Cureus

Open Access
Editorial

DOI: 10.7759/cureus.43023

Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation

Review began 08/01/2023
Review ended 08/04/2023
Published 08/06/2023

Abdulqadir J. Nashwan [1] , Jaber H. Jaradat [2]

1. Nursing Department, Hamad Medical Corporation, Doha, QAT 2. School of Medicine, Mutah University, Al Karak, JOR

Corresponding author: Abdulqadir J. Nashwan, anashwan@hamad.qa

Abstract

This editorial explores the innovative application of large language Models (LLMs) in conducting systematic reviews, specifically focusing on quality assessment and risk-of-bias evaluation. As integral components of systematic reviews, these tasks traditionally require extensive human effort, subjectivity, and time. Integrating LLMs can revolutionize this process, providing an objective, consistent, and rapid methodology for quality assessment and risk-of-bias evaluation. With their ability to comprehend context, predict semantic relationships, and extract relevant information, LLMs can effectively appraise study quality and risk of bias. However, careful consideration must be given to potential risks and limitations associated with over-reliance on machine learning models and inherent biases in training data. An optimal balance and combination between human expertise and automated LLM evaluation might offer the most effective approach to advance and streamline the field of evidence synthesis.

Categories: Healthcare Technology, Epidemiology/Public Health
Keywords: evidence synthesis, machine learning, artificial intelligence, large language models, risk of bias, quality assessment, systematic reviews

Editorial

Large language models (LLMs) present a groundbreaking opportunity to revolutionize the process of conducting systematic reviews, particularly in quality assessment (QA) and risk-of-bias (ROB) appraisal. This lies in the potential of LLMs to automate the inherently labor-intensive and often subjective process of these tasks that may lead to inconsistencies among different assessors.

Systematic reviews conduct rigorous and comprehensive assessments of existing literature on a particular medical topic [1]. They are an essential part of evidence-based medicine and involve several stages, of which the QA of included studies or the appraisal of ROB is a critical stage to ensure the credibility of the review. These reviews critically analyzed the available evidence, identified potential biases or shortcomings in individual studies, and synthesized the findings [1].

LLMs are advanced artificial intelligence (AI) models, such as Generative Pre-trained Transformer 4 (GPT-4), DALL-E, Segment Anything Model (SAM), Large Language Model Meta AI (LLaMA), Language Model for Dialogue Applications (LaMDA), Vision Transformer, etc., capable of automating various tasks in natural language processing, computer vision, etc. GPT-4 is the latest GPT model incorporated in ChatGPT, which is a multimodal foundation model that excels in generating human-like text and can process text and image inputs and beyond [2]. LLMs are trained on extensive text data, enabling them to understand context, predict semantic relationships, and extract relevant information from diverse datasets [2]. Consequently, LLMs can be effectively harnessed to automatically assess the quality of the included studies in a systematic review, streamlining the process and reducing the subjectivity associated with human independent assessors.

Although QA and ROB assessments are considered similar safeguard elements and are sometimes used interchangeably by researchers, their applications and interpretations differ. QA evaluates methodological safeguards within a study to ensure it is well-designed, conducted, analyzed, interpreted, and reported, thereby minimizing systematic errors or bias. Conversely, the ROB assessment, also known as critical appraisal, aims to understand how these safeguards may influence the study results, involving judgments about the level of bias and delving deeper into the implications of methodological safeguards on the study's outcomes. QA entails counting the safeguards present as numerical assessment, whereas a ROB assessment employs an approach to rank studies based on potential bias into low or high-risk studies [3,4]. There are numerous tools to assess QA and ROB; they differ according to the study design, such as the Joanna Briggs Institute (JBI), Assessing the Methodological Quality of Systematic Reviews (AMSTAR), Critical Appraisal Skills Program (CASP), Cochrane Risk of Bias (ROB 1 and 2) tool, Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) tool, the Newcastle-Ottawa Scale (NOS), etc. [4].

How to cite this article
Nashwan A J, Jaradat J H (August 06, 2023) Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation. Cureus 15(8): e43023. DOI 10.7759/cureus.43023

LLMs can be trained to identify common sources of bias, such as selection, performance, detection, attrition, and reporting, by recognizing specific phrases, language patterns, or missing information; the LLM can assign a risk level of bias to each study. This objective, automated assessment can minimize human assessors' subjectivity, enhance the systematic review's reliability and capture nuances that may otherwise have been overlooked [5]. The use of LLMs not only improves the efficiency and consistency of QA and ROB evaluations but also expedites the overall systematic review process. By eliminating the need for manual appraisal, LLMs enable faster evidence synthesis and knowledge dissemination, crucial in developing clinical guidelines and public health emergencies (e.g., pandemics), where timely access to highly-quality synthesized evidence is crucial. However, using LLMs holds many potential risks, limitations, and challenges. Overreliance on LLMs may weaken critical thinking skills in researchers. Risks associated with inherent biases in the training data of LLMs might influence the assessment outcomes [2,5]. Therefore, combining human expertise and automated LLM evaluation might offer the best approach, ensuring both the efficiency of automation and the nuanced understanding of human assessors.

In conclusion, harnessing LLMs in conducting quality assessments and the risk of bias in systematic reviews can be transformative. With proper caution to ensure appropriate use and mitigate potential risks, efficiency, consistency, and objectivity benefits can significantly advance the evidence synthesis field.

Additional Information

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Burns PB, Rohrich RJ, Chung KC: The levels of evidence and their role in evidence-based medicine . Plast Reconstr Surg. 2011, 128:305-10. 10.1097/PRS.0b013e318219c171
2. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al.: Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023, 9:e48291. 10.2196/48291
3. Furuya-Kanamori L, Xu C, Hasan SS, Doi SA: Quality versus risk-of-bias assessment in clinical research . J Clin Epidemiol. 2021, 129:172-5. 10.1016/j.jclinepi.2020.09.044
4. Barker TH, Stone JC, Sears K, et al.: Revising the JBI quantitative critical appraisal tools to improve their applicability: an overview of methods and the development process. JBI Evid Synth. 2023, 21:478-93. 10.11124/JBIES-22-00125
5. Karabacak M, Margetis K: Embracing large language models for medical applications: opportunities and challenges. Cureus. 2023, 15:e39305. 10.7759/cureus.39305