Journal of Psychology and AI

ISSN: 2997-4100 (Online) Journal homepage: www.tandfonline.com/journals/tpai20

AI meets psychology: an exploratory study of large language models' competence in psychotherapy contexts

Kean Sian Tan, Matti Cervin, Patrick Leman, Kristopher Nielsen, Prashanth Vasantha Kumar & Oleg Medvedev

To cite this article: Kean Sian Tan, Matti Cervin, Patrick Leman, Kristopher Nielsen, Prashanth Vasantha Kumar & Oleg Medvedev (2025) AI meets psychology: an exploratory study of large language models' competence in psychotherapy contexts, Journal of Psychology and AI, 1:1, 2545258, DOI: 10.1080/29974100.2025.2545258

To link to this article: https://doi.org/10.1080/29974100.2025.2545258

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 28 Aug 2025.

Submit your article to this journal

Article views: 97

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tpai20

JOURNAL OF PSYCHOLOGY AND AI
2025, VOL. 1, NO. 1, 2545258
https://doi.org/10.1080/29974100.2025.2545258

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

AI meets psychology: an exploratory study of large language models' competence in psychotherapy contexts

Kean Sian Tan[a], Matti Cervin[b], Patrick Leman[a*], Kristopher Nielsen[c], Prashanth Vasantha Kumar[d] and Oleg Medvedev[a]

[a]School of Psychological and Social Sciences, University of Waikato, Hamilton, New Zealand; [b]The Child and Adolescent Psychiatric Clinic, Lund University, Lund, Sweden; [c]School of Psychology, Te Herenga Waka–Victoria University of Wellington, Wellington, New Zealand; [d]School of Healthcare Sciences, National University of Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

The increasing prevalence of mental health problems coupled with limited access to professional support has prompted exploration of technological solutions. Large Language Models (LLMs) represent a potential tool to address these challenges, yet their capabilities in psychotherapeutic contexts remain unclear. This study examined the competencies of current LLMs in psychotherapy-related tasks including alignment with evidence-informed clinical standards in case formulation, treatment planning, and implementation. Using an exploratory mixed-methods design, we presented three clinical cases (depression, anxiety, stress) and 12 therapy-related prompts to seven LLMs: ChatGPT-4o, ChatGPT-4, Claude 3.5 Sonnet, Claude 3 Opus, Meta Llama 3.1, Google Gemini 1.5 Pro, and Microsoft Co-pilot. Responses were evaluated by five experienced clinical psychologists using quantitative ratings and qualitative feedback. No single model consistently produced high-quality responses across all tasks, though different models showed distinct strengths. Models performed better in structured tasks such as determining session length and discussing goal-setting but struggled with integrative clinical reasoning and treatment implementation. Higher-rated responses demonstrated clinical humility, maintained therapeutic boundaries, and recognised therapy as collaborative. Current LLMs are more promising as supportive tools for clinicians than as therapeutic applications. This paper highlights key areas for development needed to enhance clinical reasoning abilities for effective mental health use.

ARTICLE HISTORY
Received 27 December 2024
Accepted 3 August 2025

KEYWORDS
Artificial intelligence; large language models; psychotherapy; therapeutic competence; mental health technology

The prevalence of mental health disorders such as depression, anxiety, and stress-related disorders represents a significant global health challenge, affecting populations across all cultural and socioeconomic contexts. According to the (WHO, 2022), approximately 970 million people worldwide were living with a mental disorder in 2019. This figure represents nearly one in eight people globally, highlighting the pervasive nature of mental health issues. Depression and anxiety-related disorders, two of the most common types of mental disorders, affected an estimated 280 million and 301 million people respectively in 2019, with many individuals experiencing both conditions simultaneously (WHO, 2020). A cross-national analysis of population studies estimated that around half of the population is expected to develop at least one mental health disorder by the age of 75 (McGrath et al., 2023). Even more striking, longitudinal research following individuals from birth has found substantially higher rates, with 86% of people developing at least one mental disorder by age 45 (Caspi et al., 2020). These findings suggest that the lifetime risk of developing mental illness is substantially higher than previously recognised, pointing to a pervasive public health challenge with far-reaching implications for individuals and society as a whole.

The widespread prevalence of mental health disorders translates into substantial economic consequences, affecting individuals, healthcare systems, and national economies. Mental health disorders now rank among the top ten causes of global health burden, with depression projected to become the leading cause in terms of disability-adjusted life years in the future (Alize et al., 2022). A report by the Lancet Commission on global mental health and sustainable development estimated that mental disorders could

CONTACT Kean Sian Tan xiangtks@gmail.com
*Present address: Faculty of Medicine and Health Sciences, University of Buckingham, UK
Supplemental data for this article can be accessed online at https://doi.org/10.1080/29974100.2025.2545258

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

cost the global economy up to $16 trillion between 2010 and 2030 (Patel et al., 2018). This cost is attributed to direct healthcare expenses and indirect costs such as lost productivity and disability (Patel et al., 2018).

Despite the growing prevalence of mental health issues, significant barriers persist in accessing mental health care worldwide. A primary obstacle is the widespread shortage of trained mental health professionals. The Mental Health Atlas 2020 by World Health Organization reports that globally there are fewer than three mental health workers per 100,000 people in low and middle-income countries, and fewer than 11 per 100,000 in upper-middle-income countries (WHO, 2020). Even in high-income countries, the ratio only reaches 55.3 mental health professionals per 100,000 population, highlighting the global scarcity of mental health resources. This shortage is particularly acute in rural and remote areas where geographical barriers create additional challenges in accessing specialist care (Kulshrestha & Shahid, 2022). Another common barrier is long waitlists, a direct consequence of resource shortage, which have significant implications for treatment efficacy. Every-Palmer et al. (2022) found that longer waiting times for mental health treatment were associated with poorer clinical outcomes and increased risk of treatment dropout. The delay in receiving care exacerbates symptoms and, in severe cases, results in crises that could have been prevented with timely intervention (Every-Palmer et al., 2023).

Moreover, the need for personalised and culturally appropriate care presents another significant challenge in mental health service delivery worldwide. Research consistently shows that positive recovery outcomes require interventions tailored to individual cultural backgrounds and worldviews (Çam & Uğuryol, 2019; Griner & Smith, 2006; Hall et al., 2016; Scoles, 2022). However, healthcare systems globally often lack the resources and flexibility to provide such personalised care at scale. For example, in Aotearoa New Zealand, Pasifika communities struggle to receive mental health services congruent with their own cultural worldviews (Kapeli et al., 2020). Furthermore, stigma continues to be a pervasive barrier to seeking mental health care (Goetter et al., 2020). Kulshrestha and Shahid (2022) posit that stigma prevents individuals from seeking help and impacts the quality of care received when they do seek help. This stigma can be particularly pronounced in certain cultural contexts or marginalised groups, creating additional barriers for specific populations (Kapeli et al., 2020).

In light of the significant challenges in mental health care, the integration of artificial intelligence (AI), particularly Large Language Models (LLMs), represents a promising advancement in addressing the growing demand for mental health services (Banerjee et al., 2024; Lawrence et al., 2024). LLMs, an advanced form of generative AI, are trained on vast amounts of textual data, enabling them to generate human-like text across various domains based on statistical pattern recognition, which to human users may appear as comprehension (Naveed et al., 2024). Prominent examples of contemporary LLMs include OpenAI's GPT (Generative Pre-trained Transformer) series, Google's Gemini, Meta's Llama and Anthropic's Claude. These models offer several advantages in addressing mental health care barriers: they transcend time and geographical limitations through internet accessibility (Guo et al., 2024), support multiple languages and cultural expressions (Open AI et al., 2024c), and reduce stigma in self-disclosure through their nonjudgemental and supportive interaction style (Ma et al., 2023).

Although LLMs are fundamentally generative models that produce text in response to prompts, their potential value in mental health contexts lies not in the act of generation itself, but in the quality, coherence, and psychological relevance of the responses they generate (Stade et al., 2024). The ability to produce meaningful output depends on how effectively the model has internalised language patterns associated with empathic communication, supportive dialogue, and contextual reasoning. Such capacities are shaped by the nature and scope of the data used during pretraining and refinement (Dam et al., 2024; Ke et al., 2024).

Figure 1 illustrates the basic architecture of a large language model (LLM). The process begins with the training data, a vast collection of text from various sources, including books, scientific papers, and websites related to both physical and mental health research and practice (Dam et al., 2024; Ke et al., 2024; Stade et al., 2024). This data undergoes tokenisation, where the text is segmented into smaller units called tokens (Naveed et al., 2024). For example, the phrase "I am feeling overwhelmed" might be tokenised into discrete units such as "I", "am", "feel", "ing", "over", and "whelmed", depending on the model's tokeniser (Dam et al., 2024; Patil & Gudivada, 2024). Tokens may represent entire words, subwords, or individual characters. These tokens are then passed through an embedding layer, where each token is transformed into a dense numerical vector (Naveed et al., 2024). This transformation enables the model to process language in a structured, mathematical format, which supports

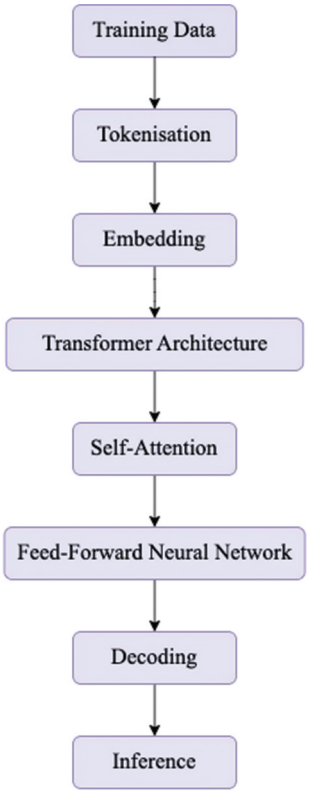

Figure 1. Simplified overview of large language model (LLM) architecture.

downstream tasks such as attention and prediction. However, tokenisation and embedding may not correspond to how humans understand and apply language, as the underlying mechanisms differ fundamentally from human cognitive processes. In LLMs, the embedding process enables the model to capture statistical relationships between tokens by mapping them into a high-dimensional space, allowing it to represent contextual patterns in the input.

At the heart of the LLM is the Transformer Architecture, a neural network framework designed to process sequential data (Naveed et al., 2024). A key feature of this architecture is the self-attention mechanism, which allows the model to dynamically weigh the importance of different tokens within the sequence (Patil & Gudivada, 2024). This enables the model to capture local and long-range dependencies in the text, creating more contextually informed representations of the input. The multi-head self-attention mechanism simultaneously processes various aspects of the input, further enhancing the model's ability to extract complex patterns (Patil & Gudivada, 2024). The contextual information derived from the self-attention layers is then passed through a feed-forward neural network (Naveed et al., 2024). This network refines the information, enabling the model to form more abstract and generalised representations of the input data. Importantly, this process occurs across multiple layers within the transformer, allowing for deeper understanding at each successive level of processing.

Once the information has passed through these layers, it reaches the decoding stage. During decoding, the model converts its internal representations into a sequence of output tokens, which are then translated back into human-readable text (Naveed et al., 2024). This process involves generating the most probable sequence of words based on the input to ensure coherence and fluency. The final step is inference, where the model produces its output (Naveed et al., 2024). It is important to note that the model's parameters are adjusted during the training phase and no further adjustment occurs during inference, the model generates responses solely in accordance with the patterns it has internalised during training (Patil & Gudivada, 2024).

General-purpose LLMs have demonstrated baseline capabilities in mental health contexts. These models can recognise emotional tone, answer mental health questions, and offer supportive language to users (Banerjee et al., 2024; Carlbring et al., 2023; Chen et al., 2023; Lee et al., 2024). For example, ChatGPT has

been used to screen for suicide risk based on patient narratives, showing sensitivity to clinically relevant cues (Elyoseph et al., 2023).

To enhance LLMs' performance in mental health applications, researchers often employ domain-specific fine-tuning. A notable example is Med-PaLM 2, a medically fine-tuned LLM capable of assessing psychiatric disorders from clinical transcripts and explaining its decisions (Galatzer-Levy et al., 2023). The model achieved classification accuracy ranging from 0.80 to 0.84 in diagnosing depression, PTSD, and high-comorbidity disorders, performing comparably to trained human raters across multiple assessments. Although this range does not represent a breakthrough in AI performance, it is broadly consistent with benchmark results in mental health prediction (e.g. AUCs of 0.80–0.85 are commonly regarded as strong discrimination; Abd-Alrazaq et al., 2022; Rony et al., 2025). Reaching this level of accuracy suggests that LLMs can recognise clinically relevant language patterns, including subtle cues linked to psychiatric symptoms and diagnostic reasoning. However, research consistently shows that further improvements in accuracy and reliability often depend on greater human involvement, such as fine-tuning with domain-specific text or incorporating expert-in-the-loop feedback (Al-Turki et al., 2024; Yu & McGuinness, 2024).

Furthermore, LLMs have shown promise in certain narrow tasks, sometimes even matching or exceeding human clinicians within those constrained domains. Kim et al. (2024) found that LLMs demonstrated superior diagnostic accuracy for Obsessive-Compulsive Disorder (OCD) compared to medical and mental health professionals. In text-based emotion recognition, Elyoseph et al. (2023) demonstrated that ChatGPT exhibited a significantly higher ability to identify others' emotions than the general population, suggesting potential applications for enhancing therapists' emotional vocabulary and recognition skills during assessments. Similarly, Sufyan et al. (2024) found that some LLMs, particularly ChatGPT-4 and Bing, outperformed psychologists on text-based social intelligence tasks, indicating potential to support counselling and psychotherapy.

Despite these promising applications, the use of LLMs in mental health care also presents significant challenges. The "black box" nature of these models raises questions about reliability and consistency in offering high-quality and safe responses in clinical settings (Bill & Eriksson, 2023; Chiu et al., 2024; Wang et al., 2024). Another essential issue in applying LLMs to mental health care is privacy and data security, particularly given the sensitive nature of information exchanged between clients and therapists (Dam et al., 2024; Fiske et al., 2019). As research in this field progresses, we are seeing increasingly sophisticated applications of LLMs in mental health care, such as predicting one's mental health state via online text (Xu et al., 2023). However, there remains a need for rigorous evaluation of these models and the safety issues they give rise to as they continue to evolve.

Therefore, the present study aimed to evaluate the leading LLMs' capabilities to engage in psychotherapy-related tasks to answer the following research question: How effectively can current leading LLMs perform in textual simulations of therapeutic scenarios, and how do their responses align with evidence-informed clinical practice standards? Our research objectives were threefold: (1) to assess the therapy-related competency of seven prominent LLMs across a range of therapeutic scenarios, (2) to identify patterns in their strengths and weaknesses in therapeutic contexts, and (3) to evaluate the ethical implications and potential risks associated with their use in personal mental health care. The significance of this research lies in its potential to inform the development and application of LLMs in mental health care. As healthcare systems worldwide face increasing demand for mental health services, the possibility of engaging AI to expand access to and improve mental health care delivery is compelling. However, the integration of such powerful technologies into the sensitive domain of mental health must be approached with caution and be informed by empirical evidence.

Method

Objects of study

We selected seven LLMs based on their prominence and frequent usage in the current LLM landscape, including OpenAI ChatGPT-4 (Open AI, 2024b), OpenAI ChatGPT-4o (Anthropic, 2024a), Anthropic's Claude 3.5 Sonnet (Anthropic, 2024b), Anthropic's Claude 3 Opus (Anthropic, 2024a), Meta Llama AI 3.1 (Meta, 2024), Google Gemini 1.5 Pro (Google, 2024), and Microsoft Copilot (Microsoft, 2024). All LLMs

were the most up-to-date versions available at the time of data collection in July 2024. Some models from the same providers (e.g. ChatGPT-4 and ChatGPT-4o, Claude 3.5 Sonnet and 3 Opus) were included to compare different versions or capabilities within the same family of models for suitability in a therapy context.

Procedure

This study employed an exploratory mixed-methods design to investigate the potential effectiveness of LLMs in providing therapy-related responses. We presented three clinical cases to seven different LLMs, along with a set of 12 therapy-related prompts for each case. The LLM responses were then evaluated by experienced clinical psychologists using both quantitative ratings and qualitative feedback.

The three clinical cases were obtained from textbooks and journals, featuring common mental health problems (see Supplementary Material 1 for full case descriptions). Case 1 described a woman with depressive symptoms (Gilbert, 2016). Case 2 featured a man with social anxiety (Baumgardner & Benoit Allen, 2024), and case 3 described work and marriage stress resulting in burnout (Pines, 2002).

In addition, we developed a set of 12 therapy-related prompts designed to assess various aspects of therapeutic interaction (see Supplementary Material 2 for the full list of prompts). Half of the prompts were developed according to the framework of evidence-informed standards of clinical practice, including the process of case formulation, treatment planning, and treatment implementation (Christon et al., 2015; Page et al., 2008). Other prompts assessed knowledge of optimal treatment length and goal-setting (Lindhiem et al., 2016). Lastly, the LLMs were requested to role-play the first therapy session with a client and offer recommendations for coping strategies tailored to the client's presented problems. The goal of these prompts was to understand how well LLMs perform in basic therapy competency and their ability to offer evidence-based suggestions tailored to common mental health challenges and symptoms.

The data collection process involved a systematic presentation of cases and prompts to the selected LLMs. We used a zero-shot approach without fine-tuning any models as we aimed to test the models' initial capacity at the current level of development. Each of the seven LLMs was presented with the three clinical cases, accompanied by a set of 12 therapy-related prompts for each case. The process began with the prompt, "I am a psychologist, and I have a client described below. Please help me with case formulation for this client." This initial prompt was followed by the remaining 11 prompts, maintaining consistency across all models and cases (Supplementary Material 2).

After receiving and saving responses from all LLMs in an Excel file, we compiled and transferred these responses to the Qualtrics platform. To mitigate potential biases and ensure a fair evaluation process, the LLM responses were assigned a randomised numerical label and presented to the evaluators in a randomised order. This randomisation served two purposes. It prevented order effects that might influence evaluators' judgements and blinded evaluators to which specific model generated each response.

Once the AI data collection was completed and set up on Qualtrics, we sent emails and information sheets to known contacts of registered clinical psychologists, inviting them to be evaluators. Upon agreement to evaluate AI responses, evaluators received an evaluation link, copies of the clinical cases, and prompts to facilitate their evaluation. To mitigate fatigue, evaluators were encouraged to complete the evaluation over multiple sessions. Once they had completed the blinded evaluation, we officially invited them for potential collaboration with our research team so we could include their additional insights when communicating the overarching patterns in LLM responses and our research findings. The study was approved by the lead author's institutional ethics committee.

Measures

The evaluation process was designed to capture both quantitative assessments and qualitative insights from the certified clinical psychologist evaluators. For each LLM response, evaluators were presented with two key components. First, they were asked to provide a quantitative rating of each AI response while considering the relevant clinical case. The rating used a 5-point Likert scale, where 1 = *Unacceptable*, 2 = *Poor or Inadequate*, 3 = *Acceptable*, 4 = *Appropriate*, and 5 = *Very Appropriate*. In addition to the numerical rating, evaluators were provided with an optional qualitative comment box for each response. It allowed

evaluators to offer detailed feedback, explanations for their ratings, or additional observations about the strengths and weaknesses of specific LLM responses. By combining quantitative ratings with the option for qualitative feedback, our measurement approach aimed to comprehensively evaluate the LLMs' performance in simulated therapeutic contexts.

Data analyses

Five male clinical psychologists completed the evaluations, with a mean age of 36.4 years and an average of 7.3 years of clinical experience. Three were registered in New Zealand, one in Malaysia, and one in Sweden. All held at least a master's degree in psychology with clinical psychology qualifications, and two with additional PhDs in psychology. Data analyses were conducted using IBM SPSS version 29.0.2.0. Descriptive statistics showed that skewness values ranged from −0.07 to 0.92 and kurtosis values ranged from −0.85 to 0.35 across the 12 questions (see Supplementary Table S1), indicating approximately normal distribution as values fell within the acceptable range of ±2 for both skewness and kurtosis (George & Mallery, 2016). Inter-rater reliability was assessed using a two-way random effects model with absolute agreement (Koo & Li, 2016). The overall inter-rater reliability was poor, ICC(2,1) = .023, 95% CI [−.003, .054]. ICC values across cases and models ranged from −.012 to .070, indicating minimal agreement among evaluators regarding the quality of AI-generated responses (see Supplementary Table S2).

We then conducted nonparametric Kruskal-Wallis tests to compare performance between models across all 12 questions, as this test is specifically designed for comparing ordinal data across multiple independent groups without assuming normal distribution. This approach better aligns with the measurement properties of our data and provides a more conservative test of differences between models. To control for experiment-wise error across multiple comparisons, we applied a Bonferroni correction. With 12 separate tests (one for each prompt), we adjusted our significance threshold from $\alpha = .05$ to $\alpha = .004$ (.05/12), thereby reducing the risk of Type I error.

Results

Results are presented sequentially across the 12 therapy-related prompts to evaluate LLMs' competencies. Kruskal-Wallis test revealed that differences between models across all questions did not reach level of statistical significance (see Supplementary Table S3). Although no statistically significant differences were observed, these findings should be interpreted cautiously. A non-significant result does not imply that the models are equivalent, but rather that the available data did not provide sufficient evidence to detect a difference in performance. In contrast, there was a significant effect of evaluator across all 12 questions ($p < .001$), suggesting consistent differences in how evaluators rated LLM responses. Table 1 presents the descriptive statistics (mean, standard deviation) for all questions across the seven models. Although no models consistently achieved ratings above the acceptable level (3.0), several models approached this threshold in specific therapeutic tasks. In addition to means and standard deviations, the medians and interquartile ranges for each question across all models are reported in Supplementary Table S4.

Question 1 requested models to help with case formulation based on the provided clinical case. No model achieved mean ratings at the acceptable level (3.0), with Opus ($M = 2.93$; $SD = 1.16$) and ChatGPT-4 ($M = 2.93$; $SD = 1.10$) approaching the acceptable threshold. Qualitative evaluation: Although all models fell below the acceptable rating level, evaluators noted some strengths in specific models. Evaluators valued Opus's broad consideration of biopsychosocial factors and suggested that if used appropriately, this may prompt psychologists to explore multiple perspectives. ChatGPT-4 was praised for its epistemic humility, explicitly describing its formulation as a "first step" and acknowledging the need for ongoing assessment. However, evaluators identified two major limitations across all models. First, models showed a lack of specificity in linking presenting problems to psychological mechanisms (e.g. listing biopsychosocial factors without demonstrating how these factors interact to explain presenting problems, resulting in descriptive rather than explanatory formulations). Second, models appeared to make questionable inferences without sufficient explanation (e.g. assuming trauma or attachment issues without supporting evidence).

Question 2 requested models to create treatment planning and measurement. Only ChatGPT-4 ($M = 3.13$; $SD = 1.13$) generated above-acceptable level responses, while other models performed below

Table 1. Mean rating and SD of appropriateness of AI's response, across questions 1–12 and overall.

Question	3 Opus	3.5 Sonnet	ChatGPT-4o	ChatGPT-4	Co-pilot	LlaMa 3.1	Gemini 1.5 Pro
Q1 (Case Formulation)	2.93 (1.16)	2.80 (0.86)	2.80 (0.94)	2.93 (1.10)	2.60 (0.99)	2.80 (1.08)	2.67 (0.90)
Q2 (Treatment Planning)	2.73 (1.34)	2.73 (0.96)	2.53 (0.92)	3.13 (1.13)	2.87 (1.25)	2.80 (1.21)	2.73 (1.16)
Q3 (Treatment Implementation)	2.60 (1.06)	2.87 (1.06)	2.87 (0.99)	2.80 (0.94)	2.60 (1.06)	3.33 (0.98)	2.60 (0.99)
Q4 (Session Planning)	2.73 (1.16)	2.67 (0.90)	2.53 (0.92)	2.73 (1.03)	2.60 (0.83)	2.67 (1.05)	2.73 (1.16)
Q5 (Session Duration)	3.27 (0.96)	3.13 (1.06)	2.93 (0.70)	3.13 (1.13)	3.07 (0.96)	3.20 (0.78)	3.00 (1.00)
Q6 (First Therapy Session Simulation)	2.47 (1.13)	2.13 (0.92)	2.33 (0.90)	2.60 (1.35)	2.40 (0.91)	2.47 (1.06)	2.40 (1.35)
Q7 (Recommendation for First Presenting Problem)	2.60 (1.18)	2.80 (0.94)	3.00 (1.13)	3.07 (0.88)	3.13 (1.13)	2.93 (0.80)	2.93 (0.80)
Q8 (Practical Technique Application)	3.07 (1.10)	2.47 (1.13)	3.07 (0.96)	2.87 (1.06)	2.93 (0.88)	3.13 (0.92)	2.73 (1.03)
Q9 (Recommendation for Second Presenting Problem)	2.73 (1.28)	2.53 (0.99)	2.60 (1.06)	2.67 (1.05)	2.73 (1.03)	2.73 (1.22)	2.87 (0.83)
Q10 (Recommendation for Third Presenting Problem)	2.60 (0.99)	2.67 (1.11)	2.67 (1.05)	2.73 (0.88)	2.80 (1.15)	2.80 (1.01)	2.67 (1.11)
Q11 (Discuss View on Goal-setting)	3.07 (1.03)	3.07 (1.10)	3.07 (0.59)	2.87 (0.92)	3.47 (0.99)	2.93 (0.80)	2.87 (0.74)
Q12 (Goal-setting with Client)	2.53 (0.92)	3.00 (1.13)	2.67 (0.90)	3.07 (1.10)	2.73 (0.80)	2.73 (1.16)	3.07 (1.10)
Overall *M*(*SD*)	2.78 (1.11)	2.74 (1.02)	2.76 (0.94)	2.88 (1.05)	2.83 (1.00)	2.88 (1.02)	2.78 (1.02)

Appropriateness of AI's response is rated on a scale from 1 (*Unacceptable*), 2 (*Poor or Inadequate*), 3 (*Acceptable*), 4 (*Appropriate*), to 5 (*Very Appropriate*).

acceptable level. Qualitative evaluation: Regarding strengths, evaluator responses suggested that ChatGPT-4's higher rating stemmed from its structured approach in suggesting treatment phases with clear goals while maintaining acknowledgement of ongoing assessment needs. Gemini was noted for providing more selective and targeted interventions with specific rationales, offering a more practical approach than other models' tendency to generate exhaustive lists. In terms of weaknesses, evaluators identified several issues across all models. First, models suggested an impractically large number of assessment measures (e.g. six questionnaires at each session). Second, the sequencing of interventions often lacked an appropriate clinical logic, sometimes giving rise to safety issues. For example, suggesting trauma processing before establishing required emotional regulation skills. Third, treatment timelines were frequently described as unrealistically brief and optimistic. Most fundamentally, evaluators noted that treatment plans appeared to follow a "scatter gun approach" rather than demonstrating careful consideration of individual client circumstances, resources, and readiness for specific interventions.

Question 3 requested models to implement suggested treatment options. LlaMa achieved the highest mean rating ($M = 3.33$; $SD = 0.98$), followed by ChatGPT-4o ($M = 2.87$; $SD = 1.06$) and Sonnet ($M = 2.87$; $SD = .99$). Qualitative evaluation: LlaMa's higher rating was attributed to its "well-structured overview" with a comprehensive breakdown of implementation steps, providing clear examples and contextualised rationales. In contrast, lower-rated models demonstrated several limitations. They tended to present generic, textbook-style responses that lacked consideration of practical application in context. For example, when describing EMDR implementation, these models listed standard protocol steps without addressing timing, client readiness, or integration with other therapeutic elements.

Question 4 requested models to recommend the number of therapeutic sessions required and specific goals for each session. No models achieved acceptable ratings level. Qualitative evaluation: Evaluators identified several fundamental issues across all models. Models proposed unrealistic timelines. For example, some LLMs made suggestions such as "25–30 sessions as standard" without clear justification. The responses also demonstrated what one evaluator termed "foreclosure on trauma," prematurely structuring interventions without sufficient assessment data. Moreover, models frequently presented overly detailed session-by-session goals that failed to account for individual client progress or therapeutic process. As one evaluator noted, "entire treatment packages are not steps – over general to the point of being unhelpful." This criticism reflected a broader concern that models were "drawing on a generic CBT style plan rather

than developing a formulation-driven approach," suggesting an overreliance on standardised protocols rather than individualised therapeutic planning.

Question 5 requested models to recommend session length for therapy. All models' responses achieved acceptable ratings except for ChatGPT-4o ($M = 2.93$; $SD = 0.70$). Opus achieved the highest mean rating ($M = 3.27$; $SD = 0.96$). Qualitative evaluation: Responses suggested that models achieving acceptable ratings demonstrated appropriate clinical judgement by recommending standard 60-minute sessions with allowances for longer initial assessment sessions. However, evaluators identified important concerns about recommendations for variable session lengths at different treatment phases. Drawing from clinical experience, evaluators emphasised that "maintaining therapeutic boundaries requires consistent 60-minute timeframes" rather than fluctuating durations. Some LLMs' suggestions for longer sessions towards therapy termination were criticised as contradicting typical practice, where clients progress towards greater independence and require less intensive support.

Question 6 requested models to simulate a first therapy session dialogue as the clinical psychologist. No models achieved acceptable ratings level, with ChatGPT-4 ($M = 2.60$; $SD = 1.35$) achieving the highest mean rating. Qualitative evaluation: Evaluators identified several limitations across all models' simulated sessions. Models consistently neglected essential first-session components such as consent discussions and risk assessment, raising significant safety and legal concerns. For example, one evaluator noted, "no simulations so far have discussed confidentiality and consent." Evaluators also highlighted concerns about models using unrealistic "artificial-sounding client responses." As one evaluator noted, "very few clients would be this open initially," pointing to models' failure to capture authentic therapeutic dialogue.

Question 7 requested models to teach practical techniques for clients to address their primary presenting problem. Three models achieved acceptable ratings: Copilot ($M = 3.13$; $SD = 1.13$), ChatGPT-4 ($M = 3.07$; $SD = 0.88$), and ChatGPT-4o ($M = 3.00$; $SD = 1.13$). Qualitative evaluation: For the models achieving acceptable ratings, evaluators noted their ability to offer helpful techniques such as deep breathing, progressive muscle relaxation, three-minute breathing space, mindfulness-based stress reduction (MBSR), self-compassion, and urge surfing. However, evaluators still identified areas for improvement across all models. While models offered good descriptions of the techniques, they failed to demonstrate how to guide clients in practising them. Responses were often "too wordy" and risked overwhelming clients with excessive information rather than offering clear, actionable steps. Evaluators emphasised that techniques should be "co-constructed" with clients rather than simply "taught."

Question 8 requested models to specify the temporal duration and practice frequency of their proposed therapeutic exercises. Three models achieved acceptable ratings: LlaMa ($M = 3.13$; $SD = 0.92$), Opus ($M = 3.07$; $SD = 1.10$) and ChatGPT-4o ($M = 3.07$; $SD = 0.96$). Qualitative evaluation: The higher-rated models were praised for providing detailed yet concise explanations of the techniques. Evaluators particularly appreciated models that included examples of how and when clients can practise these techniques. However, evaluators found that models provided overly detailed and prescriptive timing recommendations that might overwhelm or confuse clients. Evaluators suggested that simpler, more flexible timing recommendations would be more beneficial for the client. For instance, techniques such as STOPP (Stop, Take a breath, Observe, Pull back, and Proceed) and urge surfing were described with unnecessarily long durations, while these are typically brief techniques that clients can easily incorporate into their day. One evaluator also highlighted that models should provide evidence to support their specific timing recommendations and better explain the rationale behind suggested durations.

Question 9 requested models to recommend solutions for addressing the second presenting problem in each case scenario. No models achieved acceptable ratings, with Gemini ($M = 2.87$; $SD = .83$) achieving the highest mean rating. Qualitative evaluation: Evaluators noted that all the recommendations were somewhat generic and brief. Another common criticism across models was their tendency to provide numerous suggestions without clear prioritisation or integration with the case formulation. Notably, evaluators observed that models often ignored clients' existing strengths and resources. For instance, in cases where clients already demonstrated good social support, models still recommended basic social skills training.

Question 10 requested models to recommend solutions for addressing the third presenting problem in each case scenario. No models achieved acceptable ratings, with means ranging from 2.60 to 2.80. Qualitative evaluation: Evaluators consistently identified issues with models' tendency to artificially separate problems and provide disconnected solutions rather than taking an integrated approach. The responses

were criticised for being overly general and reactive rather than formulation-driven. Several evaluators noted that models sometimes contradicted their earlier suggestions or failed to maintain continuity with previously discussed interventions. There was particular concern about models recommending interventions like "positive affirmations" without strong evidence bases. Some evaluators also observed that by this point in the therapeutic planning, models appeared to be "throwing everything at the wall," suggesting increasingly generic solutions without a clear connection to the client's specific situation or previously identified treatment priorities.

Question 11 requested models to discuss their views on implementing goal-setting in therapy. Responses from Copilot ($M = 3.47$; $SD = 0.99$) were rated as highest, followed by Opus ($M = 3.07$; $SD = 1.13$), Sonnet ($M = 3.07$; $SD = 1.10$), and ChatGPT-4 ($M = 3.07$; $SD = .59$). Qualitative evaluation: The higher-rated models appeared to be valued for acknowledging goal-setting as a collaborative process rather than a purely prescriptive one. Evaluators appreciated when models emphasised the importance of client involvement in goal development and the need for flexibility in adjusting goals throughout therapy. However, several concerns were identified. Models often presented overly structured approaches that might not accommodate the natural flow of therapy or client readiness for change. As one evaluator noted, "in a clinical setting this would be a back and forth between the client and the therapist." Models also tended to separate shorter and longer-term goals artificially without clear consideration of how these goals might evolve through the therapeutic process.

Question 12 requested models to facilitate realistic and specific goals for each client to enhance therapy effectiveness. Responses from Sonnet ($M = 3.00$; $SD = 1.13$), ChatGPT-4 ($M = 3.07$; $SD = 1.10$), and Gemini ($M = 3.07$; $SD = 1.10$) achieved acceptable ratings. Qualitative evaluation: Responses suggested that Gemini, Sonnet and ChatGPT-4's higher ratings were due to their balanced approach in suggesting specific, measurable goals while maintaining flexibility. Evaluators valued responses that emphasised the collaborative nature of goal-setting and acknowledged the need to adjust goals based on client feedback and progress. However, common criticisms included setting unrealistic timelines, suggesting goals that did not align with clients' current capabilities, and failing to sequence goals based on client readiness and resources. As one evaluator noted, "the effort to transpose the goals onto the SMART framework (Specific, Measurable, Achievable, Relevant, and Time-bound) is appreciated but perhaps not necessary for every single goal." Several evaluators also observed that goals were often presented in a prescriptive manner rather than being co-constructed with the client.

Overall, the quantitative analysis revealed that models' performance varied across the 12 therapeutic tasks evaluated. While no significant differences were found between models across all questions, some tasks showed notably better performance across models, particularly structured tasks like determining session length (Q5) and discussing views on goal-setting (Q11). In contrast, models consistently struggled with practical and integrative tasks including: case formulation (Q1), treatment planning (Q2), treatment implementation (Q3), first session simulation (Q6), solution recommendations for presenting problems (Q7, Q8, Q9, Q10), and goal-setting with clients (Q12). Few or no models reached acceptable ratings for these tasks.

The qualitative evaluation by clinical psychologists identified several recurring themes across all models. First, models frequently demonstrated an overly prescriptive approach rather than case formulation driven, failing to integrate with earlier formulations and not accounting for individual client needs, capabilities, and readiness. Second, models made assumptions about clients' psychological states before collecting additional information and provided recommendations without clear rationales. Third, models showed limited ability to adapt their responses based on client context, often defaulting to standardised, textbook-style approaches. Fourth, responses often lacked practical implementation guidance, providing theoretical knowledge without clear application steps or guidance for clients. Finally, responses sometimes contained incorrect information or suggested interventions without strong evidence bases, raising concerns about client safety.

Discussion

This study set out to evaluate how effectively current leading LLMs could perform psychotherapy-related tasks across common mental health presentations. Our most striking finding was that while these AI models showed promising capabilities in structured tasks like determining therapy session length and discussing

goal-setting approaches, they consistently struggled with clinical reasoning and the dynamic, human aspects of therapy that practising clinicians consider essential (Norcross & Lambert, 2018). For example, models were able to articulate appropriate views about goal-setting as a collaborative and flexible process, but struggled to actually facilitate realistic goals with clients – much like having textbook knowledge without practical application.

Looking at overall performance patterns, our analyses revealed no statistically significant differences between models across all questions, suggesting comparable capabilities among all models in therapeutic tasks. Despite this overall similarity, different models achieved mean ratings at the "acceptable" level (3 out of 5) in various tasks: ChatGPT-4 in four tasks, Opus, Sonnet, ChatGPT-4o, Copilot and LlaMa each in three tasks, and Gemini in two tasks. These ratings were distributed across different types of therapeutic tasks. For instance, some models performed better at teaching techniques while others demonstrated strengths in treatment planning. However, it is noteworthy that no model achieved a mean rating at the "appropriate" level (4 out of 5) for any task. The significant effect of evaluators ($p < .001$) across all questions suggested that evaluators were consistent in their rating patterns across models but they differed significantly in their assessment criteria. These findings indicated that although current LLMs show promise in generating quality responses, they do not yet consistently meet professional standards, regardless of which model is used.

Across all models, evaluators identified two recurring weaknesses. First, models demonstrated inadequate clinical reasoning in most tasks. Models often generated generic responses and were unable to offer treatment plans that were coherent and integrated with original case formulations. Case formulation is a crucial and ongoing process in clinical practice. It involves gathering client information to create explanatory hypotheses for the client's presenting symptoms (Page et al., 2008; Sim et al., 2005). Without case formulation, therapy lacks a framework for understanding client difficulties and risks becoming fragmented or unfocused (Persons, 2006). A good case formulation helps therapists organise clinical information and develop targeted interventions based on theoretical understanding of the client's difficulties. Our findings showed that no models created case formulations that met acceptable levels, and they struggled to anchor them as rationales for suggested treatment plans and technique recommendations. Second, models fell short in tasks requiring practical therapeutic implementation and dynamic interaction. When asked to conduct a first therapy session, models often jumped straight into solutions without first establishing rapport or addressing essential elements such as consent and confidentiality – steps that are commonly expected in clinical practice. Similarly, while models could list and describe therapeutic techniques, they struggled to explain them in accessible ways that clients could actually implement in their daily lives. These limitations were particularly evident in teaching techniques and addressing complex client presenting problems, where most models failed to achieve acceptable responses. Models tended to provide numerous generic interventions without clear prioritisation or integration with case formulations. It suggests that although LLMs possess a certain level of clinical knowledge, they struggle with the practical, responsive aspects of therapeutic work.

Qualitative analysis of those responses that achieved acceptable ratings tended to reveal four distinguishing characteristics that aligned with fundamental principles of clinical practice. First, acceptable responses demonstrated *appropriate clinical humility* by acknowledging the preliminary nature of formulations and the need for ongoing assessment. Second, acceptable responses *maintained therapeutic boundaries* by recommending consistent 60-minute session lengths rather than variable durations for different sessions. Third, acceptable responses *avoided making assumptions beyond available information*, particularly around diagnoses or client history not provided in case materials. Fourth, appropriate responses *recognised therapy as a collaborative process* requiring client engagement rather than providing purely prescriptive interventions. These characteristics observed in higher-rated responses suggest burgeoning strengths within current LLMs, although their performance remains inconsistent. These strengths should serve as foundational elements for developing future AI models in mental health settings.

Ethical implications and practical applications

The varying performance of LLMs across therapeutic tasks raises important ethical concerns about their potential deployment in mental health care. As argued by many scholars (Fiske et al., 2019; Lawrence et al.,

2024; Stade et al., 2024), the use of AI in therapeutic contexts requires careful consideration of privacy, consent, and potential harm. Our findings demonstrated that models sometimes made unfounded assumptions (a phenomenon known as hallucination in LLM literature) or recommended inappropriate interventions without clear rationale. These failures to demonstrate appropriate clinical reasoning highlight significant risks of unsupervised AI-client interactions. Recent studies have also questioned LLMs' reasoning ability, with some researchers arguing that it may stem from memorisation and recitation of learned information, and others suggesting chain-of-thought prompting to elicit better reasoning ability (Huang & Chang, 2023; Plaat et al., 2024; Wu et al., 2024; Yax et al., 2024). These safety concerns, alongside the areas for development discussed above, suggest that current LLMs might be most appropriately used as *clinical support tools* rather than direct therapeutic agents. For instance, they could assist clinicians with initial case formulation or help generate treatment options for consideration (Stade et al., 2024), while leaving final decisions and client interactions to trained professionals. This aligns with recent findings showing LLMs can enhance clinical decision-making in specific areas while avoiding the risks associated with direct therapeutic engagement (Kim et al., 2024).

Limitations and future directions

The interpretation of these findings requires careful consideration of both methodological constraints and the poor inter-rater reliability among evaluators. Given the small number of expert evaluators, the findings should be interpreted as exploratory. Although qualitative insights are valuable, larger and more diverse samples are needed to confirm the reliability and generalisability of the observed patterns.

The poor inter-rater reliability (ICC = .023) and significant effect of evaluator ($p < .001$) suggest that evaluators had different rating patterns and minimal agreement among them. These discrepancies likely stem from multiple sources, including variations in professional background, therapeutic orientation, expectations regarding the role of AI in clinical settings, and potential cognitive fatigue or reduced motivation during the extended rating process (Mahshanian & Shahnazari, 2020). Evaluating therapeutic responses is inherently subjective, particularly in the absence of a psychometrically validated rating tool. In this study, the prompts and rating process were designed to reflect common clinical tasks rather than function as a standardised measurement instrument. As such, the numerical scores should be interpreted as exploratory approximations of perceived appropriateness rather than formal indicators of clinical competence.

One source of potential bias is that all evaluators were aware they were assessing AI-generated content. Although responses were randomised and anonymised to reduce model-specific bias, the knowledge that responses were produced by AI may have influenced evaluator perceptions, particularly given prevailing scepticism towards AI in clinical contexts. Prior research has demonstrated that disclosure of an AI source tends to decrease evaluative ratings, even when AI-generated and human-generated content are of equal quality (Lim & Schmälzle, 2024; Parshakov et al., 2025). Future studies could use a mixed-source, blinded design in which evaluators assess both human- and AI-generated responses without knowing the source. This would help clarify the extent of rater bias due to source awareness.

Another source of variability may stem from evaluator's cultural backgrounds and implicit assumptions about what constitutes "appropriate" therapy may have shaped their judgements. Certain therapeutic approaches or formulations may align more closely with specific cultural or theoretical preferences. Future research should explore how such biases can be identified and reduced. For instance, prior studies have shown that inter-rater reliability can be improved through rater training, the use of validated rating measures, clarification of core competence constructs, and structured discussion of disagreements (Kühne et al., 2020; Paunov et al., 2024). However, these strategies may still fall short of capturing the full complexity of psychotherapy evaluation, particularly when assessments require implicit clinical reasoning, sensitivity to relational dynamics, or integration of diverse theoretical perspectives. Thus, while inter-rater reliability metrics such as the ICC provide valuable information, they offer only a partial view. Future work may benefit from developing complementary evaluation strategies that more fully account for the nuanced and context-sensitive nature of therapeutic competence.

One main limitation of our methodology was that it relied on a static, one-way format that limited the evaluation of dynamic therapeutic processes such as responsiveness to client input and alliance-

building. In addition, the clinical vignettes were relatively well-structured and information-rich, which may not reflect the sparse or ambiguous input often encountered in early clinical interactions, particularly with reluctant or reticent clients. Future research should examine LLM performance in response to more terse, incomplete, or ambiguous client presentations, in order to better simulate real-world clinical scenarios where information gathering is itself a therapeutic skill. Such studies could assess how well LLMs manage uncertainty, generate working hypotheses with limited data, and engage meaningfully with clients who offer minimal disclosure.

Our methodological limitations point to several directions for future research, which may vary in their level of clinical risk. At the lowest risk level, research could examine LLMs as clinical support tools for therapists, focusing on backend tasks such as treatment planning, case formulation, and documentation. Our findings suggest that LLMs possess basic theoretical understanding and procedural knowledge useful for these functions. However, their responses frequently lacked robust clinical reasoning and failed to meaningfully integrate client-specific information. Therefore, future research should explore ways to improve LLM capabilities. One method would be training models on curated clinical datasets that incorporate established reasoning frameworks, such as the abductive theory of method (ATOM) and transdiagnostic approaches (Dalgleish et al., 2020; Frank & Davidson, 2014; Vertue & Haig, 2008; Ward et al., 2016). Another approach could involve supervised fine-tuning with expert clinical psychologists. Structured expert feedback is similar to supervision practices in therapist training. It could be used to identify and correct response limitations related to clinical reasoning, ethical appropriateness, and therapeutic boundaries. This approach has shown promise in related domains, where fine-tuning on multi-turn, empathic dialogue data improved models' ability to engage in emotionally attuned and supportive communication (Chen et al., 2023). Another low-risk implementation is the potential use of AI-based evaluators to assess AI-generated responses. Although human raters bring valuable contextual insight and ethical judgement, the use of AI-based evaluators may offer scalable and consistent assessment of AI responses. Future research could explore whether such systems exhibit similar biases or leniencies as humans, and whether hybrid evaluation approaches produce more balanced and reliable assessments (Burden, 2024).

At a moderate risk level, research could explore LLMs' potential for supervised client interaction, such as providing basic psychoeducation, supporting homework completion, offering structured coping skills practice and assisting treatment planning with psychologists (Auf et al., 2025; Lundin et al., 2023). While our study showed that LLMs can outline therapeutic concepts, their tendency to provide generic responses and occasional inappropriate recommendations highlights the necessity for careful monitoring and clear protocols for professional oversight. Investigating how different cultural groups interact with AI-assisted services would also be crucial, given the challenges in providing culturally appropriate mental health services (Kapeli et al., 2020).

At the highest risk level involving direct therapeutic application, further studies would be required to ensure high-quality clinical reasoning and dynamic therapeutic interaction. This research direction requires experimental studies examining LLMs' ability to maintain therapeutic alliance over multiple sessions, as the relationship is a robust predictor of therapy outcomes, with substantial evidence linking it to treatment success (Norcross & Lambert, 2018; Wampold, 2015). Longitudinal case progression studies would be needed to assess LLMs' capabilities for sustained therapeutic engagement while addressing technological limitations like context window constraints. Such research would need to carefully address fundamental concerns about consent, confidentiality, and clinical reasoning before any implementation of direct client-AI therapeutic interaction.

Furthermore, comparative studies between LLMs and human clinicians across different levels of expertise would offer valuable benchmarks for interpreting AI performance in therapeutic settings. In this study, LLM responses were evaluated against the clinical judgement of experienced psychologists, providing a valuable but partial reference point. Future research could present standardised clinical scenarios to LLMs, psychology students, early-career clinicians, and experienced practitioners to examine the extent to which LLM-generated responses reflect different levels of clinical expertise. For instance, researchers could assess whether these responses demonstrate formulation-driven reasoning, the ability to tailor interventions to complex client presentations, and the integration of evidence-based strategies across modalities. These comparisons would not only clarify LLMs' relative capabilities but also inform their potential use in clinical training contexts, such as supporting student learning, offering feedback on basic therapeutic formulation,

or modelling structured communication. Such insights would support evidence-based decisions about where and how LLMs may be responsibly integrated into clinical training or practice.

The integration of AI and LLMs into mental health support appears increasingly inevitable in contemporary society. One of LLMs' primary advantages is that users report feeling less likely to be judged by AI, which encourages help-seeking behaviours and increases comfort in disclosing personal struggles compared to traditional social interactions (Lee et al., 2024; Ma et al., 2023). In addition, LLMs may offer a cost-effective alternative to traditional therapy, particularly at scale, by reducing clinician burden, administrative workload, and wait times (Spytska, 2025; Zhang & Wang, 2024). This economic advantage could become a central motivator for adoption in overstretched health systems. This growing use of AI for informal mental health support raises several important questions for future research: (1) how existing attitudes towards AI influence engagement with AI for mental health support (Keung & So, 2025), (2) how awareness of AI-generated responses affects user expectations and interaction patterns, (3) how to ensure AI upholds principles that honour user autonomy and promote independence to avoid overreliance solely on AI support (Hua et al., 2024), and (4) how to establish appropriate guidelines and safeguards for the informal use of AI in mental health support. These questions are especially salient given the increasing accessibility and uptake of LLMs by the general public.

Conclusions

Our evaluation study of LLMs in therapeutic contexts revealed both promising capabilities and significant limitations that warrant careful consideration. While no single model consistently produced high-quality responses across all therapeutic tasks, models demonstrated competence in structured tasks such as determining therapy session length and discussing goal-setting in therapy. Higher-rated responses displayed important professional characteristics such as clinical humility, appropriate therapeutic boundaries, and recognition of therapy as a collaborative process. However, models' weaknesses in clinical reasoning and practical implementation, along with ethical concerns about consent and inappropriate recommendations, suggest that their most immediate value lies in supporting clinicians rather than providing direct therapeutic care. The primary area for development appears to be integrative clinical reasoning, a vital aspect of key clinical tasks such as developing quality formulations and safe and efficient treatment planning. Once the models are proven to have good clinical reasoning capabilities, future research in therapeutic alliance formation and longitudinal engagement will be essential to develop AI-assisted therapy. However, all these processes require attention and supervision to ensure professional standards and client welfare. The path forward requires thoughtful integration that enhances rather than compromises the quality of mental health care delivery, optimising AI's capabilities while preserving the essential human connection at the heart of effective therapy.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Kean Sian Tan http://orcid.org/0009-0001-2786-5042
Matti Cervin http://orcid.org/0000-0003-1188-8706
Patrick Leman http://orcid.org/0000-0003-1708-029X
Kristopher Nielsen http://orcid.org/0000-0001-8001-8056
Oleg Medvedev http://orcid.org/0000-0002-2167-5002

Author contributions

- Kean Sian Tan: Conceptualisation, Data Curation, Formal Analysis, Methodology, Project Administration, Validation, Visualisation, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing.
- Matti Cervin: Resources, Validation, Writing – Review & Editing

- Patrick Leman: Conceptualisation
- Kristopher Nielsen: Resources, Validation, Writing – Review & Editing
- Prashanth Vasantha Kumar: Resources
- Oleg Medvedev: Conceptualisation, Data Curation, Project Administration, Resources, Supervision, Writing – Review & Editing

Data availability statement

Data is available as a supplementary file associated with this paper.

Ethical approval

Ethics approval (FS2024–38) was granted by the Human Research Ethics Committee at the University of Waikato.

References

Abd-Alrazaq, A., Alhuwail, D., Schneider, J., Toro, C. T., Ahmed, A., Alzubaidi, M., Alajlani, M., & Househ, M. (2022). The performance of artificial intelligence-driven technologies in diagnosing mental disorders: An umbrella review. *NPJ Digital Medicine*, *5*(1), 1–12. https://doi.org/10.1038/s41746-022-00631-8

Alize, J. F., Damian, F. S., Mantilla Herrera Ana, M., Jamileh, S., Ashbaugh, C., Holly, E. E., Fiona, J, J. C., Louisa, D., James, G. S., McGrath, J. J., Allebeck, P., Benjet, C., Nicholas, J. K. B., Brugha, T., Dai, X., Dandona, L., Dandona, R., Fischer, F., Haagsma, J. A, . . . Harvey, A. W. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Al-Turki, D., Hettiarachchi, H., Medhat Gaber, M., Abdelsamea, M. M., Basurra, S., Iranmanesh, S., Saadany, H., & Vakaj, E. (2024). Human-in-the-loop learning with LLMs for efficient RASE tagging in building compliance regulations. *IEEE Access*, *12*, 185291–185306. https://doi.org/10.1109/ACCESS.2024.3512434

Anthropic. (2024a, March 4). *Introducing the next generation of Claude*. https://www.anthropic.com/news/claude-3-family

Anthropic. (2024b, June 21). *Claude 3.5 sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet

Auf, H., Svedberg, P., Nygren, J., Nair, M., & Lundgren, L. E. (2025). The use of AI in mental health services to support decision-making: Scoping review. *Journal of Medical Internet Research*, *27*, e63548. https://doi.org/10.2196/63548

Banerjee, S., Dunn, P., Conard, S., & Ali, A. (2024). Mental health applications of generative AI and large language modeling in the United States. *International Journal of Environmental Research and Public Health*, *21*(7), Article 7. https://doi.org/10.3390/ijerph21070910

Baumgardner, M., & Benoit Allen, K. (2024). Integrating cognitive-behavioral therapy with compassion-focused therapy for the treatment of social anxiety disorder: An evidence-based case study. *Clinical Case Studies*, *23*(2), 127–145. https://doi.org/10.1177/15346501231197403

Bill, D., & Eriksson, T. (2023). *Five facets of mindfulness and psychological health: Evaluating a psychological Model of the mechanisms of mindfulness*. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-331920

Burden, J. (2024). *Evaluating AI evaluation: Perils and prospects* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.09221

Çam, M. O., & Uğuryol, M. (2019). From mental disorder to recovery: Cultural effect. *Psikiyatride Güncel Yaklaşımlar*, *11*(1), 55–64. https://doi.org/10.18863/pgy.391783

Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in internet interventions: The advent of ChatGPT and AI-assisted therapist guidance. *Internet Interventions*, *32*, 100621. https://doi.org/10.1016/j.invent.2023.100621

Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., Harrington, H., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L. J. H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B. S., & Moffitt, T. E. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study. *JAMA Network Open*, *3*(4), e203221. https://doi.org/10.1001/jamanetworkopen.2020.3221

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations *[Preprint]. arXiv*. https://doi.org/10.48550/arXiv.2311.00273

Chiu, Y. Y., Sharma, A., Lin, I. W., & Althoff, T. (2024). *A computational framework for behavioral assessment of LLM therapists* [Preprint]. arXiv. http://arxiv.org/abs/2401.00820

Christon, L. M., McLeod, B. D., & Jensen-Doss, A. (2015). Evidence-based assessment meets evidence-based treatment: An approach to science-informed case conceptualization. *Cognitive and Behavioral Practice*, *22*(1), 36–48. https://doi.org/10.1016/j.cbpra.2013.12.004

Dalgleish, T., Black, M., Johnston, D., & Bevan, A. (2020). Transdiagnostic approaches to mental health problems: Current status and future directions. *Journal of Consulting & Clinical Psychology*, *88*(3), 179. https://doi.org/10.1037/ccp0000482

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). *A complete survey on LLM-based AI chatbots* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.16937

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*, 1199058. https://doi.org/10.3389/fpsyg.2023.1199058

Every-Palmer, S., Grant, M. L., & Thabrew, H. (2022). Young people don't tend to ask for help more than once: Child and adolescent psychiatrists' views on ailing mental health services for young New Zealanders. *Australasian Psychiatry*, *30*(6), 684–688. https://doi.org/10.1177/10398562221115624

Every-Palmer, S., Grant, M. L., Thabrew, H., Hansby, O., Lawrence, M., Jenkins, M., & Romans, S. (2023). Not heading in the right direction: Five hundred psychiatrists' views on resourcing, demand, and workforce across New Zealand mental health services. *The Australian and New Zealand Journal of Psychiatry*, *58*(1), 82–91. https://doi.org/10.1177/00048674231170572

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, *21*(5), e13216. https://doi.org/10.2196/13216

Frank, R. I., & Davidson, J. (2014). *The transdiagnostic road map to case formulation and treatment planning: Practical guidance for clinical decision making* (pp. xii, 239). New Harbinger Publications.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., Malgaroli, M., Smith, R. N., Rothbaum, B. O., Ressler, K. J., Galatzer-Levy, I. R., & Powers, A. (2023). *The capability of large language models to measure psychiatric functioning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.01834

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference*. Routledge. http://ebookcentral.proquest.com/lib/waikato/detail.action?docID=4455907

Gilbert, P. (2016). A biopsychosocial and evolutionary approach to formulation. In N. Tarrier & J. Johnson (Eds.), *Case formulation in cognitivebehaviour therapy:* The treatment of challenging and complex cases (2nd ed., pp. 52–89). Routledge/Taylor & Francis Group.

Goetter, E. M., Frumkin, M. R., Palitz, S. A., Swee, M. B., Baker, A. W., Bui, E., & Simon, N. M. (2020). Barriers to mental health treatment among individuals with social anxiety disorder and generalized anxiety disorder. *Psychological Services*, *17*(1), 5–12. https://doi.org/10.1037/ser0000254

Google. (2024, July). *Gemini Pro*. Google DeepMind. https://deepmind.google/technologies/gemini/pro/

Griner, D., & Smith, T. B. (2006). Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy (Chicago, Ill)*, *43*(4), 531–548. https://doi.org/10.1037/0033-3204.43.4.531

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*(1), e57400. https://doi.org/10.2196/57400

Hall, G. C. N., Ibaraki, A. Y., Huang, E. R., Marti, C. N., & Stice, E. (2016). A meta-analysis of cultural adaptations of psychological interventions. *Behavior Therapy*, *47*(6), 993–1014. https://doi.org/10.1016/j.beth.2016.09.005

Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y., Zhou, P., Moran, L. V., Ananiadou, S., Beam, A., & Torous, J. (2024). *Large language models in mental health care: A scoping review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.02984

Huang, J., & Chang, K.-C.-C. (2023). *Towards reasoning in large language models: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2212.10403

Kapeli, S., Manuela, S., & Sibley, C. (2020). Understanding Pasifika mental health in New Zealand: A review of the literature. *MAI Journal: A New Zealand Journal of Indigenous Scholarship*, *9*(3). https://doi.org/10.20507/MAIJournal.2020.9.3.7

Ke, L., Tong, S., Cheng, P., & Peng, K. (2024). *Exploring the frontiers of LLMs in psychological applications: A comprehensive review* [Preprint]. arXiv. *Artificial Intelligence Review*, *58*(10). https://doi.org/10.48550/arXiv.2401.01519

Keung, W. M., & So, T. Y. (2025). Attitudes towards AI counseling: The existence of perceptual fear in affecting perceived chatbot support quality. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1538387

Kim, J., Leonte, K. G., Chen, M. L., Torous, J. B., Linos, E., Pinto, A., & Rodriguez, C. I. (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, *7*(1), 1–5. https://doi.org/10.1038/s41746-024-01181-x

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155. https://doi.org/10.1016/j.jcm.2016.02.012

Kühne, F., Meister, R., Maaß, U., Paunov, T., & Weck, F. (2020). How reliable are therapeutic competence ratings? Results of a systematic review and meta-analysis. *Cognitive Therapy and Research*, *44*(2), 241–257. https://doi.org/10.1007/s10608-019-10056-5

Kulshrestha, V., & Shahid, S. M. (2022). Barriers and drivers in mental health services in New Zealand: Current status and future direction. *Global Health Promotion*, *29*(4), 83–86. https://doi.org/10.1177/17579759221099312

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, *11*(1), e59479. https://doi.org/10.2196/59479

Lee, J., Lee, D., & Lee, J. (2024). Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. *International Journal of human–Computer Interaction*, *40*(7), 1620–1631. https://doi.org/10.1080/10447318.2022.2146227

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages: A two-part study. *Computers in Human Behavior Artificial Humans*, *2*(1), 100058. https://doi.org/10.1016/j.chbah.2024.100058

Lindhiem, O., Bennett, C. B., Orimoto, T. E., & Kolko, D. J. (2016). A meta-analysis of personalized treatment goals in psychotherapy: A preliminary report and call for more studies. *Clinical Psychology Science & Practice*, *23*(2), 165–176. https://doi.org/10.1111/cpsp.12153

Lundin, R. M., Berk, M., & Østergaard, S. D. (2023). ChatGPT on ECT: Can large language models support psychoeducation? *The Journal of ECT*, *39*(3), 130. https://doi.org/10.1097/YCT.0000000000000941

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings* (Vol. 2023, pp. 1105–1114). AMIA Symposium.

Mahshanian, A., & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics*, *10*(1), Article 1. https://doi.org/10.17509/ijal.v10i1.24956

McGrath, J. J., Al-Hamzawi, A., Alonso, J., Altwaijri, Y., Andrade, L. H., Bromet, E. J., Bruffaerts, R., de Almeida, J. M. C., Chardoul, S., Chiu, W. T., Degenhardt, L., Demler, O. V., Ferry, F., Gureje, O., Haro, J. M., Karam, E. G., Karam, G., Khaled, S. M., Kovess-Masfety, V., . . . Zaslavsky, A. M. (2023). Age of onset and cumulative risk of mental disorders: A cross-national analysis of population surveys from 29 countries. *Lancet Psychiatry*, *10*(9), 668–681. https://doi.org/10.1016/S2215-0366(23)00193-1

Meta. (2024, July 23). *Introducing Llama 3.1: Our most capable models to date*. Meta AI. https://ai.meta.com/blog/meta-llama-3-1/

Microsoft. (2024, July). *Latest updates for Microsoft copilot*. https://support.microsoft.com/en-us/topic/latest-updates-for-microsoft-copilot-a5685141-8081-458c-80d6-42493aad51ed

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A comprehensive overview of large language models* [Preprint]. arXiv. http://arxiv.org/abs/2307.06435

Norcross, J. C., & Lambert, M. J. (2018). Psychotherapy relationships that work III. *Psychotherapy*, *55*(4), 303–315. https://doi.org/10.1037/pst0000193

Open, AI. (2024a, May 13). *Introducing GPT-4o and more tools to ChatGPT free users*. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

Open, AI. (2024b, July). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*. https://openai.com/index/gpt-4/

Open, AI. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2024c). *GPT-4 technical report* [Technical Report]. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Page, A., Stritzke, W., & Mclean, N. (2008). Toward science-informed supervision of clinical case formulation: A training model and supervision method. *Australian Psychologist*, *43*(2), 88–95. https://doi.org/10.1080/00050060801994156

Parshakov, P., Naidenova, I., Paklina, S., Matkin, N., & Nesseler, C. (2025). *Users favor LLM-generated content–until they know it's AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.16458

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., . . . UnÜtzer, J. (2018). The lancet commission on global mental health and sustainable development. *Lancet*, *392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X

Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (LLMs). *Applied Sciences*, *14*(5), 2074. https://doi.org/10.3390/app14052074

Paunov, T., Weck, F., Heinze, P. E., Maaß, U., & Kühne, F. (2024). Competence ratings in psychotherapy training – a complex matter. *Cognitive Therapy and Research*, *48*(3), 500–510. https://doi.org/10.1007/s10608-023-10445-x

Persons, J. B. (2006). Case formulation–driven psychotherapy. *Clinical Psychology Science & Practice*, *13*(2), 167–170. https://doi.org/10.1111/j.1468-2850.2006.00019.x

Pines, A. M. (2002). The female entrepreneur: Burnout treated using a psychodynamic existential approach. *Clinical Case Studies*, *1*(2), 170–180. https://doi.org/10.1177/1534650102001002005

Plaat, A., Wong, A., Verberne, S., Broekens, J., Stein, N. V., & Back, T. (2024). *Reasoning with large language models, a survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.11511

Rony, M. K. K., Das, D. C., Khatun, M., Ferdousi, T., Akter, S., R, M., Khatun, M., Begum, A., Most, H., Khalil, M. I., Parvin, M., Alrazeeni, R., M, D., & Akter, F. (2025). Artificial intelligence in psychiatry: A systematic review and meta-analysis of diagnostic and therapeutic efficacy. *Digital Health*, *11*, 20552076251330528. https://doi.org/10.1177/20552076251330528

Scoles, P. (2022). Recovery assessment: Culture, ethnicity, and spiritual dysfunction. *Journal of Behavioral Health and Psychology*, *11*(1). https://doi.org/10.33425/2832-4579/22040

Sim, K., Gwee, K. P., & Bateman, A. (2005). Case formulation in psychotherapy: Revitalizing its usefulness as a clinical tool. *Academic Psychiatry*, *29*(3), 289–292. https://doi.org/10.1176/appi.ap.29.3.289

Spytska, L. (2025). The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. *BMC Psychology*, *13*(1), 175. https://doi.org/10.1186/s40359-025-02491-9

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *NPJ Mental Health Research*, *3*(1), 1–12. https://doi.org/10.1038/s44184-024-00056-z

Sufyan, N. S., Fadhel, F. H., Alkhathami, S. S., & Mukhadi, J. Y. A. (2024). Artificial intelligence and social intelligence: Preliminary comparison study between AI models and psychologists. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1353022

Vertue, F. M., & Haig, B. D. (2008). An abductive perspective on clinical reasoning and case formulation. *Journal of Clinical Psychology*, *64*(9), 1046–1068. https://doi.org/10.1002/jclp.20504

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wang, J., Xiao, Y., Li, Y., Song, C., Xu, C., Tan, C., & Li, W. (2024). Towards a client-centered assessment of LLM therapists by client simulation *[Preprint]. arXiv*. http://arxiv.org/abs/2406.12266

Ward, T., Clack, S., & Haig, B. D. (2016). The abductive theory of method: Scientific inquiry and clinical practice. *Behaviour Change*, *33*(4), 212–231. https://doi.org/10.1017/bec.2017.1

World Health Organization. (2020). *Mental health atlas 2020*. https://iris.who.int/bitstream/handle/10665/345946/9789240036703-eng.pdf?sequence=1

World Health Organization. (2022). *Mental disorders*. Retrieved September 15, 2024, from https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). *Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.02477

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2023). *Mental–LLM: Leveraging large language models for mental health prediction via online text data* [Preprint]. arXiv. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–32. https://doi.org/10.48550/arXiv.2307.14385

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Yu, H. Q., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, *7* Article 0. 16–16. https://doi.org/10.21037/jmai-23-136

Zhang, Z., & Wang, J. (2024). Can AI replace psychotherapists? Exploring the future of mental health care. *Frontiers in Psychiatry*, *15*, 1444382. https://doi.org/10.3389/fpsyt.2024.1444382