Models for exploring the credibility of large language models for mental health support – Protocol for a scoping review

Dipak Gautam, Philipp Kellmeyer

Submitted to: JMIR Research Protocols
on: June 03, 2024

Table of Contents

Models for exploring the credibility of large language models for mental health support – Protocol for a scoping review

Dipak Gautam[1] BS; Philipp Kellmeyer[2, 3, 4] Dr med, MD, MPhil

[1]University of Manneim Mannheim DE
[2]School of Business Informatics and Mathematics University of Manneim Mannheim DE
[3]Department of Neurosurgery University of Freiburg - Medical Center Freiburg im Breisgau DE
[4]Institute for Biomedical Ethics and History of Medicine University of Zuricj Zurich CH

Corresponding Author:
Philipp Kellmeyer Dr med, MD, MPhil
School of Business Informatics and Mathematics
University of Manneim
B6, 26
Mannheim
DE

Abstract

Background: The rapid evolution of Large Language Models (LLMs), such as BERT and GPT, has introduced significant advancements in natural language processing. These models are increasingly integrated into various applications, including mental health support. However, the credibility of LLMs in providing reliable and explainable mental health information and support remains underexplored.

Objective: This scoping review aims to systematically explore and map the factors influencing the credibility of LLMs in mental health support. Specifically, the review will assess LLMs' reliability, explainability, and ethical implications in this context.

Methods: The review will follow the PRISMA extension for scoping reviews (PRISMA-ScR) and the Joanna Briggs Institute (JBI) methodology. A comprehensive search will be conducted in databases such as PsycINFO, Medline via PubMed, Web of Science, IEEE Xplore, and ACM Digital Library. Studies published from 2019 onwards in English and peer-reviewed will be included. The Population-Concept-Context (PCC) framework will guide the inclusion criteria. Two independent reviewers will screen and extract data, resolving discrepancies through discussion. Data will be synthesized and presented descriptively.

Results: The review will map the current evidence on the credibility of LLMs in mental health support. It will identify factors influencing the reliability and explainability of these models and discuss ethical considerations for their use. The findings will provide practitioners, researchers, policymakers, and users insights.

Conclusions: This scoping review will fill a critical gap in the literature by systematically examining the credibility of LLMs in mental health support. The results will inform future research, practice, and policy development, ensuring the responsible integration of LLMs in mental health services.

(JMIR Preprints 03/06/2024:62865)
DOI: https://doi.org/10.2196/preprints.62865

Preprint Settings

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

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

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

Original Manuscript

Models for exploring the credibility of large language models for mental health support – Protocol for a scoping review

Abstract

Background: The rapid evolution of Large Language Models (LLMs), such as BERT and GPT, has introduced significant advancements in natural language processing. These models are increasingly integrated into various applications, including mental health support. However, the credibility of LLMs in providing reliable and explainable mental health information and support remains underexplored.

Objective: This scoping review aims to systematically explore and map the factors influencing the credibility of LLMs in mental health support. Specifically, the review will assess LLMs' reliability, explainability, and ethical implications in this context.

Methods: The review will follow the PRISMA extension for scoping reviews (PRISMA-ScR) and the Joanna Briggs Institute (JBI) methodology. A comprehensive search will be conducted in databases such as PsycINFO, Medline via PubMed, Web of Science, IEEE Xplore, and ACM Digital Library. Studies published from 2019 onwards in English and peer-reviewed will be included. The Population-Concept-Context (PCC) framework will guide the inclusion criteria. Two independent reviewers will screen and extract data, resolving discrepancies through discussion. Data will be synthesized and presented descriptively.

Results: The review will map the current evidence on the credibility of LLMs in mental health support. It will identify factors influencing the reliability and explainability of these models and discuss ethical considerations for their use. The findings will provide practitioners, researchers, policymakers, and users insights.

Conclusions: This scoping review will fill a critical gap in the literature by systematically examining the credibility of LLMs in mental health support. The results will inform future research, practice, and policy development, ensuring the responsible integration of LLMs in mental health services.

Introduction

The emergence of generative AI and the rapid evolution of Large Language Models (LLMs) are introducing new complexities and accelerating technological advancements. Our understanding of the inner workings of these systems remains limited. However, there is a widespread rush across all sectors to adopt these technologies, often overlooking the ethical considerations and potential threats to data privacy and confidentiality (Sebastian, 2023). The transformative power of this technology is undeniable, with the potential to revolutionize nearly every industry and facet of human life.

As generative AI becomes increasingly integrated into various tools and applications, its presence in our everyday lives grows. The deployment of AI in smartphones, social media, and platforms like OpenAI's ChatGPT, Google's Gemini, and Meta's Llama is a testament to this trend. These

technological advancements are becoming a staple in our personal and professional spheres, a presence set to expand further. Their influence on our choices is significant and will only intensify in the coming years, potentially impacting our mental health and overall well-being (Nepal et al., 2024).

A recent research published on the alignment of outcomes of LLMs to human intentions (Liu et al., 2024) concluded that, based on publicly available users' opinions about their use of LLMs, in general, users tend to trust those LLMs more that demonstrate higher alignment to human intentions. In a Harvard Business Review (Candelon et al., 2023), the authors of the review "AI can be both accurate and transparent", which was a study to examine the trade-off between accuracy and explainability, tested a wide array of AI models on nearly 100 representative datasets and found that 70% of the time, a more explainable model could be used without compromising accuracy. This suggests that a reliable LLM can be developed and aligned with transparency and social norms. Another study (Mökander & Floridi, 2021) suggested that ethics-based auditing can be a governance mechanism for building and deploying LLMs and potentially bridge the gap between principles and practice in AI ethics. It argues that ethics-based auditing will improve the quality of decision-making, users' satisfaction with privacy and confidentiality at the center, influence laws and policies that govern these systems and minimize human harm.

Given that this technology is still nascent and research into its effects on society is scarce, evaluating its implications and integration into the medical and healthcare sectors, particularly in mental health support, is essential. It is imperative to recognize and scrutinize this progression, understand its benefits (Srikanth et al., 2021) and associated risks, and identify potential measures to prevent and reduce the adverse effects of these technologies on human lives. The ethical dilemmas surrounding the use of these systems, the safeguarding of user privacy and confidentiality, and the accountability of developers remain largely uncharted territories (Xu et al., 2024). It can be said that a lot of the published literature and journals either provide general solutions or look into a specific domain, such as applicability or accuracy.

This exploration seeks to shed light on the current state of LLMs, focusing on their reliability and explainability, especially in providing support for mental health. It will try to add more consistent factors that can account for the credibility of LLMs. It is also worth noting that many studies suggest that there are very few to almost no studies on the credibility of LLMs, particularly in medicine (Gama et al., 2022), (Sharma et al., 2022) and mental health support. This review intends to address this gap.

Objectives and research questions

The study's overall objective is to explore the current state of evidence on the credibility of Large Language Models by comprehensively reviewing research on credibility factors such as reliability and explainability. A secondary objective is to derive insights into ethical implications for the responsible use of LLMs in mental health support.

To this end, the following research questions will be pursued in the scoping review:

- How credible are LLMs in providing mental health information and support?

    - What factors influence the reliability of large-language models for mental health support?
    - How explainable (XAI) are LLMs providing mental health information and support?

- What are the users' perceptions of using LLMs as reliable and explainable sources of mental health support?
- What ethical implications should be considered for the responsible use of LLMs in mental health support?
    - How can we ensure privacy and confidentiality when users interact with LLMs for sensitive mental health issues?
    - How can we ensure that the shared sensitive information by the user is secure and private?

Methods

Design

The PRISMA extension for scoping reviews (PRISMA-ScR) (Tricco et al., 2018) will be used as a basic tool, and the Joanna Briggs Institute's (JBI) approach to scoping reviews will be followed (Peters et al., 2021), (Peters et al., 2020), (Peters et al., 2015).

Research Strategy and Terms

Information Sources

The following e-databases will be included as sources of information: PsycINFO, Medline via PubMed, Web of Science, IEEE Xplore, and ACM Digital Library. Additionally, screening of websites such as Google Scholar, Semantic Scholar, JMIR, Internet Archive Scholar, and Springer will serve as supplementary sources.

Search Strategy

The sources will be searched with a combination of relevant search terms before the scoping review. An iterative approach will be followed to develop the search strategy. At first, search terms used in previous studies and reviews related to the credibility of LLMs in mental health support will be identified. Then, an initial search in Medline via PubMed and ACM Digital Library will be conducted after analyzing text words (title and abstract) and indexed terms, according to Joanna Briggs Institute methodology. The initial search strategy for Medline via PubMed can be found in Appendix 1 below. This approach will use all received search terms in all databases. Afterward, the references for all included contributions will be checked. The terms will be adapted to the essential search particulars like wildcards (*), truncations, and Boolean operators in each electronic database.

For a more precise explanation of the inclusion criteria, the PCC, which stands for Population, Concept, and Context scheme, will be followed. The table below shows the most important criteria based on the PCC framework.

Population	Mental health practitioners, researchers, educators, students, and adults (age group: 18-65)
Concept	Non-participatory, Exploratory, co-creation, co-design
Context	Exploration of credibility (reliability, explainability) of Large Language Models for mental health support
Types of Sources	Secondary sources, electronic databases, studies published in the last five years (2019 to 2024), Studies published in English or have an English translation available. Full-text articles

The kinds of literature and papers that provide information on at least one research question from above will be included. Detailed inclusion and exclusion criteria are written below for each review.

Eligibility Criteria

Papers to be included must meet the following criteria:

1. *Study design*: It should be empirical, exploring the application of Large Language Models (LLMs) in the context of mental health support, including both qualitative and quantitative studies.
2. *Model type*: The study should use LLMs, like GPT-2 or later Models, BERT, or similar. Studies that do not use LLMs or BERT models will be excluded.
3. *Mental health focus*: The study must be related to mental health support. It could include a wide range of mental health conditions and support strategies, like therapy, counseling, self-help strategies, intellectual support strategies, or clinical mental medication.
4. *Credibility assessment*: The study should measure or explore dimensions of credibility, such as accuracy, reliability, explainability, correctness, or user satisfaction.
5. *Publication date*: LLMs are rapidly evolving in their capabilities. Thus, only studies published since 2019 will be considered.[1]
6. *Language*: The study should be published in English or have an English translation.
7. *Peer-review*: The studies should be peer-reviewed to ensure the quality and reliability of the results and findings.

Study Selection

The saved papers will be checked for duplicates. The open-source application Zotero will be used as a bibliographic tool. Articles screening will be done using the open-access tool Rayyan.ai. Two independent reviewers will screen all titles and abstracts separately for inclusion or exclusion. The same will be done for the full-text screening. An apparent reason for excluding a study will be assessed in each of these steps. The search results and the study inclusion or exclusion process will

[1] We chose 2019 as the starting year for including studies in our scoping review because it marks a significant point in the innovation dynamics of large language models (LLMs). The introduction of BERT in late 2018 and the foundational work on Transformer models in 2017 catalyzed a rapid evolution in the field. These developments laid the groundwork for subsequent influential models such as GPT-2, which emerged in 2019. The advancements from 2018 onward have significantly impacted computational research, including potential applications in mental health support, making this an appropriate and justifiable starting point for our review.

be transparently reported in full in the final scoping review, which will be presented in a PRISMA-ScR flowchart.

Data Extraction

We will use the literature review matrix method (Goldman et al., 2004) to organize and chart the data extracted from the included research papers. Two reviewers will chart the data independently, continually update the matrix in an iterative process, and discuss the results while all the changes are detailed in the scoping review. After independent extraction, the two reviewers will compare their extracted data. Any discrepancies or disagreements will be discussed and resolved through consensus. A third reviewer will be consulted to decide if a consensus cannot be reached. The data extraction form will be continually updated in an iterative process as new insights are gained and new studies are reviewed. Reviewers will keep detailed records of any changes to the extraction process and document the rationale. The extracted data will be managed and stored using Zotero (https://www.zotero.org/) for reference management and Rayyan.ai (https://www.rayyan.ai/) for screening and collaboration. A draft extraction form/literature matrix is added below in *Appendix 2*.

Data Analysis and Presentation

The extracted data will be presented logically and descriptively, including diagrams and tables based on the objectives and research questions of the scoping review. Data synthesis and presentation will follow an inductive approach. A summary description and discussion of the findings according to the research questions, the flowchart, and the entire research process will be provided in text form and described narratively.

Results

The analysis and review preparation are planned from May to October 2024. The target date for the submission of the scoping review is 31st October 2024.

Discussion

This scoping review aims to systematically explore and map the factors influencing the credibility of Large Language Models (LLMs) in providing mental health support. By identifying and synthesizing the current evidence on the reliability and explainability of LLMs, this review seeks to offer valuable insights for stakeholders, including mental health practitioners, researchers, policymakers, and end-users. The findings will contribute to understanding how LLMs can be effectively integrated into mental health services while maintaining ethical standards and user trust.

Potential Implications

The results from this review are expected to have several important implications. For mental health practitioners, understanding the credibility of LLMs can inform decisions about integrating these technologies into therapeutic practices. For researchers, the review will highlight gaps in the literature, suggesting areas for future investigation. Policymakers could use the insights to develop guidelines and regulations to ensure the responsible use of LLMs in mental health support. For users, this review aims to provide clarity on the reliability of LLMs as a supplementary resource for mental

health information and support.

Limitations

While scoping reviews offer a comprehensive overview of existing literature, they have inherent limitations. The rapid evolution of LLM technologies may result in newly published studies being overlooked, and the inclusion of diverse study designs can introduce biases, complicating direct comparisons. Additionally, the review will only consider studies published in English, potentially excluding relevant research in other languages. Despite a thorough search strategy, some pertinent sources might not be identified.

Conclusion

This scoping review addresses a critical gap in the current literature by evaluating the credibility of LLMs in mental health support. It aims to be the first to systematically map out the reliability and explainability of LLMs in this context. The findings will provide a foundation for further research and inform the development of practices and policies to leverage LLMs' potential while safeguarding ethical standards and user privacy.

Acknowledgments

This scoping review is a master thesis paper and is part of a student assessment under the supervision of Prof. Dr. med. Philipp Kellmeyer. Data and Web Science Group, School of Business Informatics and Mathematics, University of Mannheim, Germany.

Data Availability

All data collected, used, and analyzed in this scoping review will be included as supplementary files with the scoping review report.

Conflict of Interest

There are not any conflicts of Interest.

Ethics

Since the data used in the review is collected from secondary sources and primary data is not collected, formal ethical approval is not required.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers

GPT: Generative Pre-trained Transformers

JBI: Joanna Briggs Institute

LLM: Large Language Model

PCC: acronym for Population, Concept and Context

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

XAI: Explainable Artificial Intelligence

References

Candelon, F., Evgeniou, T., & Martens, D. (2023, May 12). AI Can Be Both Accurate and Transparent. *Harvard Business Review*. https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent

Gama, F., Tyskbo, D., Nygren, J., Barlow, J., Reed, J., & Svedberg, P. (2022). Implementation Frameworks for Artificial Intelligence Translation Into Health Care Practice: Scoping Review. *Journal of Medical Internet Research*, *24*(1), e32215. https://doi.org/10.2196/32215

Goldman, K. D., & Schmalz, K. J. (2004). The Matrix Method of Literature Reviews. Health Promotion Practice, 5(1), 5-7. https://doi.org/10.1177/1524839903258885

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2024). *Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment* (arXiv:2308.05374). arXiv. http://arxiv.org/abs/2308.05374

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Nepal, S., Pillai, A., Campbell, W., Massachi, T., Choi, E. S., Xu, X., Kuc, J., Huckins, J. F., Holden, J., Depp, C., Jacobson, N., Czerwinski, M. P., Granholm, E., & Campbell, A. (2024). Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3613905.3650767

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare*, *13*(3), 141–146. https://doi.org/10.1097/XEB.0000000000000050

Peters, M. D. J., Marnie, C., Colquhoun, H., Garritty, C. M., Hempel, S., Horsley, T., Langlois, E. V., Lillie, E., O'Brien, K. K., Tunçalp, Özge, Wilson, M. G., Zarin, W., & Tricco, A. C. (2021). Scoping reviews: Reinforcing and advancing the methodology and application. *Systematic Reviews*, *10*(1), 263. https://doi.org/10.1186/s13643-021-01821-3

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*(10), 2119–2126.

https://doi.org/10.11124/JBIES-20-00167

Raymond, L. (2010). *Academic Guides: Common Assignments: Literature Review Matrix*. https://academicguides.waldenu.edu/writingcenter/assignments/literaturereview/matrix

Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4454761

Sharma, M., Savage, C., Nair, M., Larsson, I., Svedberg, P., & Nygren, J. M. (2022). Artificial Intelligence Applications in Health Care Practice: Scoping Review. *Journal of Medical Internet Research*, *24*(10), e40238. https://doi.org/10.2196/40238

Srikanth, T. K., Rao, G. N., Parthasarathy, R., Raj, D., Math, S. B., Mehrotra, S., Tirthahalli, J., Kumar, N. C., Sudhir, P., & Jayarajan, D. (2021). Leveraging technology to improve quality of mental health care in Karnataka. *Companion Publication of the 13th ACM Web Science Conference 2021*, 107–114. https://doi.org/10.1145/3462741.3466652

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–32. https://doi.org/10.1145/3643540

Appendices

1. Search strategy:

Search conducted in Medline via PubMed: 31 May 2024

Foundational Keyword	Items found	Search Query	Notes/Field Tags
Credibility (#1)	2,739,856	"reliable*"[Title/Abstract] OR "rely*"[Title/Abstract] OR "correct*"[Title/Abstract] OR "robust*"[Title/Abstract] OR "explainable*"[Title/Abstract] OR "explain*"[Title/Abstract] OR "interpret*"[Title/Abstract] OR "interpretable*"[Title/Abstract] OR "credible*"[Title/Abstract]	Title/Abstract
Large Language Models (#2)	747,930	"Large Language Models"[Title/Abstract] OR "LLM"[Title/Abstract] OR "LLMs"[Title/Abstract] OR "Artificial Intelligence"[Title/Abstract] OR "AI"[Title/Abstract] OR "Generative Artificial Intelligence"[Title/Abstract] OR "GenAI"[Title/Abstract] OR "GAI"[Title/Abstract] OR "Explainable AI"[Title/Abstract] OR "XAI"[Title/Abstract] OR "Explainable Machine Learning"[Title/Abstract] OR "XML"[Title/Abstract] OR "Machine	Title/Abstract

		Learning"[Title/Abstract] OR "ML"[Title/Abstract] OR "Interpretable AI"[Title/Abstract] OR "chatbot*"[Title/Abstract] OR "Artificial Intelligence"[MeSH Terms] OR "Machine Learning"[MeSH Terms]	
Mental health support (#3)	1,060,031	"Mental health support"[Title/Abstract] OR "mHealth"[Title/Abstract] OR "Mental health Information"[Title/Abstract] OR "mental health"[Title/Abstract] OR "Psychological Counseling"[Title/Abstract] OR "Counseling"[Title/Abstract] OR "Psychotherapy"[Title/Abstract] OR "mental health services"[Title/Abstract] OR "User Perception"[Title/Abstract] OR "User Satisfaction"[Title/Abstract] OR "Patient Perception"[Title/Abstract] OR "Patient Satisfaction"[Title/Abstract] OR "perceive*"[Title/Abstract] OR "satisfy*"[Title/Abstract] OR "mental health"[MeSH Terms] OR "mental health services"[MeSH Terms] OR "Counseling"[MeSH Terms] OR "Psychotherapy"[MeSH Terms] OR "Patient Satisfaction"[MeSH Terms]	Title/Abstract
Merged	2,447	#1 AND #2 AND #3	All Fields
Merged with filters	1,578	#1 AND #2 AND #3, Filters: in the last 5 years	All Fields

2. Literature review matrix:

Author/ Date	Theoretical/ Conceptual Framework	Research Question(s)/ Hypotheses	Methodology	Analysis & Results	Conclusions	Implications for Future research	Implications For practice