Evaluating the Readability and Reliability of Large Language Model Generated Information About Anxiety and Depression

Authors:

1. **Gargi Porwal, B.S.A**
   Department of Computer Science
   The University of Texas at Austin, Austin, Texas, USA

2. **Kapil Laddha, DDS**
   Department of Dentistry
   University of Colorado School of Dental Medicine, Aurora, Colorado, USA

3. **Jitendra Jeenger, MD**
   Department of Psychiatry
   Geetanjali Medical College and Hospital, Udaipur, Rajasthan, India

Abstract

Artificial intelligence (AI) large language models (LLMs) hold great potential to transform psychiatry and mental health care by delivering relevant and tailored mental health information. This study aimed to evaluate the quality of mental health information generated by LLMs by determining their level of accessibility, reliability, and interpreting any bias present. Generative Pre-trained Transformer-4 (GPT-4) (San Francisco, California: Open AI), Gemini 1.5 Flash (Mountain View, California: Google LLC), and Large Language Model Meta AI 3.2 (Llama 3.2) (Menlo Park, California: Meta Inc.) were prompted with 20 questions commonly asked about anxiety and depression. The responses were evaluated using Flesch-Kincaid readability tests to quantify their ease of understanding through grade level and reading ease score measures. The text was subsequently analyzed using a modern DISCERN score to assess the reliability of the health-related information presented. Finally, the LLM responses were evaluated for stigmatizing language using communication and language guidelines for mental health. A significant difference in grade levels and reading ease scores was observed between GPT-4 and Llama ($p < 0.01$) and Gemini and Llama($p < 0.01$), with both GPT-4 and Gemini having higher readability scores. No significant difference was observed in the grade level and reading ease scores between GPT-4 and Gemini ($p > 0.05$). All three models reported moderate reliability scores but no significant differences were observed ($p > 0.05$). GPT-4, Gemini, and Llama included stigmatizing phrases in 10%, 15%, and 20% of their responses respectively. These phrases were attributed to descriptions of mental health conditions and substance use; however, no significant differences were observed in the proportions of stigmatizing phrases present across the three models ($p > 0.05$). Furthermore, comparisons between anxiety and depression responses for each model revealed no significant differences in readability, reliability, or bias. All models demonstrated the ability to generate mental health information that generally satisfied criteria for accessibility, reliability, and minimal bias, with GPT-4 and Gemini reporting higher readability than Llama. However, the lack of additional resources cited in responses and the occasional presence of certain stigmatizing phrases, indicate that additional model training and fine-tuning, specifically for mental health applications, may be necessary before these tools can be deployed on a large scale for mental health care.

Introduction

Globally, there is a mental health crisis, as half of all individuals are expected to experience a mental health disorder in their lifetimes [1]. Access to mental health care remains unable to meet this increasing demand due to several challenges, including a shortage of trained professionals, financial barriers, and stigma that prevents many individuals from seeking help [2,3]. As societal gaps limit access to care, people are increasingly turning to digital tools to receive timely support [4].

Generative artificial intelligence (AI) large language models (LLMs) are quickly becoming one such tool. With their ability to interpret, summarize, and generate texts, LLMs have become popular tools for web-based information-seeking and have been utilized for mental health support by people experiencing distress but lacking access to immediate help. These tools provide support in various ways, including helping teens cope with anxiety and depression, guiding users through cognitive behavioral therapy (CBT) techniques, and providing crisis interventions[5-7].

AI tools provide avenues for scalable and personalized interventions that broaden access to mental health care to remote and underserved populations. While LLMs hold promise for enhancing diagnostic accuracy and delivering personalized interventions, they also present risks, including generating hallucinations that may convey inaccurate information or content that perpetuates bias or violates ethical concerns [8]. Therefore, it becomes imperative to evaluate the accessibility, accuracy, and clarity of the information generated by LLMs to ensure it is understandable by diverse populations with varying health literacy levels. By analyzing readability and complexity, we can determine if the information is presented in a manner that minimizes the risk of misinterpretation and promotes effective understanding.

Given the opportunities and risks associated with LLMs, it is important to assess how models respond to common questions and concerns about mental health conditions, particularly anxiety and depression, which are among the most prevalent disorders [9]. In this study, Generative Pre-Trained Transformer-4 (GPT-4)(San Francisco, California: Open AI), Gemini 1.5 Flash (Mountain View, California: Google LLC), Large Language Model Meta AI (Llama) 3.2 (Menlo Park, California: Meta Inc.) were prompted with queries about anxiety and depression, and their responses were evaluated for readability, reliability, and presence of stigmatizing language.

Methods

Open AI's GPT-4, Google's Gemini 1.5 Flash, and Meta's Llama 3.2 were used to perform this cross-sectional study.

The tools were prompted with 20 questions about mental health conditions, including 10 questions each about anxiety and depression. The questions addressed diagnostic procedures, treatment options, and guidance on accessing support resources for both conditions [10-14].

Table 2 in the Supplementary Materials section provides a list of all the questions used in this study.

The responses generated by the LLMs were analyzed using the Flesch-Kincaid calculator, which evaluated word count, sentence count, average words per sentence, and average syllables per word. These metrics were utilized to compute the Flesch-Kincaid grade level and reading ease scores, which quantify the readability and ease of understanding of each response.

A modified DISCERN score assessed the reliability of each response [15]. The tool utilized a five-question scale, with each question scored as 0 or 1 based on whether the response met specified criteria. A final score of 5 indicates high reliability, whereas a score of 0 suggests the text lacks scientific reliability.

Responses were evaluated for biases using six principles derived from media mental health communication guidelines: stigma reduction, promotion of help-seeking behavior, avoidance of sensationalism, adherence to evidence-based content, person-centered language, and empowering language [16,17]. Each response was rated as 0 or 1, where 0 indicated that the response contained at least 1 stigmatizing phrase, and 1 suggested that the response did not contain any stigmatizing language. The rating was performed by comparing each response against the list of preferred and problematic phrases listed in Mindframe's guide, *Our words matter: Guidance of Language Usage* [16]. All authors rated the responses according to these guidelines, and the majority rating was used for further analysis.

A one-way ANOVA compared the mean Flesch-Kincaid grade level, reading ease score, and modified DISCERN score among GPT-4, Gemini, and Llama. Significant differences identified by the ANOVA test were further analyzed using a post-hoc Tukey HSD test for pairwise comparisons to determine which specific pairs among GPT-4, Gemini, and Llama showed statistically significant differences. An unpaired t-test compared the means values of each variable for depression and anxiety responses within each model. A two-proportion Z-test was conducted to assess whether the adherence rates to communication guidelines significantly differed among the GPT-4, Gemini, and Llama models. Additionally, the test examined whether adherence varied between responses addressing anxiety and depression within each model. Statsmodels Python package (Python Software Foundation) was used to perform the statistical significance tests. P-value < 0.05 was considered statistically significant.

Results

Readability and Ease of Understanding

GPT-4, Gemini, and Llama were each prompted with 20 questions about anxiety and depression.

Significant differences were identified for mean word count ($p = 2.46e-10$), mean sentence count ($p = 1.89e-8$), and average syllables per word ($p = 0.0148$) across responses generated

by GPT-4, Gemini, and Llama. Post-hoc pairwise comparisons revealed statistically significant differences in word count across all three models. For sentence count, GPT-4 had a significantly higher mean compared to Gemini and Llama, but no significant difference in sentence count was observed between Gemini and Llama. The average syllables per word were significantly higher for Llama than GPT-4, but no significant differences were found between GPT-4 and Gemini or Gemini and Llama. No significant differences were observed for the average number of words per sentence.

There were significant differences in the Flesch-Kincaid grade level ($p = 0.00176$) and reading ease scores ($p = 0.00453$). Llama had a significantly higher grade level and lower reading ease score than both GPT-4 and Gemini (Figure 1A, 1B). GPT-4 had a lower grade level and higher reading ease score than Gemini but these differences were not statistically significant.

Comparisons between responses addressing anxiety and depression for each model revealed that Gemini generated significantly higher word and sentence counts for anxiety-related responses compared to those for depression. Similarly, GPT-4 produced significantly higher average words per sentence for anxiety responses than for depression. For each LLM, the grade level was higher and the reading ease was lower for anxiety responses than for depression responses, but these differences were not statistically significant. No significant differences were observed in average syllables per word for any of the models. Table 1 shows the means and standard deviations for each readability variable assessed in this study.

	GPT-4		**Gemini 1.5 Flash**		**Llama 3.2**		
Variables	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**P-value**
Word Count	424.6	144.86	182.25	65.69	266.25	43.25	0.0000000002461*
Sentence Count	46.95	18.23	20.85	9.71	25.65	6.49	0.0000000189*
Average Words per Sentence	9.4	1.96	9.38	2.57	10.81	2.35	0.0904
Average Syllables per Word	1.95	0.15	1.98	0.15	2.08	0.13	0.01485*
Flesch-Kincaid Grade Level	11.03	1.93	11.3	1.78	13.1	1.95	0.001768*
Flesch-Kincaid Reading Ease	32.75	12.69	30.66	12.19	20.32	11.74	0.004531*
modified DISCERN	3.2	0.61	3.15	0.81	2.95	0.69	0.5412

Table 1: Mean and Standard Deviation of Readability and Reliability Measures.
*** - P-value < 0.05 and difference is considered statistically significant.**

Text Reliability

GPT-4 had a higher mean modified DISCERN score than Gemini and Llama, but these differences were not statistically significant (Table 1). Although modified DISCERN scores were higher for anxiety prompts compared to depression prompts for all 3 models, these differences were not significant (Figure 1C).

LLMs were able to effectively answer each prompt and cite reliable sources used to generate responses. The most common DISCERN criteria the LLMs failed to meet was addressing areas of uncertainty in their responses. 85% of the responses for each model did not meet this criteria. Listing additional sources of information for patient reference was another criteria that the LLM tools failed to meet adequately. GPT-4 and Gemini cited additional references in 35% of the responses, while Llama did so for only 15%.

Communication Guideline Adherence

Although the relevance of each guideline varied for different topics addressed in the prompts, some common patterns were observed in the responses. Responses employed person-centered and empowering language, and consistently included disclaimers emphasizing the importance of seeking professional help as well. Phrases such as “seeking help is a sign of strength, not a sign of weakness” and “many people with anxiety lead fulfilling lives” were observed across all three models, with particular emphasis on mental health conditions being treatable.

Despite these patterns, however, the LLMs were not able to incorporate non-stigmatizing language consistently throughout all their responses. Stigmatizing phrases appeared in 10% of GPT-4’s responses, 15% of Gemini’s responses, and 20% of Llama’s responses. This was especially evident in phrases describing anxiety and depression, with GPT-4 using the term “impairing”, and Gemini and Llama both describing the conditions as “debilitating.” Information about alcohol and substance use also included the term “substance abuse.”

The differences in the proportion of responses with stigmatizing language were not statistically significant. No significant differences were observed in the use of stigmatizing language in responses related to anxiety and depression across the models (Figure 1D).

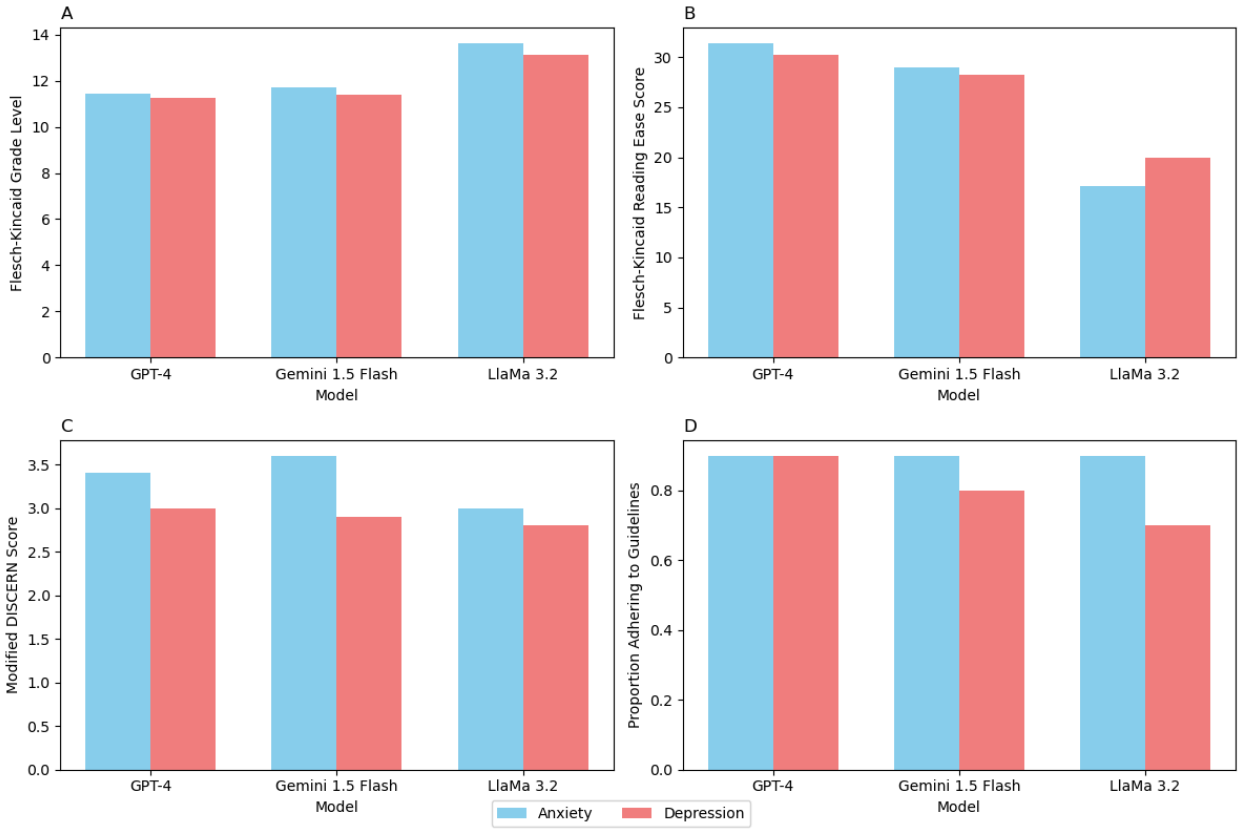

Figure 1: Comparison of GPT-4, Gemini, and Llama in (a) Mean Grade Level, (b) Mean Reading Ease Score, (c) Mean Modified DISCERN Score, and (d) Proportion of Responses Adhering to Communication Guidelines for queries about anxiety and depression.

Discussion

LLMs have the potential to transform population mental health by enhancing early diagnosis, personalizing treatments and interventions, and broadening access to care. It is therefore essential to assess the quality of information provided by these tools to determine whether it is accurate, unbiased, and understandable by diverse populations. This study evaluated the readability and reliability of information provided by GPT-4, Gemini, and Llama in response to common questions about anxiety and depression.

GPT-4 had a significantly higher word count and sentence count compared to both Gemini and Llama. This suggested the model may have been trained to provide longer responses. However, this pattern did not result in GPT-4's responses being more difficult to read or overly complex in structure or language. Responses generated by Llama had a significantly higher grade level and lower reading ease score than GPT-4 and Gemini, indicating that the readability of mental health text generated by Llama had significantly lower readability. Mental health information must be accessible to diverse populations, particularly underserved communities where literacy

rates are low. These findings suggest that GPT-4 may be better suited for mental health applications compared to Gemini or Llama as it generates detailed responses with high readability. However, further training may be required to tailor the information and ensure it is understandable by individuals of different backgrounds and literacy levels.

The information generated by LLMs regarding the cause of anxiety and depression, the chronicity and recurrence of the conditions, and treatment options were verified with scientific journals, indicating the information was generally accurate. However, when asked about common treatments for anxiety and depression, GPT-4 and Llama included alternate therapy such as acupuncture and herbal supplements among the treatment options. However, the existing literature presents mixed evidence regarding their efficacy, suggesting the material presented by these tools may not always be reliable, and consulting expert sources may be necessary [18,19]. LLMs are trained on texts from a wide range of sources, including scientific journals, medical websites, and publicly available content, and they prioritize generating a comprehensive response rather than critically evaluating the strength and quality of evidence. As a result, the information presented may include varying levels of scientific evidence. Further fine-tuning on accurate and relevant data about mental health will be necessary, so the tools provide information that is evidence-based.

The tools often lacked additional resources or references in their responses, which is particularly important in the context of mental health. Individuals in distress benefit greatly from access to external resources they can turn to for further support. Providing such resources is crucial for making mental health education and care more accessible and effective. This suggests that simple queries, such as the ones used in this study, could not generate tailored responses and further model fine-tuning or prompt engineering may be necessary to personalize responses.

The tone and style of the language in the generated responses was evaluated against communication guidelines for mental health. All models emphasized the importance of seeking professional help and incorporated person-centered, empathetic empowering language into their responses by consistently underscoring that anxiety and depression were treatable and there were multiple avenues available for help. However, stigmatizing language was occasionally observed in descriptions of mental health conditions and substance use, indicating that societal biases and stereotypes may be present in the training data for LLMs. This highlights the need of further training to both identity and mitigate such biases and ensure LLMs communicate mental health information in a more inclusive manner.

These findings highlight the opportunities presented by generative AI in the mental health realm. LLMs such as GPT-4 have demonstrated the ability to generate quality mental health information that is accessible, easily understandable, and person-centered. However, critical areas for refinement need to be addressed to ensure their outputs are consistently reliable and unbiased.

Drawbacks in the responses highlight the need to further improve LLMs for mental health applications. Fine-tuning models by providing training data relevant to mental health information offers an avenue to improve LLM performance by specifying the emotional patterns and type of information that best aligns with mental health applications. This technique can focus the responses of LLMs to be more accurate, unbiased, and safe by training it on textual data that is evidence-based and language that de-stigmatizes mental health [20]. Cultural humility and personalization can also be introduced through this mechanism by incorporating mental health information specific to different races, ethnicities, genders, and geographies in the training data. This can help ensure LLM responses are best tailored to the background of the client interacting with it. However, fine-tuning tasks can be computationally expensive and, therefore, may have limited scalability depending on the availability of resources. Prompt engineering offers another avenue to improving LLM task performance that is accompanied by less computational costs, making it more feasible in some cases. Recent studies have shown that prompt engineering has been as effective as fine-tuning to improve LLM performance on medical question-answering [21]. Through techniques such as in-context learning, LLMs can be trained on example queries to learn to generate context-appropriate responses with the desired style, tonality, content, and literacy level [20,22]. In addition to improving accessibility and accuracy, further evaluation of whether LLMs adhere to ethical standards and respect user privacy is crucial before their widespread adoption for mental health applications.

This scope of this study posed limitations that should be considered. Only 20 questions about anxiety and depression were used in this study. This limited set of prompts may not completely capture the variability in LLM performance and cannot be generalized to the broader range of mental health conditions. Additionally, evaluating the LLM responses using standardized measures of readability, reliability, and bias was valuable in quantifying these aspects of the text, but did not consider the variable understanding and reactions people may have in real-world settings depending on their literacy levels and cultural backgrounds. Additionally, the evaluation of bias was constrained to language and common stigmatizing phrases without considering gender, race, ethnicity, and socioeconomic factors that might also impact the quality of responses. Future studies should focus on exploring these areas of bias in LLM generated text and include the perspectives of individuals from diverse cultural and socioeconomic backgrounds when evaluating LLM responses.

Conclusion

This study evaluated the efficacy of LLMs (GPT-4, Gemini, and Llama) in addressing questions about anxiety and depression, and assessed both the readability and reliability while interpreting bias present in the responses. Overall, the models generated quality responses that had varying levels of readability, with GPT-4 and Gemini presenting responses that were most readable and accessible. No significant differences in grade level or reading ease was observed between responses for anxiety and depression for any model, indicating that the models performed consistently across these two mental health conditions.

All three models provided information that was generally accurate; however, some of the responses had limited information about additional resources, raising concerns about the reliability of these tools for direct deployment in mental health applications. Additionally, some responses contained stigmatizing phrases about mental health, highlighting the need to improve the language of LLMs when responding to mental health concerns. Overall, LLMs offer significant opportunities in making mental health information more widely accessible, but it will be critical to ensure they are further trained for mental health purposes so they adhere to ethical standards and propagate accurate and reliable information.

References

1. McGrath JJ, Al-Hamzawi A, Alonso J, et al. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. Lancet Psychiatry. 2023;10(9):668-681. doi:10.1016/S2215-0366(23)00193-1
2. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med. 2018;48(9):1560-1571. doi:10.1017/S0033291717003336
3. Wainberg ML, Scorza P, Shultz JM, et al. Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. Curr Psychiatry Rep. 2017;19(5):28. doi:10.1007/s11920-017-0780-z
4. Sorkin DH, Janio EA, Eikey EV, et al. Rise in Use of Digital Mental Health Tools and Technologies in the United States During the COVID-19 Pandemic: Survey Study. J Med Internet Res. 2021;23(4):e26994. doi:10.2196/26994
5. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj Ment Health Res. 2024;3(1):48. doi:10.1038/s44184-024-00097-4
6. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res. 2024;3(1):12. doi:10.1038/s44184-024-00056-z
7. Otal HT, Stern E, Canbaz MA. LLM-Assisted Crisis Management: Building advanced LLM platforms for effective emergency response and public collaboration. 2024 IEEE Conference on Artificial Intelligence (CAI). 2024 Jun 25;851–9. doi:10.1109/cai59869.2024.00159
8. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications. 2024;11(1). doi:10.1057/s41599-024-03811-x
9. Mental health by the numbers. 2023. (Accessed December 6, 2024, at https://www.nami.org/about-mental-illness/mental-health-by-the-numbers/)
10. Depression FAQs. 2020. (Accessed December 4, 2024 at https://adaa.org/understanding-anxiety/depression/faqs)
11. Frequently asked questions. 2021. (Accessed December 4, 2024, at https://adaa.org/understanding-anxiety/faqs)
12. Symptoms of anxiety and how to know when you need help. 2020. (Accessed December 4, 2024, at https://health.ucdavis.edu/blog/cultivating-health/symptoms-of-anxiety-and-how-to-know-when-you-need-help/2024/08)
13. Ask the expert: Common questions about depression. 2020. (Accessed December 4, 2024, at https://msutoday.msu.edu/news/2020/ask-the-expert-common-questions-about-depression)
14. Anxiety and depression: The most common questions asked by parents. 2023. (Accessed December 4, 2024, at

https://www.seattlechildrens.org/healthy-tides/anxiety-and-depression-the-most-common-questions-asked-by-parents/)
15. Uzun O. Assessment of Reliability and Quality of Videos on Medial Epicondylitis Shared on YouTube. Cureus. 2023;15(4):e37250. doi: 10.7759/cureus.37250
16. Our Words Matter: Guidelines for language use. 2023. (Accessed December 4, 2024, at https://mindframe.org.au/our-words-matter-guidelines-for-language-use)
17. Volkow ND, Gordon JA, Koob GF. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. Neuropsychopharmacology. 2021;46(13):2230-2232. doi: 10.1038/s41386-021-01069-4
18. Smith CA, Armour M, Lee MS, Wang LQ, Hay PJ. Acupuncture for depression. Cochrane Database Syst Rev. 2018;3(3):CD004046. doi: 10.1002/14651858.CD004046.pub4
19. Natural remedies for depression: Are they effective?. 2018. (Accessed December 6, 2024, at https://www.mayoclinic.org/diseases-conditions/depression/expert-answers/natural-remedies-for-depression/faq-20058026)
20. De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. arXiv preprint. 2023. doi:10.48550/arxiv.2311.14693
21. Maharjan J, Garikipati A, Singh NP, Cyrus L, Sharma M, Ciobanu M, Barnes G, Thapa R, Mao Q, Das R. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Sci Rep. 2024;14(1):14156. doi: 10.1038/s41598-024-64827-6
22. Priyadarshana YHPP, Senanayake A, Liang Z, Piumarta I. Prompt engineering for digital mental health: a short review. Front Digit Health. 2024;6:1410947. doi: 10.3389/fdgth.2024.1410947