. Virtual assisted technologies as a helping tool for therapists in assessment of anxiety. outcomes of a pilot trial with chatbot assistance. In *Contemporary Methods in Bioinformatics and Biomedicine and Their Applications*, pages 60–66, Cham. Springer International Publishing.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a client-centered assessment of llm therapists by client simulation. *Preprint*, arXiv:2406.12266.

Heng Yao, Alexandre Gomes de Siqueira, Adriana Foster, Igor Galynker, and Benjamin Lok. 2020. Toward automated evaluation of empathetic responses in virtual human interaction systems for mental health scenarios. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *Preprint*, arXiv:2410.01553.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Annual Meeting of the Association for Computational Linguistics*.

A Publications related to Feature Categories

Technology	Percentage	Publications
AI-based	48.6%	(Dupuy et al., 2019, 2020; Bowers et al., 2024; Batyrkhan Omarov, 2023; Ab Razak et al., 2023; Loizou et al., 2024; Tanana et al., 2019; Holt-Quick and Warren, 2021; Darnell et al., 2021; Pereira et al., 2023; Allen, 2022; Ali et al., 2023; Zheng et al., 2024; Yao et al., 2020; Tu et al., 2024; Gilbert et al., 2024; Todorov et al., 2022; Reger et al., 2021)
LLMs (GPT)	29.7%	(Maurya, 2023a,b; Maurya et al., 2024; Dergaa et al., 2024; Qiu and Lan, 2024; Li et al., 2024; Chen et al., 2023; Yao et al., 2020; Louie et al., 2024; Wang et al., 2024; Elyoseph et al., 2024)
Rule-based	8.1%	(Campillos-Llanos et al., 2020; Dupuy et al., 2021; Haut et al., 2023)
Hybrid	2.7%	(Kellen R. Maicher and Danforth, 2022)
Mixed/VR	10.8%	(Seo et al., 2023; Chaby et al., 2022; Ochs et al., 2019; Battegazzorre et al., 2021)

Table 5: Technology categories and associated publications

Application	Percentage	Publications
Training	86.5%	(Maurya, 2023b; Dupuy et al., 2019; Kellen R. Maicher and Danforth, 2022; Bowers et al., 2024; Batyrkhan Omarov, 2023; Loizou et al., 2024; Maurya, 2023a; Dupuy et al., 2019; Kellen R. Maicher and Danforth, 2022; Bowers et al., 2024; Maurya et al., 2024; Dergaa et al., 2024; Campillos-Llanos et al., 2020; Loizou et al., 2024; Tanana et al., 2019; Seo et al., 2023; Chaby et al., 2022; Holt-Quick and Warren, 2021; Dupuy et al., 2020, 2021; Darnell et al., 2021; Allen, 2022; Qiu and Lan, 2024; Li et al., 2024; Chen et al., 2023; Yao et al., 2020; Ali et al., 2023; Louie et al., 2024; Yao et al., 2024; Wang et al., 2024; Ochs et al., 2019; Maurya, 2023b; Elyoseph et al., 2024; Gilbert et al., 2024; Haut et al., 2023; Reger et al., 2021)
Education	8.1%	(Ab Razak et al., 2023; Pereira et al., 2023; Battegazzorre et al., 2021)
Assessment	5.4%	(Batyrkhan Omarov, 2023; Todorov et al., 2022)

Table 6: Application areas and associated publications

Dataset	Percentage	Publications
No Dataset	18.9%	(Maurya, 2023a,b; Maurya et al., 2024; Dergaa et al., 2024; Bowers et al., 2024; Elyoseph et al., 2024; Todorov et al., 2022)
Internal	56.8%	(Dupuy et al., 2019; Kellen R. Maicher and Danforth, 2022; Campillos-Llanos et al., 2020; Loizou et al., 2024; Tanana et al., 2019; Seo et al., 2023; Chaby et al., 2022; Holt-Quick and Warren, 2021; Dupuy et al., 2020, 2021; Darnell et al., 2021; Allen, 2022; Li et al., 2024; Chen et al., 2023; Ali et al., 2023; Louie et al., 2024; Yao et al., 2020; Tu et al., 2024; Ochs et al., 2019; Gilbert et al., 2024; Haut et al., 2023)
Mixed	16.2%	(Batyrkhan Omarov, 2023; Ab Razak et al., 2023; Pereira et al., 2023; Wang et al., 2024; Battegazzorre et al., 2021; Reger et al., 2021)
Public	8.1%	(Qiu and Lan, 2024; Yao et al., 2024; Zheng et al., 2024)

Table 7: Dataset availability and associated publications

Evaluation Category	Percentage	Publications
Qualitative	8.1%	(Maurya, 2023a; Dergaa et al., 2024; Maurya, 2023b)
Quantitative	5.4%	(Kellen R. Maicher and Danforth, 2022; Todorov et al., 2022)
Mixed	86.5%	(Dupuy et al., 2019; Bowers et al., 2024; Batyrkhan Omarov, 2023; Maurya et al., 2024; Ab Razak et al., 2023; Campillos-Llanos et al., 2020; Loizou et al., 2024; Tanana et al., 2019; Seo et al., 2023; Chaby et al., 2022; Holt-Quick and Warren, 2021; Dupuy et al., 2020, 2021; Darnell et al., 2021; Pereira et al., 2023; Allen, 2022; Qiu and Lan, 2024; Li et al., 2024; Chen et al., 2023; Yao et al., 2024; Ali et al., 2023; Louie et al., 2024; Zheng et al., 2024; Yao et al., 2020; Wang et al., 2024; Tu et al., 2024; Ochs et al., 2019; Battegazzorre et al., 2021; Elyoseph et al., 2024; Gilbert et al., 2024; Haut et al., 2023; Reger et al., 2021)

Table 8: Evaluation categories and associated publications