Towards the use of LLMs to generate search strings in Systematic Literature Review in Software Engineering: How the medical field is studying the uses of LLMs to elaborate search strings

Leonardo Tironi
Federal Institute of Paraná (IFPR)
Londrina, Brazil
leotironineto@gmail.com

Anderson Yoshiaki Iwazaki
University of São Paulo (ICMC/USP)
São Paulo, Brazil
iwazaki.anderson@usp.br

Abstract

A Systematic Literature Review (SLR) is a secondary study method that aggregates, evaluates, and interprets evidence relevant to a specific research topic. One of the key activities involved in this method is creating a search string to identify relevant studies across various databases. However, developing an effective search string is a complex task. The main challenges include the difficulty of mapping relevant terms and synonyms. A poorly formulated search string can either yield a large volume of irrelevant studies or result in the omission of important evidence. In the field of medicine, studies have addressed this challenge using Large Language Models (LLMs). These models have shown disruptive potential for automating tasks within the SLR process, particularly in formulating complex search queries specific to medicine. Although the creation of search strings has also been examined in software engineering, existing studies often lack strong theoretical foundations from the medical perspective. This paper aims to explore how the medical field is utilizing LLMs to develop search strings and to assess the potential application of this technology in software engineering. To achieve this goal, we conducted a quasi-systematic mapping study involving one round of forward and backward snowballing, starting with a single seed paper to analyze the approaches and models used. Our findings indicate that LLMs, including both proprietary models and fine-tuned Open-Source models, have been applied in medicine through various techniques, such as prompt engineering in Zero-Shot or Few-Shot scenarios, as well as fine-tuning. While these models can accelerate the formulation of search strings, researchers are still essential to ensure the quality of the queries. The application of LLMs for search string formulation in medicine is a promising area of research, and a transition to software engineering is feasible. However, this transition requires specific adaptations to create prompts tailored to Software Engineering and search strings for effective model fine-tuning. This study provides a foundation for researchers to investigate how LLMs are utilized in SLR within medicine and the potential for their application in software engineering.

Keywords

Large Language Models, Software Engineering, search strings, Secondary Studies, Medicine

1 Introduction

The number of empirical research studies in Software (SE) is constantly expanding, making it challenging for researchers to keep up to date [3]. In this scenario, Secondary Studies (SS) emerge as an alternative to compile, evaluate, and interpret the resources created by the primary studies (also known as experimental, quasi-experimental, controlled observational studies, or others) for a specific question or topic [9]. As a specific category of SS, the Systematic Literature Review (SLR) employs a rigorous methodology to synthesize current evidence, identify research gaps, and mitigate for the SE community (academic and industry) [9, 12].

The SLR process can be separated into three main steps: Planning, Execution, and Reporting [5, 9]. During the planning phase, the main objective is the development of the research protocol, which includes the research objective, research questions, database selection, search process, and the formulation of the search string [5, 9, 12].

In the Planning phase, formulating a search string is one of the most critical steps [3, 5]. This step involves creating a set of keywords and relevant synonyms for the SLR to find studies in the databases [9]. Researchers often encounter difficulties in identifying all important terms and synonyms, particularly due to variations in language across different papers, where the same concept may be referred to by different names 80% of the time [15] and, as a consequence, an incomplete search string can compromise the quality of the SLR by either overlooking relevant studies or retrieving a significant amount of irrelevant ones.

To address this challenge, researchers have been developing AI tools aimed at creating effective search strings [3]. Some of the methods used include Term Frequency-Inverse Document Frequency, Continuous Bag of Words, and Skip-Gram [3]. However, none of these methods uses LLMs to generate search strings. In the medical field, studies have begun utilizing Large Language Models (LLMs) as alternatives for generating search strings [1]. The use of LLMs can streamline the process of creating search strings by giving new tools for query generation, enabling researchers to leverage either General Purpose LLMs or fine-tuned LLMs to draft queries.

Although the literature in medicine demonstrates the benefits of using LLMs to create search strings, research on their application for generating search strings in Software Engineering remains largely unexplored. In this context, this paper aims to examine how LLMs can be used to generate search strings in the medical field and to investigate the potential for applying LLMs in the Software Engineering domain as well. To obtain an overview of the use of LLMs in medicine, a quasi-systematic mapping was conducted. This process was conducted based on the well-established guidelines defined by Kitchenham *et al.* [10].

Our results provide a detailed overview of the current state-of-the-art in applying LLMs for generating search strings in the field of medicine. We found that while LLMs can significantly accelerate

the formulation of queries, they often exhibit low recall rates and are prone to producing inaccuracies or "hallucinations". As a result, the literature strongly recommends that LLMs serve as assistants rather than replacements for human researchers. It is essential for researchers to validate and refine the generated search strings during the process. Based on these findings, we also discuss the implications, limitations, and potential roadmap for implementing LLM-assisted techniques in SLR within the context of Software Engineering.

The rest of the paper is organized as follows: Section 2 presents the background and related work concerning search string formulation and the use of Large Language Models in research methodology. Section 3 details the methodology applied for the quasi-systematic mapping of the medical literature. Section 4 presents the results obtained from the selected studies. Section 5 presents the discussion and discusses how these findings can be translated to the Software Engineering domain. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Background Related Work

To understand how LLMs are used to generate search strings for SLRs, it is important to know what an SLR is, how AI has been used in SLRs, the challenges of building search strings, and the possibility of automating this process.

2.1 Systematic Literature Review

The objective of a SLR is to evaluate and interpret all available research relevant to a given topic or question [9]. The studies used to conduct an SLR are called primary studies. [9]. The most common reasons to write a systematic review are to summarize the current evidence, to identify gaps in current research and to provide a background for new research [9]. The process to formulate a search string is a complex step in the planning phase [3]. A badly formulated search string can compromise the research by missing relevant evidence or retrieving a large number of irrelevant studies [3].

2.2 Large Language Models

Artificial intelligence is a large field of knowledge in the Software Engineering domain. Adapting language for computers has been a field of study inside the Artificial Intelligence domain for decades and the language modeling advanced quickly in the last years [2]. The development of LLMs went through 4 main phases. (1) Statistical Language Model is a model developed in the field of Natural Language Processing. (2) Neural Language Models introduced the use of vectors to represent words or tokens. (3) Pre-trained Language Models created a new pre-training phase with unlabeled texts to learn the language structures and a fine-tuning phase for specific tasks. Finally (4) LLMs are the Language Models, usually with more with more than 100 billion parameters trained with a large dataset with hundreds of Gigabytes of data [2].

LLMs can be categorized into Proprietary Models, such as ChatGPT-4 and Claude-Sonnet, or Open-Source models such as Mistral and Llama [19]. While the Proprietary Models can provide high capabilities, the researcher control is limited and this type of model lacks transparency [13]. The Open-Source models offer researchers the control of the code, weights of the model and the possibility of fine-tuning the model.

To test and verify the capabilities of the models, an input, called a prompt, is required. The prompt can be formulated to give context to the model and generate more reliable outputs utilizing Zero-Shot, Few-Shot, Chain-of-Thought (CoT) and Multi-prompt strategies [4, 16–19].

2.3 Artificial Intelligence in the Systematic Literature Review

To reduce the necessary work and time consumed during the SLR, artificial intelligence tools have been developed to help researchers conduct a SLR. Some uses are for the study selection process, where text mining and machine learning have helped researchers with abstract screening [11].

Some tools support search string construction, such as the Term Frequency-Inverse Document Frequency, Continuous Bag of Words and Skip-Gram to suggest keywords [3]. In recent papers in the medical field, researchers tested the use of LLMs to generate search strings [1, 4, 13, 17–19].

2.4 Related Work

The use of LLMs in medical research is growing. Beyond the studies formally selected through our mapping, which are presented and analyzed in Section 4, other works examine the adoption of LLMs across different stages of SLRs in medicine and provide context for our investigation [6–8, 14].

Huesing *et al.* [8] described various applications of LLMs in research. One application involved generating and refining search queries through prompts and seed studies. Their findings indicated that the models still faced issues with hallucination and required human verification at each step of the process.

Gartlehner *et al.* [6] discussed the responsible use of AI and LLMs in SLRs, arguing that these technologies present opportunities to assist researchers in certain steps of SLRs in the medical field. They emphasized that LLMs can not perform complete steps independently or replace human researchers; instead, they should serve as tools to reduce the time needed to conduct secondary studies.

Simmons *et al.* [14] tested the use of LLaMa 3-70B to extract 173 data fields from 24 articles using a Retrieval-Augmented Generation (RAG) approach. Their results showed that 68% of data field extractions were considered acceptable based on the interpretations of two independent human reviewers who reached a consensus, with a third reviewer resolving any disagreements.

The study by Hu and Geng [7] explored the use of LLMs across different stages of SLRs, including literature search, title and abstract screening, full-text screening, and data extraction. The agreement overlap between human researchers and ChatGPT ranged from 20% to 68%.

Unlike these works, which address the broader role of LLMs across multiple SLR activities, our study focuses specifically on how have LLMs been used to generate search strings in medicine and analyzes this evidence through a quasi-systematic mapping to assess its transition to Software Engineering.

3 Methods

To understand the adoption of LLMs in the field of medicine, a quasi-systematic mapping study was undertaken. This process followed the guidelines established by Kitchenham *et. al.* [10], while the specific search strategy was guided by the snowballing procedures proposed by Wohlin *et. al.* [20].

3.1 Planning

This study aims to understand the current usage of LLMs in the medical field. It focuses on evaluating the application of LLMs in SLR, specifically in the process of developing search strings. To achieve this goal, two research questions (RQ) were formulated:

- **RQ1**: "How have LLMs been studied to generate search strings in the medical field?"
  Rationale: This question seeks to investigate how the medical field has adopted LLMs for generating search strings, while assessing the benefits and limitations of these models.
- **RQ2**: "Which LLMs have been utilized in these studies?"
  Rationale: This question aims to identify the most commonly used LLMs in the medical field for generating search strings and provide an overview of their usage.

3.2 Study Selection

To identify relevant studies, this research utilized a snowballing search strategy, which included both forward and backward iterations [20]. To ensure we selected the most pertinent studies, we chose the study by Adam *et al.* [1], titled "*Literature Search Sandbox: A Large Language Model that Generates Search Queries for Systematic Reviews*" as our seed study.

This quasi-systematic mapping was conducted using only one round of forward and backward snowballing. However, the studies found by this strategy were relevant to verify the main uses of LLMs in search string generation in the medical field.

3.3 Selection Criteria

To identify relevant studies, we developed specific selection criteria. These criteria include one Inclusion Criterion (IC) and three Exclusion Criteria (EC), as outlined in Table 1. It is important to note that these criteria were applied during the selection process to exclude studies that are not pertinent to our RQ while including those that are relevant.

Table 1: Inclusion and Exclusion Criteria

ID	Description
IC1	The study analyses the uses of LLMs in SLR in medical field
EC1	The study does not analyses the uses of LLMs in SLR in medical field
EC2	The study is not written in Portuguese or English
EC3	The study is not openly accessible or available through institutional access

3.4 Data Extraction

To address our RQ, we identified three Data Extraction (DE) categories to gather the essential information from the studies (see Table 2). The first DE focused on the use of LLMs in developing search strings within the medical field. The analysis centered on three key aspects: (i) studies that utilized LLMs to create the search string; (ii) studies that partially relied on LLMs for this purpose; and (iii) studies that did not use LLMs at all in the development of the search string.

During the second DE, we examined whether the studies included the search string developed using the LLMs. To clarify this data extraction process, we categorized the findings into three main responses: (i) Yes — studies that presented the search string developed using the LLMs; (ii) Partially — studies that partially presented the search string developed using the LLMs; and (iii) No — studies that did not present the search string developed using the LLMs.

Finally, during DE3, we analyzed the main LLMs used for developing the search strings. We found the models used by these studies and they can be divided into two main groups: (i) Proprietary Models – studies that utilized Proprietary Models such as ChatGPT and Claude-Sonnet for search string elaboration; (ii) Open Source Models -— studies that employed the Open Source models such as Llama and Mistral for search string development.

3.5 Conducting

Figure 1 illustrates the three main steps undertaken during the conducting phase. In the first step, a snowballing technique was used to find a total of 81 studies. After removing duplicates, 80 studies remained for the next stage, including the seed paper. In the second step, we reviewed the titles and abstracts of these studies and applied the IC/EC, resulting in the selection of 12 studies. Finally, in the last step, we conducted a thorough reading of the texts and identified seven studies that were deemed relevant to our RQs. It is important to note that all of the studies included are from 2023, 2024 or 2025.

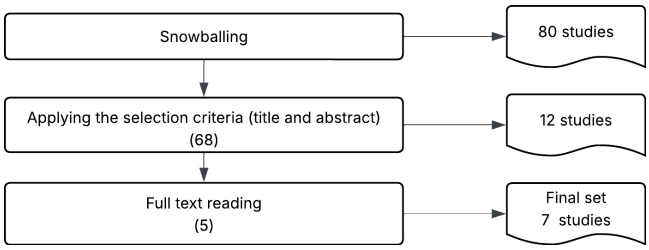

Figure 1: Study selection process

4 Results

As previously detailed, a total of seven studies were selected to analyze and answer the two RQs elaborated in this study, as detailed below.

4.1 Overview of the Selected Studies

The quasi-systematic mapping selected seven studies, all published between 2023 and 2025. Table 5 consolidates, for each study, how

Table 2: Data Extraction and Descriptions

RQ	DE	Description	Answer options
RQ1	DE1	The study analyses the uses of LLMs in SLR in medical field	Yes, Partially, No
RQ1	DE2	The study analyses the creation of search strings	Yes, Partially, No
RQ2	DE3	Which LLMs have been utilized?	ChatGPT; Llama; Claude; Others.

it uses LLMs (RQ1) and which models it employs (RQ2). The paragraphs below extend this summary with the main approaches and findings of each study.

Adam *et al.* [1], the seed study of this mapping, investigated the use of LLMs to generate search strings in medicine. They fine-tuned an LLM on a dataset of search queries from PROSPERO, a repository of SLRs for health research, reaching a sensitivity of 86%. Compared to human-generated queries, however, the model led to a 156% increase in the number of papers to read, reinforcing the need for human validation.

Wang *et al.* [17] tested the capability of ChatGPT (GPT-3.5) to generate and refine Boolean Queries from a title, topic, or non-refined search string. They found that its recall was inferior to that of existing tools, that the model was non-deterministic, and that it hallucinated in 55% of cases, using terms not found in the Medical Subject Headings.

Shen *et al.* [13] proposed a framework in which LLMs automate several stages of the SLR, including search string generation, reporting accuracy ranging from 61% to 99%. The model built Boolean Queries from PICO elements (Patient, Intervention, Comparison, and Outcome); key terms were extracted and enhanced for query construction, although human verification of the results remained necessary.

Wang *et al.* [18] developed a pipeline based on LLMs to accelerate SLR phases, generating the Boolean Query from the research question. The approach kept humans in the loop and used a Chain-of-Thought (CoT) method to iteratively enrich the queries. GPT-4 alone reached a recall of 0.074, whereas the CoT-based system, combined with human researchers and other strategies, reached 0.782, with a hallucination rate ranging from 0.06 to 0.138.

Wang *et al.* [19] proposed LEADS, a model fine-tuned specifically for medical literature that outperformed general-purpose LLMs such as GPT-4 in study recall. Leveraging the PICO framework to extract key elements and build search strings, LEADS saved experts between 20.8% and 26.0% of their time across different SLR steps; for search string generation it achieved recall scores of 24.68 and 32.11 for publication and trial searches, respectively, surpassing GPT-4's scores of 5.79 and 6.74.

Chelli *et al.* [4] conducted a comparative analysis of hallucination rates and reference accuracy, using GPT-4 and Bard to replicate medical SLRs by prompting the models to generate literature references. They observed hallucination rates of 28.6% for GPT-4 and 91.4% for Bard, with both models showing a precision below 14%.

Wang *et al.* [16] explored generating natural language queries from Boolean Queries, showing that LLMs such as ChatGPT (GPT-3.5) and Alpaca can grasp the research context and construct natural language queries. This indicates the potential for further investigation in the Software Engineering domain.

RQ1: How have LLMs been studied to generate search strings in the medical field?

Five prompting strategies were utilized to generate search queries in the field of medicine (see Table 3 – column 1).

Table 3: Prompt strategies used to elaborate search string in the medical field

Strategy	Studies
Zero-Shot	[4, 16, 17, 19]
Few-Shot	[13, 17–19]
Fine-tune	[1, 16, 19]
Multi-prompt	[17, 18]
Chain-of-thought (CoT)	[18]

The Zero-Shot strategy involves creating a prompt for the model without providing any examples of the input or the expected output. This approach was adopted to assess the capabilities of general-purpose models. Studies utilizing this strategy have shown that LLMs can understand frameworks like PICO and formulate syntactically correct Boolean Queries. However, the results indicate that this approach results in a high rate of hallucinations, with 55% of tests generating non-existent terms, which compromises the reliability of the search strings.

Few-Shot is a strategy that provides the model with a limited number of examples of the prompt and expected output to guide the model's generation pattern [13]. Authors used this strategy to demonstrate the syntax and objectives of the query to the LLM [17]. As the main result, it was observed that integrating example queries significantly improves overall query effectiveness, with higher recall [17].

Fine-tuning models is an effective strategy to address the limitations of general models. Research has suggested fine-tuning LLMs on specialized datasets. For example, the "Literature Search Sandbox" [1] used 10,346 search queries from the PROSPERO dataset to fine-tune a Mistral-7B model. This approach achieved a high sensitivity of 86%, but it also resulted in a wider range of results, which increased the workload for human reviewers. Similarly, Wang et al.[19] developed a foundation model based on Mistral-7B, fine-tuned with a dataset of 453,625 publications specifically for medical literature mining. This model outperformed general-purpose models in its task.

The multi-prompt strategy addresses the limitations of using a single prompt for generating queries. This approach breaks down the task of creating a search string into smaller, manageable steps by utilizing a sequence of prompts [17, 18]. Research by Wang *et al.* [17] demonstrated the effectiveness of this method, showing that a multi-prompt strategy can increase precision by up to 39.1% and recall

by 11.2% compared to using a single prompt. This improvement is largely due to the reduction of syntactical errors and enhanced term coverage.

CoT introduces a new strategy for prompting language models, enabling them to generate intermediate steps rather than just delivering the final output. A study by Wang *et al.* [18] found that incorporating this approach into a pipeline can enhance document retrieval. While the baseline recall was 0.074, the recall improved to 0.782 when using CoT, Retrieval-Augmented Generation, and human interaction. This demonstrates that CoT effectively addresses low recall without requiring a fine-tuned model.

RQ2: Which LLMs have been utilized in these studies?

The studies adopted two types of LLM: proprietary models (seven of seven studies) and open-source models (four of seven), as shown in Table 4.

Table 4: Types of LLMs used in Systematic Literature Review in the medical field

Model	Studies
Proprietary Model	[1, 4, 13, 16–19]
Open-Source Model	[1, 16, 18, 19]

Proprietary Models are LLMs with closed architectures, meaning their weights and internal code are not accessible for researchers to inspect, modify, or train directly. Most studies that used these models demonstrated their ability to follow complex instructions in Zero-Shot or Few-Shot scenarios. However, they often struggle with reliability, exhibiting high rates of hallucination and a tendency to generate non-existent terms. The main Proprietary Models referenced in these studies include GPT-3.5 and GPT-4 from OpenAI, Bard (Gemini) from Google, and Claude Sonnet and Haiku from Anthropic.

Open-source models allow researchers to modify the weights or the code to meet specific domain needs. These models enable the use of specialized datasets to fine-tune large language models (LLMs), which helps generate relevant and context-sensitive search queries in the medical field. Researchers using this type of model often implement it in its original form [16] or customize it through fine-tuning based on their specific requirements [1, 16, 18, 19]. The main Open-Source models identified in the selected studies are *Mistral-7B* (along with its instruction-tuned variants), *LLaMA*, and *Alpaca*.

Table 5 outlines the primary strategies and models employed in each study.

5 Discussion

The use of LLMs in SLRs within the medical field has been a topic of discussion since 2023. Adopting these models could represent a promising research avenue in medicine. This quasi-systematic mapping analysis highlights the models that have been utilized and how they are tested in various studies.

The main contribution of this work is the identification of the patterns used in medicine to serve as a guide for Evidence-Based Software Engineering. Most studies have demonstrated that in the medical field, various strategies and models can be used as primary tools to develop search strings. As previously observed in medicine, the main strategies adopted are:

- **Zero-Shot** – This strategy is used in the medical field to verify the baseline capabilities of general-purpose LLMs. Similarly, it could be applied in software engineering to evaluate how LLMs generate domain-specific terminology and to establish a baseline accuracy for comparing more complex strategies.
- **Few-Shot** – In the medical field, this strategy presented better results than the Zero-Shot prompting. In the software engineering field, this strategy could be tested to verify whether providing the model with guidelines and examples can lead to more precise search strings.
- **Fine-tune** – Fine-tuning is a strategy in the medical field used to train models to specialize in specific domains. This approach has shown excellent results, particularly in the development of search strings. In the field of software engineering, this type of model may not be ideal since its application requires curation of a domain-specific training dataset derived from computational academic research. However, with this dataset, the fine-tuning strategy is promising.
- **Multi-Prompt** – The multi-prompt model breaks down each step of the SLR into smaller steps, making more manageable tasks. This approach has proven to be effective in various contexts, especially in the recall and the reduction of hallucinations since the researcher can verify each generated term before the LLM continues to the next step. In software engineering, the multi-prompt model can be applied by dividing the creation of a search string into smaller stages, such as extracting the main concepts from the research questions and keywords, then generating synonyms and assembling the terms to finish the formulation of the search string.
- **Chain-of-Thought** – The CoT strategy can be effectively used alongside other approaches. This method enhances the capabilities of models by prompting them to generate intermediate reasoning. Studies have demonstrated that CoT increases transparency in a model's reasoning process, allowing researchers to understand how the model arrives at its responses. Additionally, when combined with other strategies, CoT has been shown to improve recall and reduce hallucinations. In the field of software engineering, this approach can be utilized to verify how a model generates text based on a given prompt, helping researchers comprehend its behavior when creating search strings with higher recall.

In the field of medicine, the types of LLMs utilized are:

- **Proprietary Models** – These types of models can be used as baselines, similar to how medicine has employed and validated them against fine-tuned models based on Open-Source technologies. In the field of software engineering, such models could help researchers determine which models enhance search strings and allow for comparisons with fine-tuned models.
- **Open-Source Models** – The field of medicine has adopted these models primarily for fine-tuning purposes, and it has been verified that fine-tuned models exhibit achieve better performance for SLR than non-fine-tuned models. In the field

Table 5: Summary of Selected Studies on LLMs for Systematic Reviews in Medicine

Studies	How the study analyses the uses of LLMs (RQ1)	Which LLMs have been utilized (RQ2)
[1]	Used a fine-tuned model based on the PROSPERO dataset to generate search strings based on title and key questions	GPT-4, Fine-tuned Mistral-Instruct-7b.
[17]	Tested the capability of GPT-3.5 to generate and refine Boolean Query based on the title, topic or a non-refined search string.	ChatGPT (GPT-3.5).
[13]	Created a framework of LLMs to automate various steps of the SLR including the generation of the search string	ChatGPT-4 (GPT-4).
[18]	Created a pipeline based on LLMs to accelerate SLR phases. The Booleang String is generated based on the research question.	TrialMind (fine-tuned Mistral-7B), Claude-3-Sonnet, GPT-4.
[19]	Developed a system (LEADS) to help researchers in substeps of a SLR, including the creation of a search string.	Fine-tuned Mistral-7B, Llama, Mistral, BioMistral, MedAlpaca, GPT-4o, GPT-3.5, Haiku-3.
[4]	Conducted a comparative analysis of hallucination rates and accuracy for query generation.	ChatGPT (GPT-3.5/4) and Bard (Gemini).
[16]	Explored generating natural language queries based on the Boolean Query.	ChatGPT and Alpaca.

of software engineering, this model could serve as a foundational for training on software engineering data, particularly for academic research.

Together, these strategies and model types can be combined to create new frameworks and pipelines for search string formulation. Specifically, a promising approach for software engineering would integrate the most effective strategies from medicine: utilizing *Few-Shot* to provide research context, *fine-tuning* to adapt models to the computing domain, *CoT* to enforce logical reasoning before outputting the final query, and *multi-prompting* to divide the generation process, keeping the researcher aware of every step. At the same time, the use of different models can be tested to verify how well each model performs in the software engineering research, specifically, Proprietary Models and fine-tuned Open-Source models.

Research in the medical field has adopted the PICO framework, along with various medical datasets, to enhance model performance. In software engineering, different strategies can be utilized to help models generate search strings, including the use of relevant keywords, the research questions and objectives. Additionally, fine-tuning models with data from software engineering research can also be an effective approach when developing search strings.

Despite significant advances in medicine regarding the use of LLMs for generating search strings, studies still show that human intervention is essential. Similarly, in software engineering, it has become clear that researchers must remain actively involved in the process, using LLMs as tools, since the models are non-deterministic and can generate hallucinations.

5.1 Threats to Validity

The execution of a single round of forward and backward snowballing poses a threat to validity, as it could result in the loss of relevant studies. However, the seven selected studies offer a comprehensive overview of how medicine has adopted LLMs for SLR and the search string creation, which may facilitate future applications in the field of software engineering.

5.2 Future Work

For future work, we propose an empirical study to assess the capabilities of LLMs in generating search strings for software engineering research. By comparing general-purpose and fine-tuned models across distinct prompting strategies (e.g., CoT, Few-Shot, and multi-prompting), we aim to systematically evaluate their strengths and limitations. Furthermore, to enhance the methodological rigor of future systematic mappings, we propose integrating additional rounds of snowballing to mitigate potential biases and maximize the exhaustiveness of the selected literature.

6 Conclusion

This paper aimed to identify the current state of the use of LLMs in SLR within the field of medicine, specifically focusing on how different models are utilized during the formulation of search strings. Our analysis reveals that various strategies and models have been employed in the medical field. Some studies highlight the use of different models, and in certain cases, human intervention is necessary to identify the most relevant search strings. In software engineering, the use of both general-purpose models and fine-tuned models, along with techniques such as CoT, Few-Shot, and Multi-Prompt, could facilitate the search string development process.

Artifact Availability

The replication package for this study can be found at [1]. This package includes the processes of backward and forward snowballing, as well as the study selection log. It provides details on the inclusion and exclusion decisions, along with the rationale for each candidate study. The data is released under the CC BY 4.0 license.

References

[1] Gaelen P Adam, Jay DeYoung, Alice Paul, Ian J Saldanha, Ethan M Balk, Thomas A Trikalinos, and Byron C Wallace. 2024. Literature search sandbox: a large language model that generates search queries for systematic reviews. *JAMIA Open* 7, 3 (oct 2024), ooae098. arXiv:https://academic.oup.com/jamiaopen/article-pdf/7/3/ooae098/59342336/ooae098.pdf doi:10.1093/jamiaopen/ooae098

[1] https://github.com/LeonardoTironi/CBSOFT2026

[2] Yadagiri Annepaka and Partha Pakray. 2025. Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems* 67, 3 (2025), 2967–3022. doi:10.1007/s10115-024-02310-4

[3] Leonardo Cairo, Glauco de F. Carneiro, Miguel P. Monteiro, and Fernando Brito e Abreu. 2019. Towards the Use of Machine Learning Algorithms to Enhance the Effectiveness of Search Strings in Secondary Studies. In *Proceedings of the XXXIII Brazilian Symposium on Software Engineering* (Salvador, Brazil) *(SBES '19)*. Association for Computing Machinery, New York, NY, USA, 22–26. doi:10.1145/3350768.3350772

[4] Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research* 26 (may 2024), e53164. doi:10.2196/53164

[5] Vinicius dos Santos, Anderson Y. Iwazaki, Katia R. Felizardo, Érica F. de Souza, and Elisa Y. Nakagawa. 2024. Sustainable systematic literature reviews. *Information and Software Technology* 176 (2024), 107551. doi:10.1016/j.infsof.2024.107551

[6] Gerald Gartlehner, Barbara Nussbaumer-Streit, Candyce Hamel, Chantelle Garritty, Ursula Griebler, Valerie Jean King, Declan Devane, and Chris Kamel. 2025. Responsible Integration of Artificial Intelligence in Rapid Reviews: A Position Statement From the Cochrane Rapid Reviews Methods Group. *Cochrane Evidence Synthesis and Methods* 3, 6 (Nov. 2025), e70063. doi:10.1002/cesm.70063

[7] Bo HU and Liuna Geng. 2024. Accelerating Systematic Reviews with Large Language Models: Current Practices and Recommendations. doi:10.31219/osf.io/wm2av

[8] Lennart Hüsing and Tim Schupp. 2025. LLMs in Research Methodology: Opportunities and Challenges. In *Proceedings of the 2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE '25)*. IEEE Press, Ottawa, ON, Canada.

[9] Barbara Kitchenham. 2004. Procedures for Performing Systematic Reviews. *Keele, UK, Keele Univ.* 33 (08 2004).

[10] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-based software engineering and systematic reviews*. Vol. 4. CRC press, Boca Raton, USA.

[11] Christopher Marshall, Barbara Kitchenham, and Pearl Brereton. 2018. Tool Features to Support Systematic Reviews in Software Engineering - A Cross Domain Study. *E-Informatica Software Engineering Journal* 12 (01 2018). doi:10.5277/e-Inf180104

[12] Mehwish Riaz, Muhammad Sulayman, Norsaremah Salleh, and Emilia Mendes. 2010. Experiences conducting systematic reviews from novices' perspective. (04 2010).

[13] Jiashu Shen, Zhiyao Luo, Danping Jia, Siran Wang, Feng Sun, and Jinge Wu. 2025. Large language model enhanced framework for systematic reviews and meta-analyses. *BMJ Digital Health & AI* 1, 1 (10 2025), e000017. doi:10.1136/bmjdhai-2025-000017

[14] Zalaya Simmons, Beti Evans, Tamsyn Harris, Harry Woolnough, Lauren Dunn, Jonathon Fuller, Kerry Cella, and Daphne Duval. 2025. Assessing the Feasibility and Acceptability of a Bespoke Large Language Model Pipeline to Extract Data From Different Study Designs for Public Health Evidence Reviews. *Cochrane Evidence Synthesis and Methods* 3, 6 (nov 2025), e70061. doi:10.1002/cesm.70061

[15] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics* 6, 3 (sep 2005), 239–251. doi:10.1093/bib/6.3.239 PMID: 16212772.

[16] Shuai Wang, Harrisen Scells, Bevan Koopman, Martin Potthast, and Guido Zuccon. 2023. Generating Natural Language Queries for More Effective Systematic Review Screening Prioritisation. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Beijing, China) *(SIGIR-AP '23)*. Association for Computing Machinery, New York, NY, USA, 73–83. doi:10.1145/3624918.3625322

[17] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? arXiv:2302.03495 [cs.IR] https://arxiv.org/abs/2302.03495

[18] Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine* 8, 1 (aug 2025), 509. doi:10.1038/s41746-025-01840-7

[19] Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K. Gill, Sun-Hyung Kim, Yijia Li, Yi Liu, Yiming Luo, Hanley Ong, Justin F. Rousseau, Irfan Sheikh, Jenny J. Wei, Ziyang Xu, Christopher M. Zallek, Kyungsang Kim, Yifan Peng, Zhiyong Lu, and Jimeng Sun. 2025. A foundation model for human-AI collaboration in medical literature mining. *Nature Communications* 16, 1 (2025), 8361. doi:10.1038/s41467-025-62058-5

[20] Claes Wohlin, Marcos Kalinowski, Katia Romero Felizardo, and Emilia Mendes. 2022. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology* 147 (2022), 106908. doi:10.1016/j.infsof.2022.106908