A Comprehensive Literature Review on [LLMs For Mental Health Therapeutics]

Your Name
Your Affiliation

April 22, 2025

Abstract

This literature review examines the evolving landscape of [LLMs For Mental Health Therapeutics] research. It synthesizes key findings from [Number] studies published between [Year Range], focusing on [Key Themes]. The review identifies [Key Gaps/Contradictions] and proposes [Potential Future Directions].

Contents

1 Introduction

article lipsum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2 History of LLMs

article lipsum

2.1 Subtopic 1.1

Einstein introduced the theory of relativity in 1905 Arriba-Pérez and García-Méndez (2024).

2.2 Subtopic 1.2

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

3 Efficacy of Mental Health Chat Bots

Recent advancements in large language models (LLMs) have opened up more relevance in the field of digital mental health, offering scalable solutions for early detection, diagnosis and support. But the efficacy of the models is a very important consideration.

This section explores the data sources, fine-tuning methods, performance trade-offs, and ethical implications of using AI-driven chatbots in mental health contexts. We highlight the impact of training data biases, discuss the balance between speed and accuracy and examine real-world risks including misinformation, user trust and demographic disparities.

4 Model Training and Evaluation of LLMs in Mental Health

4.1 Model Training and Biases

Chatbots and AI models designed for mental health support are predominantly trained on datasets derived from social media platforms, psychotherapy transcripts, and electronic health records (EHRs). A majority of models in the literature have been fine-tuned specifically for mental health prediction tasks using benchmark datasets such as *Dreaddit*, *DepSeverity*, *SDCNL*, and *CSSRS-Suicide* Arriba-Pérez and García-Méndez (2024). These datasets are often annotated by human experts and sourced primarily from Reddit, enabling models to perform tasks such as binary stress classification, depression detection, and suicide risk estimation. While these data sources provide scalability, they also introduce significant demographic and cultural biases.

Bias remains a persistent challenge. Studies have highlighted how the overreliance on Reddit-based data skews model generalizability, often failing to capture the nuances of underrepresented groups Li et al. (2023); Gbollie et al. (2023). Diagnostic models trained on such datasets may inadvertently reinforce users' distress rather than challenge cognitive distortions, potentially increasing false positives. Furthermore, a lack of diversity in training data has led to measurable racial and gender biases in chatbot responses Arriba-Pérez and García-Méndez (2024). Even expert-annotated corpora are susceptible to embedded stereotypes and confirmation bias, raising concerns about the validity of ground-truth labels Greco et al. (2023).

4.2 Trade-Offs in Performance and Design

Performance trade-offs are evident across model size, inference method, and processing strategy. Larger general-purpose models such as GPT-4 and FLAN-T5 offer faster response times, while smaller fine-tuned models like Mental-Alpaca often achieve higher task-specific accuracy Xu et al. (2024). Improved prompt engineering can enhance inference speed, but often at the cost of response quality. Stream-processing models allow for real-time interaction but demand continuous updating, whereas batch-processing models deliver greater accuracy by processing data in aggregate — albeit with delayed outputs McGorry et al. (2025).

Fine-tuning strategies play a central role in improving performance. Approaches such as instruction tuning and LoRA (Low-Rank Adaptation) have been applied to enhance

LLMs for cognitive behavioral therapy (CBT) tasks. These models optimize predictions using cross-entropy loss, reducing divergence from human-labeled data Na (2024).

4.3 Accuracy, Coherence, and Logical Structure

Transformer-based models have shown high accuracy in mental health tasks — with some detecting depression at rates exceeding 96%, and chatbot-based models for cognitive impairment reaching over 80% accuracy and 85% recall Greco et al. (2023); McGorry et al. (2025). Fine-tuned models such as Mental-FLAN-T5 consistently outperform zero-shot models like GPT-4 by up to 4.8% in balanced accuracy Xu et al. (2024). However, zero-shot performance remains inconsistent, often fluctuating between 50% and 83% based on prompt structure.

In terms of coherence, rule-based conversational agents like Woebot and Wysa tend to outperform generative models in logical consistency and structure Stade et al. (2024). While GPT-4 responses are generally fluent, they can be overly generic and lack context-specific insight. In contrast, fine-tuned models offer more domain-relevant outputs. Notably, GPT-based models are sensitive to prompt phrasing, leading to variability in output quality and occasional overgeneralization — particularly in nuanced cases such as suicide risk Sejnowski (2023).

4.4 Risk of Misinformation and Ethical Implications

Numerous studies caution against the misuse of LLMs in high-stakes contexts. Despite their sophistication, generative models are prone to producing "falsely reasonable" yet incorrect outputs. This poses a significant risk in mental health settings, where the consequences of misinformation can be severe. The absence of robust, systematic evaluation frameworks for mental health reasoning further compounds the issue Greco et al. (2023). In addition, the use of sensitive training data raises major concerns about privacy, consent, and ethical safeguards Arriba-Pérez and García-Méndez (2024).

The potential for discriminatory outputs is fairly high given the lack of fairness assessments across different subgroups of population. Vulnerable users, including those in crisis, are particularly susceptible to receiving inaccurate or even harmful advice.

4.5 Real-Time Performance vs. Quality

While real-time models like GPT-4 excel in responsiveness, fine-tuned models such as Mental-FLAN-T5 yield more accurate and context-sensitive results with lower computational demands Xu et al. (2024). Enhancing model accuracy through prompt engineering often increases processing time. However, instruction-fine-tuned models manage to balance inference speed with depth of response more efficiently than zero-shot systems.

4.6 Handling Ambiguity in User Prompts

Fine-tuned models have demonstrated superior contextual understanding of vague or ambiguous queries related to stress or depression. In contrast, zero-shot models frequently default to binary or overly simplified interpretations. Few-shot learning methods have shown to improve GPT-4's performance by around 4.1% on ambiguous mental health queries Xu et al. (2024), indicating the need for more adaptable model architectures in this space.

4.7 User Willingness and Perceptions

User receptivity remains high, with 81.1% of university students reporting willingness to engage with mental health chatbots — though only 4% have done so in practice Gbollie et al. (2023). Younger demographics (18–35) show higher adoption rates, while users in acute distress often prefer human counselors due to trust and safety concerns.

Preferences are influenced by multiple factors. Stigma around seeking therapy makes some users favour AI over traditional face-to-face approaches. Conversational AI models like GPT-4 are generally preferred due to flexible and natural interaction. Rule-based model responses can often be repeated due to a fixed response database. However, concerns around privacy, transparency, and the potential for AI-generated errors remain key barriers to adoption.

4.8 Outlook and Efficacy

Emerging models such as CBT-LLM exemplify domain-specific fine-tuning tailored for psychological support. Evaluations show that such models outperform generic LLMs in both fluency and therapeutic relevance Na (2024). Despite their promise, several researchers question whether LLMs truly "understand" mental health concerns or merely reflect user input in plausible ways — a phenomenon likened to a "reverse Turing test" Sejnowski (2023).

A broader body of research suggests that while LLMs may augment therapists and automate diagnostics, they are not yet reliable standalone tools. Studies have shown marked symptom improvement using chatbot interventions (Hedge's g = 0.64 for depression and 0.7 for distress) Stade et al. (2024), but little impact on long-term psychological well-being. Lastly, transformer-based models like BERT continue to demonstrate high efficacy for mental health classification, though their success remains closely tied to data quality and fine-tuning depth Greco et al. (2023); Li et al. (2023).

5 Privacy Concerns

article natbib

... Some text ... Gbollie et al. (2023). More text... (Sejnowski, 2023) ...

6 Conclusion

article lipsum

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat

sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

References

Arriba-Pérez, F. and García-Méndez, S. (2024). Leveraging llms for real-time mental health predictions. *Arabian Journal for Science and Engineering*.

Gbollie, E. F., Bantjes, J., and Jarvis, L. (2023). Intention to use digital mental health solutions. *Digital Health*, 9:1–19.

Greco, C. M., Simeri, A., and Tagarelli, A. (2023). Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters*, 167:204–211.

Li, H., Zhang, R., and Kraut, R. E. (2023). Systematic review and meta-analysis of ai-based conversational agents for mental health. *npj Digital Medicine*, 6:236.

McGorry, P., Gunasiri, H., Mei, C., Rice, S., and Gao, C. X. (2025). The youth mental health crisis: analysis and solutions. *Frontiers in Psychiatry*, 15:1517533.

Na, H. (2024). Cbt-llm: A chinese large language model for cognitive behavioral therapy. arXiv.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural Computation*, 35:309–342.

Stade, E. C., Stirman, S. W., and Ungar, L. H. (2024). Large language models could change the future of behavioral healthcare. *npj Mental Health Research*, 3:12.

Xu, X., Yao, B., and Dong, Y. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. In *Proceedings of ACM Interact.*