On the difficulties of conducting and replicating systematic literature reviews studies using LLMs in software engineering

Katia Romero Felizardo, Anderson Deizepe
Universidade Tecnológica Federal do Paraná (UTFPR)
Departament of Computing
Cornélio Procópio, Brazil
katiascannavino@utfpr.edu.br
andersondeizepe@alunos.utfpr.edu.br

Daniel Coutinho
Pontifical Catholic University of Rio de Janeiro (PUC–RIO)
Informatics Department
Rio de Janeiro, Brazil
dcoutinho@inf.puc-rio.br

Genildo Gomes, Maria Meireles
Federal University of Amazonas (UFAM)
Computing Institute – ICOMP
Amazonas, Brazil
genildo.gomes, maria.meireles@icomp.ufam.edu.br

Marco Gerosa, Igor Steinmacher
Northern Arizona University (NAU)
School of Informatics, Computing, and Cyber Systems
Flagstaff, Arizona
marco.gerosa, igor.steinmacher@nau.edu

***Abstract*—The Software Engineering (SE) community has adopted Systematic Literature Reviews (SLRs) to summarize the state-of-the-art in specific research topics. SLRs offer benefits such as synthesizing evidence from diverse studies to generate auditable results following a reproducible approach, and identifying research gaps for future exploration. However, the process is effort-intensive, prone to errors, and lays various challenges during their conduction. To overcome some of these issues, there is a growing belief that Large Language Models (LLMs) can support systematic literature reviews. While the literature has shown promising results in social sciences, more evidence of its accuracy is needed in technical fields like SE. In this context, studies and replications are essential in verifying the benefits and drawbacks of applying LLMs in systematic literature reviews. This paper discusses the difficulties in conducting and replicating studies that adopt LLMs to support systematic literature in SE. As an implication, we identified the challenges of adopting LLM in SLRs and offered a list of open issues for future research.**

***Index Terms*—LLM, AI, systematic literature review, SLR, difficulties, replication**

I. Introduction

Systematic Literature Reviews (SLRs) gather information from various primary studies to produce auditable results and identify research gaps and perspectives for future research. The SLR process is time and effort-consuming [1]. Selecting studies can be arduous when an SLR involves a large volume of studies [2]; extracting evidence is defiance since SE researchers must examine each study to identify the appropriate information [3] and synthesizing data is also not a trivial activity due to the heterogeneity of evidence [4]. Due to these issues, tool support is essential.

Large Language Models (LLMs) are advanced Artificial Intelligence models designed to understand human inputs and generate human-like text, enabling them to perform various tasks, such as image and video generation, translation, or even data analysis [5]. By demonstrating remarkable success in various applications [6], [7], LLMs present a promising avenue for assisting humans, especially in tasks involving understanding, synthesizing, and generating human language [8]. Software Engineering (SE) researchers have employed LLMs to support SLR [9]–[12]. Using LLMs to support SLRs in SE offers several advantages, including scalability. LLMs can rapidly process large volumes of data, aiding researchers in identifying relevant studies, summarizing findings, and extracting insights, thereby reducing the manual effort required in traditional SLRs. Given these implications, it is important to assess the current strengths and limitations of AI's capacity to support different stages of SLR conduction, as it has shown promising results in various natural language processing tasks [13]. LLMs' capabilities (and limitations) in automating the study selection phase of the SLR process have been acknowledged in multiple studies, particularly in the medical field [14]. In SE research, a few emerging studies have begun to focus on the initial selection stage—analyzing titles, keywords, and abstracts [9], [10]. Conducting and replicating SLRs involving LLMs introduces new challenges that have yet to be thoroughly examined. Therefore, further studies are needed to advance the scientific understanding of LLM support in the SE SLR process. Additionally, replication studies are crucial to validate initial findings, ensuring the robustness and reliability of LLM applications in SLRs while identifying potential limitations and areas for improvement.

This paper aims to bring up the difficulties in conducting and replicating SLR studies involving LLMs. To this end, we investigate the research question: ***What are the difficulties in conducting and replicating SLR studies using LLMs?***

II. Method

We searched and selected articles related to SLR conduction with LLM support (in the SE field) to investigate the difficulties in conducting and replicating them. The search was conducted in September 2024 using the search string: *((“systematic literature review” OR “systematic review” OR SLR) AND (“large language model” OR LLM OR ChatGPT or GPT OR Llama OR Gemini) AND (“software engineering”) AND (LIMIT-TO (SUBJAREA “COMP”)).*

We collected 89 articles via Scopus, a representative digital library for the SE field [15]. We then used the following Inclusion Criteria (IC) to filter the studies: **IC_1.** studies that used LLMs to support the SLR process (reducing the sample to 21 candidate studies); **IC_2.** studies within the SE context (2 studies included – (Study 1 [9] and Study 2 [10])).

After analyzing whether the papers met the inclusion criteria, one author extracted, and two others verified the collected data. Our goal was to gather the following information: the context of the study and its objective; the SLR activity supported by the LLM; the study design to understand how the LLM was applied to support the SLR process; the LLM model, including its parameters and prompts (essential for characterizing the LLM and its replicability); the metrics used to evaluate the LLM’s performance; and the sample size, including the number of studies included and excluded. Both articles were included. We describe them in the following.

Study 1 [9] explored how LLMs can accelerate title-abstract screening through two perspectives: simplifying abstracts for human researchers and automating title-abstract screening entirely. They performed a study where human researchers performed title-abstract screening for 20 articles with original and simplified abstracts from a prior SLR. The study replicated human researchers’ tasks using GPT–3.5-turbo-0613, GPT–3.5-turbo-16k-0613, and GPT-4–turbo-0613, and explored different prompting techniques, such as Zero-shot, One-shot, Few-shot, Chain-of-Thought, Few-Shot with Chain-of-Thought. Redesigning the prompts was also the focus of the investigation. Precision and Recall were the metrics adopted.

Study 2 [10] explores the impact of prompt variations in the initial stage of the study selection activity — filtering based on the title, abstract, and keywords. Six variant prompts (simple, simpleX, positive, positiveX, balanced, balancedX) were analyzed using five SLR datasets. Each simple, positive, balanced prompt includes an additional variant where exclusion criteria are explicitly stated (denoted by the suffix “X”). Simple prompts do not include any shots, positive prompts use only positive shots, and balanced prompts use an equal number of positive and negative shots. The number of shots was limited to two positives and two negatives to ensure conciseness. The five real SLR datasets totalized 5144 studies (SLR1: 205 + SLR2: 292 + SLR3: 875 + SLR4: 1089 + SLR5: 2683). The data from this study, comprising 463 included and 4681 excluded studies, was openly accessible, and the selection criteria were meticulously detailed. The metrics adopted were Precision, Recall, Negative Predictive Value, Specificity, Work Saved over Sampling, Balanced Accuracy, and Matthews Correlation Coefficient (MCC).

III. Discussing difficulties of conducting and replicating SLR studies with LLMs

This section discusses the difficulties in conducting and replicating SLR studies using LLMs.

1. Prompt sensitivity. LLMs are responsive to even slight variations in prompt formatting. For instance, differences between the prompts of Study 1 [9] and 2 [10] include the description (or not) of the context (*e.g., I am screening articles for a systematic literature review* — only in Study 2 [10]) and the topic of the SLR (*e.g., The topic of the systematic review is <TOPIC>* — only in Study 2 [10]). The details of the task provided to the LLM also varied. For example, in Study 1 [9], the prompt mentioned only the task, *e.g., “Your task is to include or exclude a research article based on the inclusion criteria. If none of the inclusion criteria applies, the article is excluded.”* In Study 2 [10], the prompt, in addition to the task, detailed how to perform it and the response format, *e.g., “Decide if the following article should be included or excluded from the systematic review. Only answer Include or Exclude. Be lenient. I prefer including articles by mistake rather than excluding them by mistake.”* In both studies, specific terminology from the SLR area was used, and the task was explicitly described, avoiding unnecessary complexity. However, the prompt of Study 2 [10] was enriched with details about how to perform the task, and the contextualization of SLR was mentioned. Specific constraints were imposed during the selection activity to maintain consistency and control. These constraints encompassed classifying articles exclusively as included or excluded.

How to write prompts to support SLR is a challenge, and the impacts of variations in prompt wording are unknown. For example, can the sentence (*e.g., “Be lenient. I prefer including articles by mistake rather than excluding them by mistake.”*) change the LLM classification to avoid the exclusion of relevant studies? Another concern would be about the behavior of LLMs if they were requested to classify articles using a rate of agreement for studies inclusion following a five or seven-point Likert scale to provide a more granular understanding of the LLM agreement related to the selection criteria. Given the fragility that prompts impose, future research should focus on understanding the factors that influence prompt performance in LLMs for SLRs, such as variants of prompt writing, to propose a prompt template replicable to different SLRs. Better ways of evaluating prompts are also needed.

2. The inherent randomness of LLMs. By design, when users utilize the default parameters, the LLM will provide different answers to the same prompt, and randomness prevents the replicability of SLRs. While this can be mitigated in some tasks (e.g., classification) by changing some parameters (*e.g.,* temperature, top-p, etc.), it is a problem that is not entirely fixable in the currently available LLMs for all tasks, since tasks that require creativity might perform worse when this randomness is removed. For example, Study 1 [9] set the

temperature to 0 to control/reduce the response creativity (or variability). However, more (focused) studies are required to understand better configuration settings of LLMs to explain the advantages and disadvantages of changing these settings and how they impact LLM outputs in each phase of the SLR process.

3. Limited information provided to LLM to classify studies. If a researcher is unsure about excluding an article after reading its title and abstract, Kitchenham et al. [16] recommend reading other sections of the paper. In both studies analyzed [9], [10], the information provided to LLM was only title-abstract and keywords. We believe that to improve the accuracy of LLM classifications, SE researchers could supplement their understanding by incorporating additional information from the candidate studies, for example, prompt relevant parts of the articles, such as the results and conclusions sections, since these sections often provided more detailed and context-rich information compared to abstracts alone.

4. Application of exclusion criteria boundaries. Another issue related to information limitation (as discussed in 3.) is that applying some exclusion criteria also requires information not available in the title-abstract-keywords, for example, publication date, type of study (*e.g.,* grey literature), or the language in which it was written. Collecting this information to make it available to LLM requires additional research effort. Another selection criterion typically adopted in SLRs is excluding previous versions of a duplicate article. In this sense, LLM currently lacks making connections across articles. In the two studies analyzed [9], [10], this type of exclusion criteria was not applied. Therefore, finding ways to support LLMs in identifying these duplicates is an open challenge.

5. Burden to choose few-shot examples. In the context of SLRs (study selection), a shot consists of the information about one article and the expected decision based on the selection criteria application (a 2-tuple article, decision). Therefore, positive shots are articles that should be included, and negative shots are the ones that should be excluded—a decision point to select helpful shots. For Study 2 [10], the authors selected the positive shots considering the list of articles used to calibrate the search string (oracle); this list was defined during the planning stage of the SLR underneath evaluation. However, they mentioned that the negative shots were more challenging to obtain since articles clearly outside the scope of the SLR do not help distinguish relevant from irrelevant articles. A helpful negative shot is ambiguous relevance or unclear exclusion criteria application, which can lead to conflicting judgments among reviewers. However, choosing a helpful negative shot involves reading articles manually, which increases the time and effort required from SLR researchers. To propose methods and techniques to support negative shots is a topic for further research.

6. LLMs are "black boxes" – lack of transparency. LLMs do not explain why a particular article was classified as included or excluded, operating as "black boxes." In none of the studies [9], [10] interpretability mechanisms were inserted, such as fact-checking [17] or argument mining [18], to improve the LLM-based study selection. Moreover, it is unclear how "contamination" from training data can affect the outcomes, given that the training set is unknown. How to deal with these challenges related to lack of transparency is still open for future research.

7. LLM configuration may not necessarily be available over time. The parameters applied for Study 1 [9] were GPT-3.5-turbo-0613, GPT-3.5-turbo-16k-0613, GPT-4-turbo-0613. For Study 2 [10] the parameters used were GPT-3.5-turbo-0613 (t (0.0), seed (128), context (4K), cut-off date (June 13, 2023)); ii) gpt-3.5-turbo-16k-0613 (t (0.0), seed (128), context (16K), cut-off date (June 13, 2023)); and GPT-4-turbo-0613 (t (0.0), seed (128), context (8K), cut-off date (June 13, 2023)). With the rapid improvements that have been happening to LLMs, despite current availability, there is no guarantee that the used models will be available in the future. This is especially true for the models utilized in the two SLRs, given that they are closed-source and cloud-based. In articles from other SE topics, this situation is even worse, as some studies directly utilize ChatGPT (*i.e.*, the website) instead of utilizing the GPT models directly (*i.e.*, via API). By doing that, they forgo control over the parameters and version of the model that they are utilizing. For example, GPT 3.5 Turbo is no longer available when utilizing ChatGPT.

8. High costs for conduction and replication. The current state-of-the-art LLMs can be pretty expensive to run. For instance, GPT 4, which Study 1 [9] utilizes, currently costs 10 USD/1M tokens and 30 USD/1M tokens for input and output, respectively. While this can be mitigated by executing new studies that utilize smaller models, researchers would have to sacrifice better results for lower costs. Even open-source LLMs (not used by any of the two studies), which can be run locally, have high hardware costs. This creates a problem in which it can become prohibitively expensive to replicate studies utilizing LLMs, especially as the amount of data scales. Study 2 [10], for example, ran six variations of prompts to evaluate 5,144 studies. Adding another dimension to this experiment (like trying different models) can significantly increase costs.

9. Limitation of studies for supporting the SLR process as a whole. The two studies using LLM for SLR in SE [9], [10] focused on the study selection stage [9], [10], analyzing titles, keywords, and abstracts. However, according to Waseem et al. [19], the impacts of LLMs on SLRs include other SLR phases, from generating research questions and keywords for search strings to their employment in extracting and synthesizing information. Alshami et al. [20] broadly investigated the potential of LLMs through all SLR activities. The findings exposed that LLMs can help generate research questions and suggest boolean research terms, but LLMs are restricted to data extraction. Relying on assistance for generating search strings, Wang and his coauthors [21] emphasize that LLMs can follow complex instructions, making them a valuable tool for facilitating researchers performing SLR. Other researchers found that LLMs helped enhance the overall quality of search strings by providing synonyms [20], [21]. Despite LLMs

capabilities (and limitations) in (semi–) automating the SLR process have been recognized in several studies, they were performed in the medical field [19], [22]. We encourage the SE community to investigate the LLM application in the SLR context to push the state of the art forward and provide more and more evidence of its adoption.

10. Scarcity of an SLR data repository. In SE, unlike medicine, we do not have a Systematic Review Data Repository. Therefore, researchers from Studies 1 [9] and 2 [10] used replication packages to share data among researchers conducting and replicating studies with LLM to support SLRs. The development of a dedicated SLR repository for SE is critical and urgent. A central database of SLR data can be updated and augmented continuously. Another advantage is that an open repository containing data from multiple SLRs can increase transparency and trust in the review process and provide comparison and reuse data in conducting and updating SLRs. This is a call for the community, which should be more committed to reproducibility, providing detailed, traceable, and well-documented replication packages, openly available. This is the first step not only to assess the LLMs' potential, but to enable better-conducted replications and SLR updates.

IV. Final Remarks

The advancement of fundamental AI methods, particularly the latest LLMs, provides opportunities for AI-based SLRs. However, significant concerns exist about AI applications to support the SLR process, including adapting prompts and evaluating its use in the SE domain. This study provides an overview of the difficulties of conducting and replicating studies adopting LLMs to aid SLR. Our contributions include listing the difficulties in validating the use of LLM in the SLR context, highlighting the challenges in validating the application of LLM in selecting studies, offering valuable insights into the current research landscape, and suggesting future directions for research in the SE domain.

Our research is limited to one empirical study type, SLR, and two articles [9], [10]. Considering these two studies, we discussed difficulties in conducting and replicating studies using LLMs. Therefore, representativeness is limited, so results may not represent the full spectrum of the SE domain. Exploring other SLR activities and more studies is necessary and should be considered for future research.

Finally, we believe SLR processes will incorporate AI tools to support SE researchers in planning and conducting their reviews. The responsible implementation of LLMs in SE offers a promising opportunity to enhance SLR conduction and replication. However, critically evaluating advantages and weaknesses is key for appropriately integrating LLMs in SE.

References

[1] A. Natukunda and L. Muchene, "Unsupervised title and abstract screening for systematic review: a retrospective case-study using topic modelling methodology," *Systematic Review*, vol. 3, pp. 1–16, 2023.

[2] K. R. Felizardo, E. Y. Nakagawa, S. G. MacDonell, and J. C. Maldonado, "A visual analysis approach to update systematic reviews," in *18^{th} International Conference on Evaluation and Assessment in Software Engineering (EASE' 14)*, pp. 1–10, 2014.

[3] M. Riaz, M. Sulayman, N. Salleh, and E. Mendes, "Experiences conducting systematic reviews from novices' perspective," in *14^{th} Conference on Evaluation and Assessment in Software Engineering (EASE'10)*, pp. 44–53, 2010.

[4] D. S. Cruzes and T. Dybå, "Synthesizing evidence in software engineering research," in *ACM-IEEE Symposium on Empirical Software Engineering and Measurement (ESEM'10)*, pp. 1–10, 2010.

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[6] A. Robinson, W. Thorne, B. P. Wu, A. Pandor, M. Essat, M. Stevenson, and X. Song, 2023. arXiv:2308.06610.

[7] D. Wilkins, 2023. https://arxiv.org/pdf/2311.07918.

[8] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2024.

[9] A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, "The promise and challenges of using llms to accelerate the screening process of systematic reviews," in *28^{th} International Conference on Evaluation and Assessment in Software Engineering (EASE'24)*, (Salerno, Italy), pp. 1–10, ACM, 2024.

[10] E. Syriani, I. David, and G. Kumar, "Screening articles for systematic reviews with chatgpt," *Journal of Computer Languages*, vol. August 2024, no. 80, p. 101287, 2024.

[11] K. R. Felizardo, M. S. Lima, A. Deizepe, T. Conte, and I. Steinmacher, "Chatgpt application in systematic literature reviews in software engineering: an evaluation of its accuracy to support the selection activity," in *18^{th} International Symposium on Empirical Software Engineering and Measurement (ESEM'24)*, (Barcelona, Spain), pp. 1–10, ACM, 2024.

[12] K. R. Felizardo, M. S. Lima, A. Deizepe, T. Conte, I. Steinmacher, and M. P. Barcellos, "Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering," in *18^{th} International Symposium on Empirical Software Engineering and Measurement (ESEM'24) – Emerging Results, Vision and Reflection Papers Track*, (Barcelona, Spain), pp. 1–10, ACM, 2024.

[13] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," in *2023 Systems and Information Engineering Design Symposium (SIEDS)*, vol. 1, pp. 274–279, 2023.

[14] S. Wang, H. Scells, S. Zhuang, M. Potthast, B. Koopman, and Zuccon, 2023. arXiv:2401.06320.

[15] C. Wohlin, M. Kalinowski, K. R. Felizardo, and E. Mendes, "Successful combination of database search and snowballing for identification of primary studies in systematic literature studies," *Information and Software Technology*, vol. 147, p. 106908, 2022.

[16] B. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC, 2015.

[17] J. Vladika and F. Matthes, "Scientific fact-checking: A survey of resources and approaches," in *Annual Meeting of the Association for Computational Linguistics*, 2023.

[18] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 45, pp. 765–818, 2020.

[19] M. Waseem, A. Ahmady, P. Liangz, M. Fehmidehx, P. Abrahamsson, and T. Mikkonen, "Conducting systematic literature reviews with chatgpt," in *17^{th} International Symposium on Empirical Software Engineering and Measurement (ESEM'23)*, pp. 1–10, 2023.

[20] A. Alshami, M. Elsayed, E. Ali, A. E. E. Eltoukhy, and T. Zayed, "Harnessing the power of chatgpt for automating systematic review process: Methodology, case study, limitations, and future directions," *Systems*, vol. 11, no. 7, 2023.

[21] S. Wang, H. Scells, B. Koopman, and G. Zuccon, "Can chatgpt write a good boolean query for systematic review literature search?," in *46^{th} International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, pp. 1426—-1436, 2023.

[22] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield, "Can large language models replace humans in systematic reviews? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages," *Research Synthesis Methods*, vol. 1, no. 1, pp. 1–11, 2024.