ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity

Katia Romero Felizardo*
katiascannavino@utfpr.edu.br
Universidade Tecnológica Federal do Paraná
Cornélio Procópio, PR, Brazil

Anderson Deizepe
Universidade Tecnológica Federal do Paraná – UTFPR
Cornélio Procópio, PR, Brazil
deizepeanderson@gmail.com

Márcia Sampaio Lima
Universidade do Estado do Amazonas
Manaus, AM, Brazil
msllima@uea.edu.br

Tayana Uchôa Conte
Universidade Federal do Amazonas
Manaus, AM, Brazil
tayana@icomp.ufam.edu.br

Igor Steinmacher
Northern Arizona University
Flagstaff, AZ, USA
Igor.Steinmacher@nau.edu

ABSTRACT

Context: The Systematic Literature Review (SLR) process involves searching, selecting, and synthesizing relevant literature on a specific research topic for evidence-based decision-making in Software Engineering (SE). Due to the time-consuming of the SLR process, tool support is essential. **Gap:** ChatGPT is a significant advancement in Natural Language Processing (NLP), and it can potentially accelerate time-consuming and propone-error activities, such as the selection activity of the SLR process. Therefore, having a tool to assist in the selection process appears beneficial, and we argue that ChatGPT can facilitate the analysis of extensive studies, saving time and effort. **Objective:** We aim to evaluate the accuracy (i.e., studies correctly classified) of using ChatGPT–4.0 in SLR in SE, particularly to support the first stage, based on the title, abstract, and keywords. **Method:** We assessed the accuracy of utilizing ChatGPT for selecting studies, the first stage, to be included in two SLRs (SLR1 and SLR2), in contrast to the conventional method of reading the title and abstract. **Results:** The accuracy of ChatGPT supporting the initial selection activity was 75.3% (SLR1 – 101 correct selections: 48 inclusions and 53 exclusions; 33 incorrect selections: 17 inclusions and 16 exclusions) and 86.1% (SLR2 – 386 correct selections: 113 inclusions and 273 exclusions; 62 incorrect selections: 27 inclusions and 35 exclusions). **Conclusions:** Our accuracy results indicate that it is not advisable to outsource the selection process to ChatGPT completely. However, it could be valuable as a support tool, aiding novice or experienced researchers when they are in doubt.

CCS CONCEPTS

• **General and reference → General literature**.

ESEM '24, Sun 20 - Fri 25 October 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

ChatGPT, Systematic literature review, Selection of studies, Software Engineering

ACM Reference Format:
Katia Romero Felizardo, Anderson Deizepe, Márcia Sampaio Lima, Tayana Uchôa Conte, and Igor Steinmacher. 2024. ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity. In *Proceedings of 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The Software Engineering (SE) community has increasingly conducted Systematic Literature Reviews (SLRs) to summarize evidence from relevant studies and highlight the state-of-the-art in a given research topic [46]. While SLRs offer advantages such as handling information from various studies in an unbiased and repeatable manner to produce auditable results and identifying research gaps and perspectives for future research, they also present several challenges, with only punctual solutions available for some of these issues [16].

Santos and his collaborators [16] emphasize that, although the SE academic community has increasingly conducted SLR, the process support still needs to improve. The several specific solutions proposed for those problems have yet to be widely influential. The issues related to the SLR issues in the context of SE include the following [14, 36, 37, 90]: i) SLR conduction is still time- and effort-consuming [6, 16, 22, 61, 78]; ii) the documentation of SLR process is poor [16]; iii) primary studies quality assessment is not appropriate, especially for qualitative studies [16, 95]; iv) learning the SLR process and defining a research protocol is challenging for novice researchers [95]; and v) the access and acquisition of relevant studies across multiple e-libraries [4, 80, 95]. Given the current problems with SLR execution in SE, different tools have been proposed and updated to cope with the challenges. The existing tools semi-automate some phases of the SLR process [13, 29, 39, 58, 88], focusing on: i) supporting the protocol creation [28, 104]; ii) retrieving relevant studies from the digital libraries [7, 30, 40, 50, 57, 59,

73, 74, 76, 85–87, 89]; iii) recommending appropriate studies for inclusion [1, 2, 12, 17, 19, 23, 26, 27, 31, 48, 51, 54, 64–68, 77, 79–81, 84, 91, 99, 102, 102, 103]; iv) assessing of quality of the studies [8]; v) extracting data from the primary studies [9, 38, 43, 62, 71, 82, 92]; vi) synthesizing the evidence [24, 33, 48, 63]; and vii) reporting SLR conduction and its results [47].

ChatGPT, a representative Large Language Model (LLM) that OpenAI developed, is an alternative tool that has emerged to support the SLR process [10, 35, 53, 96, 98]. Although ChatGPT is relatively new, several studies—especially in the medical field—acknowledge its capabilities (and limitations) in automating SLRs [10, 11, 35, 60, 72, 96, 98, 101]. However, further research to empirically investigate "how" we can do it at different activities of the SLR process is recommended [10], especially in the SE field. Still, some SLR practices in SE and medical research differ in some points. Primarily, the volume of empirical studies in SE is smaller compared to the medical field. Additionally, the research methodologies employed by SE researchers are generally less stringent than those applied by their medical counterparts [46].

As a first step in this investigation, we focus on of the one most time-consuming and error-prone Systematic Literature Review (SLR) activities: the study selection phase [99]. Traditionally, this phase begins with manually reviewing all retrieved articles' titles, abstracts, and keywords. In the second stage, researchers read the full text of articles classified as 'included' from the first stage. The number of studies included in SLRs in Software Engineering (SE) varies widely, ranging from tens to thousands of studies. Working through these studies, especially in large numbers, requires significant effort from SE researchers. Furthermore, many studies retrieved by the search process are irrelevant to the underlying research question, leading to additional wasted effort [25]. A tool that automates some steps of the selection process could significantly benefit researchers by saving time and reducing effort.

Given this context, our study aims to **assess the accuracy of ChatGPT to support the first stage of the selection activity in SLRs in SE.** We will answer the following research question (RQ): "*To what extent is ChatGPT able to support the SLR process, especially concerning the selection activity?*"

To achieve our goal, we replicated the first selection phase of two SLRs with the support of ChatGPT. To do so, our work went through three (3) major stages: i) extracting the SLR data for replication, including lists of included and excluded studies, as well as inclusion and exclusion criteria (we used the list of included and excluded studies as our benchmark); ii) the selection activity was replicated by one researcher prompting ChatGPT with information about selection criteria using a predefined template; iii) results were analyzed by compared to the original study to assess the accuracy of ChatGPT. The LLM's outcomes were compared with the benchmark set by the SLR authors, with correct classification occurring when it aligned with human judgment and incorrect classification otherwise.

Following the selection process and considering the benchmark, we reached an overall accuracy of 80.7%, considering ChatGPT inclusions/exclusions. Specifically for SLR1, ChatGPT correctly classified 101 articles (48 true positives (TP) + 53 true negatives (TN)). On the other hand, 33 articles were incorrectly classified (16 false negatives (FN) + 17 false positives (FP)), reaching 75.3% of accuracy. Regarding SLR2, the accuracy of ChatGPT was 86.1%, with 386 correct classifications (113 TP; 273 TN) and 62 incorrect ones (35 FN; 27 FP) compared to the benchmark.

The contributions of this work are to point out the benefits and threats of using ChatGPT–4.0 to support the selection activity and provide empirical evidence on the adoption of ChatGPT in the SLR context within the SE area. We recomend that SE researchers could regard ChatGPT as a supplementary "opinion" to mitigate biases and inaccuracies in the included evidence, thereby resulting in fair outcomes.

This paper is then organized as follows. Section 2 the tools already available to assist in the automation of SLR in SE. Section 3 details how we plan the ChatGPT accuracy evaluation to select studies instead of the traditional method. Section 4 presents and contextualizes our findings, focusing on describing quantitative data related to the accuracy of ChatGPT in selecting studies. Subsection 4.1 offers a discussion of our results and describes limitations and potential issues threatening the validity of our findings. Finally, Section 5 explains the relevance of our findings for SE researchers and outlines possible future research directions.

2 RELATED WORK

2.1 Tools to support the automation of SLR in SE

Developing the protocol, searching for evidence, selecting relevant studies, extracting data, and synthesizing the evidence are the SLR activities with the most potential for automation [14, 20, 88, 94].

Felizardo et al. [20] affirm that Visual Text Mining (VTM) techniques have already been used to support formulating research questions [28], searching [57], selecting studies [2, 19, 26, 65], and synthesizing [24]. Some authors point out that automating data extraction and synthesis via machine learning [83, 94] is an important research direction.

There are several SLR support tools available [3, 5, 55, 56]. In particular, some of the tools that support researchers in conducting SLR in SE are presented below [13, 20, 29, 39, 58]:

(1) *Linked Data*[1] – suggests text mining to semi-automate the selection activity [91].
(2) *PEx/Revis*[2] – provides visual representations of studies to support selection activity [19, 21, 24, 25].
(3) *Review Tolkit*[3] – supports simple literature filtering, design of a taxonomy, classification of literature and analysis of the classification by generated diagrams [32].
(4) *SluRp*[4] – supports the whole SLR process, the management of a large number of studies and shares tasks among a group of researchers [13].
(5) *SLRONT*[5] – describes common terminologies and their relationships during SLR process [89].
(6) *SLR-Tool*[6] – supports the whole SLR process and provides uses text mining to refine search results [29].

[1] No prototype available
[2] *http://vicg.icmc.usp.br/vicg/tool/1/projection-explorer-pex*
[3] https://github.com/sebastiangoetz/slr-toolkit
[4] https://uhra.herts.ac.uk/handle/2299/14730
[5] No prototype available
[6] https://alarcos.esi.uclm.es/

(7) *StArt*[7] – assists SLR conduction from protocol creation to results presentation through graphics [17, 18, 39].
(8) *UNITEX*[8] – automatically extracts knowledge from studies using text mining [92].
(9) SLR automation tools examples in Medicine and other domains can be found on the following websites. [9]
(10) *Parsifal*[10] – supports researchers in performing SLR within the context of SE.

Unfortunately, most of these tools are still in the early stages of development (prototype) and are often neglected or abandoned. Another issue involving automation tools for the SLR in the SE area is the lack of large-scale validation [20].

2.2 The use of LLMs to assist in the SLR process

As shown in Table 1, several studies have investigated the adoption of LMMs to support the SLR process in different research domains, including Dental [53], Educational [44], and Medical [10, 11, 35, 60, 72, 96, 98, 101]. Some studies have evaluated whether LLMs could support the SLR process as a whole [10, 53, 98], while others focus on specific SLR activities such as suggesting new SLR topics [35], searching for evidence [96], selecting relevant studies [34, 45, 75, 97, 100], extracting [45, 53, 98], and synthesizing data [101].

According to Waseem et al. [98], the impacts of LLMs on SLRs include outlining research questions, formulating search strings, collecting data, and synthesizing relevant literature. Moreover, LLMs can reduce the researchers' workload in time-consuming and error-prone activities. By automating these activities, ChatGPT allows researchers to save time to focus their attention on higher-level tasks, such as critically evaluating the literature. The authors emphasize the importance of researchers and LLMs collaborating to ensure accurate results. For example, LLMs can identify terms, concepts, and themes for synthesis while researchers review and analyze them to confirm their relevance to the research questions. Other researchers [10, 53] reinforce that LLMs can accelerate the SLR process while requiring less human labor.

Huotala et al. [41] investigated to what extent LLM can facilitate title-abstract selection compared to human researchers. They concluded that using LLMs to automate title-abstract selection seems promising but does not significantly improve human accuracy. Other authors [34, 45, 75, 97, 100] from the field of medicine also believe that LLMs have the potential to support the application of inclusion and exclusion criteria selection, save time and effort for researchers, and contribute to the overall quality of SLRs. However, they commonly advise that the studies were conducted under specific conditions offering preliminary evidence, and more studies are needed before the wide use of LLMs.

Najafali and his collaborators [60] agree that LLMs in their current state are limited, and considerable improvements are needed to execute the entire SLR process singlehandedly. For them, LLMs provide a limited reference list and incomplete manuscripts when tasked with doing so. It also makes general claims to answer the research questions.

In line with Najafali et al. [60], Anghelescu et al. [11] reaffirm that one of LLMs's drawbacks is distinguishing between truth and falsehood in their answers. Similarly, other researchers [72] note that the output appears to be valid at a high level, but much is erroneous and needs vetting. In this context, a concern is that what literature LLMs are trained in is unclear, and it probably does not include articles with closed access. Consequently, the answers are based on a restricted, not updated, evidence base [53]. Another challenge concerns originality, transparency, credibility, and bias issues.

Alshami et al. [10] also scrutinized the SLR process's quality regarding reproducibility, transparency, and bias when utilizing LLMs against traditional SLR methods. They broadly investigated the potential of LLMs through all SLR activities, from generating research questions and keywords for search strings to their employment in extracting and synthesizing information. The findings exposed that LLMs can help generate research questions and suggest boolean research terms, but LLMs are restricted to data extraction. Comparing the accuracy of utilizing LLMs for selecting studies from abstracts to be included in an SLR, in the face of the conventional methods, ChatGPT exhibits an overall accuracy of 88%, leading to significant time and effort savings.

Gupta et al. [35] explored how accurate LLMs could be in suggesting SLR topics. They concluded that, overall, 55% of SLR topics derived from LLMs were novel. Relying on assistance for generating search strings, Wang and his coauthors [96] emphasize that LLMs can follow complex instructions, making them a valuable tool for facilitating researchers performing SLR.

Regarding data extraction, by sharing the complete text and specifying the desired information or data to be extracted, AI can be of assistance [53]. Concerning synthesis, Hoai et al. [101] point out that LLMs can generate conclusions similar to those established by humans. However, they observed that LLMs cannot summarize information when more than three types of information are compared.

The role of LLMs as decision support tools for the SLR process and their potential and risks are summarized in Table 2.

The previously mentioned studies raised important questions about using ChatGPT in the SLR process. The most listed benefit reported by researchers was that ChatGPT speeds up the SLR process with a lower level of human effort [53, 98].

Researchers found that ChatGPT helped enhance the overall quality of search strings by providing synonyms [10, 96]. Additionally, ChatGPT is a valuable tool for extracting essential information from texts [98]. Another benefit identified was support in summarizing data [101] Furthermore, ChatGPT demonstrated its utility as a tool assistant for suggesting new SLR topics [35].

Limited research has been driven on using ChatGPT to assist in research question formulation [10] or selecting studies [41]. Most existing literature [60, 96, 98] agrees that human verification is mandatory to minimize errors. Moreover, researchers fear ChatGPT could provide generic or even erroneous answers. Researchers advertise that if ChatGPT does not "know" the answer, it will "fabricate" one using a relevant language related to the question, which is conveyed in a persuasive tone [11, 60, 72]. Furthermore, using

[7] http://lapes.dc.ufscar.br/tools/start_tool
[8] https://unitexgramlab.org/pt
[9] https://parsif.al/ https://www.rayyan.ai/
https://aut.ac.nz.libguides.com/systematic_reviews/tools
https://ktdrr.org/resources/sr-resources/tools.html
https://libguides.jcu.edu.au/rapidreview/tools
[10] https://parsif.al/

Table 1: Literature on adopting ChatGPT to aid the SLR process.

Title	Domain	Ref.
Application ChatGPT in conducting systematic reviews and meta-analyses	Dental	[53]
Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?	Medical	[72]
Automated paper screening for clinical reviews using large language models: data analysis study	Medical	[34]
Automated title and abstract screening for scoping reviews using the GPT-4 large language model	Medical	[100]
Bio–SIEVE: exploring instruction tuning large language models for systematic review automation	Medical	[75]
Can ChatGPT write a good boolean query for systematic review literature search?	Medical	[96]
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	Medical	[45]
Conducting systematic literature reviews with ChatGPT	Computer Science	[98]
Comparing meta-analyses with ChatGPT in the evaluation of the effectiveness and tolerance of systemic therapies in moderate-to-severe plaque psoriasis	Medical	[101]
Expanding cosmetic plastic surgery research with ChatGPT	Medical	[35]
Harnessing the power of chatGPT for automating systematic review process: methodology, case study, limitations, and future directions	Medical	[10]
PRISMA systematic literature review, including with meta-analysis vs. Chatbot/GPT (AI) regarding current scientific data on the main effects of the calf blood deproteinized hemoderivative medicine (actovegin) in ischemic stroke	Medical	[11]
The promise and challenges of using LLMs to accelerate the screening process of systematic reviews	Software Engineering	[41]
Truth or lies? The pitfalls and limitations of ChatGPT in systematic review creation	Medical	[60]
Zero-shot generative large language models for systematic review screening automation	Computer Science	[97]

ChatGPT raised concerns about reliability. Since only versions after 2017 ChatGPT have Internet access to update their knowledge, the reference list used as a source of evidence before this date can be limited and outdated [53, 60]. Therefore, we should carefully evaluate ChatGPT responses because they may not always be trustworthy.

Table 2: Using LLMs to support the SLR process and outline their possible benefits and risks.

Using ChatGPT for assist SLR	
Advantages	• Reduce workload [10, 53] • Outline research questions [10] • Use correct Boolean operators and syntax [96] • Suggest new SLR topics [35] • Suggest terms of research interest for search string [96] • Suggest synonyms for search string [10, 96] • Extract data to support summarization [53, 98] • Synthesise data to answer questions [101]
Limitations	• Provide generic information [53, 60, 72] • "Fabricate" information and convey it in a persuasive tone [11, 60, 72] • Lack of transparency [72] • Originality issues [72] • Restricted and not updated evidence base [53, 60] • Offer a limited reference list [60] • Limited summarization data [10, 101] • Write an incomplete report [60]

Another main concern focused on the implications of using ChatGPT concerning originality [53]. Researchers also doubt reproducibility, transparency, and bias [10, 53].

Similar to the previously mentioned studies, we also investigated the use of LLM in the context of SLRs. Although LLMs hold promise in supporting various other activities within the SLR process, including summarizing and synthesizing evidence, we focus on its application in selecting studies. In particular, in the SE area, Huotala et al. [41] also evaluated LLM to assist title-abstract selection; however, unlike Huotala et al. [41], our study did not focus on prompt optimization and adoption of LLMs to simplify abstracts for human selection.

3 METHODOLOGY

In this paper, we focus on understanding the extent to which ChatGPT can support researchers in selecting relevant studies during the first phase of the SLR process. The study selection is a labor-intensive and time-consuming activity that often involves classifying thousands of candidate studies. Reducing this effort would largely benefit the researchers by reducing their workload with tasks that may be automated. Therefore, in this section, we detail the method followed to assess the accuracy of ChatGPT to help with this activity.

3.1 Replicated SLRs

We used a convenience sample of two SLRs in which the authors of the present study took part. We selected these two given the familiarity of the researchers with the studies, in addition to having all the data required to perform the study (a complete list of studies returned from the search, detailed selection process, and inclusion/exclusion decision from the first stage of the selection

activity, based on the title, abstract, and keywords). Still, we had access to the authors in case we needed to clarify any specific point of the studies. We detail both SLRs below.

SLR 1 – This is a tertiary study (i.e., an SLR of SLRs) on the "convergence of Human-Computer Interaction and Artificial Intelligence" [93]. They analyzed the article's title, abstracts, and keywords of 134 articles.

The selection process for this study followed rigorous SLR guidelines [46]. Four reviewers (one expert in Artificial Intelligence, one in Human-Computer Interaction, and two in secondary studies) independently screened all the potential articles. Meetings were performed to solve conflicts. 134 studies (sample) were randomly chosen from 536 candidate studies (90% of confidence level, 5% of error) as our dataset.

The data provided made it possible to evaluate the classification by contrasting ChatGPT's results with the classification of four experts. Each reviewer independently classified each article into one of the three scores: 0–exclude, 1–uncertain, and 2–include. For this SLR, articles were included whenever the total score was equal or higher to four (e.g., at least two reviewers agreed with the article's inclusion) and excluded otherwise.

The data (including and excluding studies) from the first stage of the selection activity, based on the title, abstract, and keywords reading, was openly available. The selection criteria were detailed. Secondary studies were included in the tertiary study [93] only if they met the two inclusion criteria, as stated in Table 3.

SLR 2 – This SLR aimed to identify primary studies describing user profiles in games or gamified environments and evaluate the impact of game elements within these environments based on the users' profiles [69]. The researchers analyzed the article's title, abstracts, and keywords of 448 primary studies.

In this SLR, the selection process also followed rigorous SLR guidelines [46]. The researchers started the SLR's search process by defining an oracle (known collection of studies). They defined the search string by extracting terms from the studies' keywords, titles, and abstracts in the oracle [70], which helped them validate the search string results. In addition, researchers rigorously validated the review protocol and search strategy definition and refined the selection process using appropriate inter-rated agreement measures. The data (included and excluded studies) from the first stage of the selection activity was also available. The selection criteria were detailed. Primary studies were included in the SLR if they met two inclusion criteria, as stated in Table 3. It was possible to evaluate the classification by comparing ChatGPT's results with a benchmark [69]. The researchers' classification is assigned for each article: 0 – exclude and 1 – include. Articles were included whenever they met all the inclusion criteria. Our evaluation considered the SLR1 and SLR2 expert classifications (i.e., benchmark) as "true" values.

3.2 Accuracy assesssment

As shown in Figure 1, we followed three (3) main steps to assess the accuracy of adopting ChatGPT in the selecting studies, as described below.

Table 3: Inclusion criteria (IC).

	SLR 1
IC1	• It is a secondary study (Systematic Review, Mapping Study, Rapid Review, or Systematic Mapping)
	AND
IC2	• It presents findings for converging Human-Computer Interaction and Artificial Intelligence
	SLR 2
IC1	• The article discusses the relationship between game elements and user types.
	AND
IC2	• The article discusses how to use the relationship between game elements and user types to analyze user engagement.

Stage 1: Input – To conduct the selection using ChatGPT, we got the title, abstract, and keywords of all articles for the initial selection (SLR 1 – 134; SLR 2 – 448).

Stage 2: Prompting ChatGPT –We provided a clear and specific input prompt following a prompt and predicted format [52]. To construct the prompt, we used verbatim excerpts extracted from the protocol. Initially, we inserted the study's goal, the research questions, and the selection criteria. Then, we iteratively removed each item until we had not observed a negative impact on the classifications. The goal was to create the best prompt based on the initial protocol defined by the authors. We observed that the context could be limited to presenting the role that the LLM would be playing, along with the selection criteria, without losing accuracy.

The exclusion criteria for the first round selection on SLR1 and SLR2 were the negative for of the inclusion criteria. Thus, we decided to include only the inclusion criteria in our prompts. Other exclusion criteria, such as "the article is a newer version of another article," were not treated as examples of criteria applied in the second selection stage while reading the full text.

We asked ChatGPT to answer using a rate of agreement for studies inclusion following a 7-point Likert scale (1– strongly disagree, 2– disagree, 3– somewhat disagree, 4– neither agree nor disagree, 5– somewhat agree, 6– agree, and 7– strongly agree). We used a 7-point Likert scale because seven points provide a more granular understanding of the ChatGPT agreement related to the selection criteria. We imposed specific constraints during the selection process to maintain consistency and control. These constraints encompassed classifying articles exclusively into the seven (7) rates. The template used to prompt ChatGPT is presented here:

```
"Assume you are a software engineering researcher conducting
a systematic literature review (SLR). Consider the
title, abstract, and keywords of a primary study.

Title: <the title of the primary study>
Abstract: <the abstract of the primary study>
Keywords: <the keywords of the primary study>

Using a 1-7 Likert scale (1 - Strongly disagree, 2 -
Disagree, 3 - Somewhat disagree, 4 - Neither agree nor
```

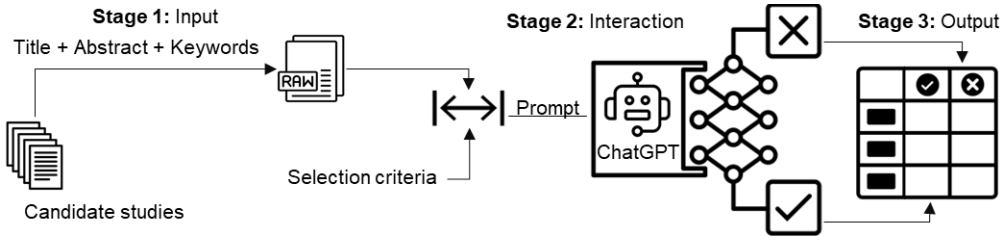

Figure 1: Method research: assessing the accuracy of adopting ChatGPT in selection studies.

`disagree, 5 - Somewhat agree, 6 - Agree, and 7 - Strongly agree) rate your agreement with the following statement (only number): <IC1 - Description of inclusion criteria 1.>.`

`Using a 1-7 Likert scale (1 - Strongly disagree, 2 - Disagree, 3 - Somewhat disagree, 4 - Neither agree nor disagree, 5 - Somewhat agree, 6 - Agree, and 7 - Strongly agree) rate your agreement with the following statement (only number): <IC2 - Description of inclusion criteria 2.>.`

In the context of LLMs, temperature, and top_p (nucleus sampling) are parameters that can be adjusted to fine-tune the balance between generating coherent and diverse response outputs, and changing these parameters allows control over the stability of the generated responses. A lower temperature (e.g., 0.2) produces less random and more deterministic output, generating high-confidence predictions. Conversely, a higher temperature (e.g., 0.7) leads to more randomness, encouraging the model to explore a broader range of possible outputs. Top_p controls randomness, and instead of sampling from the entire probability distribution of words, top_p sampling considers only the most likely words whose cumulative probability exceeds a certain threshold (p). This threshold dynamically adjusts based on the probability distribution, ensuring the model generates diverse yet relevant responses. We set the temperature to 0 and the top_p to 0.1.

Stage 3: Analysis – In our analysis, we consider as "included" the studies with a threshold greater than or equal to five through the Likert scale (5 – somewhat agree) for each inclusion criterion; we classified the other cases as excluded.

As illustrated in Figure 2, from a "conservative" perspective, we also analyzed our data considering the inclusion of studies with a rate greater than or equal to four instead of five. This conservative perspective considers that, in case of doubt, the paper should be included in the next step. A rate of 4 could indicate doubt.

We then compared the output with our benchmark (original output from the human analysis). Matches were considered true positives (TP) or negatives (TN), and the outputs that were not matched were classified as false positives (FP) or negatives (FN). Finally, the accuracy was calculated using the formula (TP + TN) / (TP + TN + FP + FN).

4 RESULTS

A summary of the results is presented in Table **??**. ChatGPT achieved correct classification when it matched human classification for the first selection round in the original paper (True Positive/True Negative) and incorrect otherwise (False Positive/False Negative). Concerning SLR 1, while the reviewers included 51 studies, ChatGPT included 75; 83 were excluded by reviewers and 59 by ChatGPT. Looking at ChatGPT classifications, 101 studies were correctly classified, and 33 were not. More specifically, the ChatGPT decisions were: correct classifications (TP+TN) – 101 (48+53); incorrect classifications (FP+FN) – 33 (17+16). The average execution time and tokens used were 1.72s and 918.2s, respectively.

Regarding SLR 2, the original reviewers included 148 studies, while ChatGPT included 113. Analyzing ChatGPT's classifications, it correctly classified 386 studies (113 TP + 273 TN) and incorrectly classified 62 studies (27 FP + 35 FN). The average execution time was 1.41s and 915.79 was the average token count.

We used Cohen's kappa [15] to determine the level of agreement, denoted as k, between (1) the SLR1 benchmark and ChatGPT and (2) the SLR2 benchmark and ChatGPT. For SLR1, the kappa value was $k = 0.59$. For SLR2, the kappa value was $k = 0.68$. These values indicate a moderate and a substantial agreement, respectively [49].

4.1 Discussions

In this paper, we bring evidence that ChatGPT may not provide accurate selection classifications for SLRs. Therefore, its adoption presents issues discussed in the sequence that must be thoroughly improved before widespread implementation in SLRs in the SE area.

• **Accuracy** – Similar to our results (75.3% and 86.1%), the accuracy of other studies published in the literature concerning applying LLMs to support title-abstract study selection has varied between 60 and 83.5% [34, 45, 75, 97, 100].

• **Incorrect classifications** – One observed limitation of the ChatGPT adoption in selecting studies is its potential for generating incorrect classifications. Researchers are aware that if ChatGPT does not "know" the answer, it will "fabricate" one [11, 42], making it difficult to distinguish between truth and falsehood answers (hallucinations). Therefore, instead of SE researchers adopting ChatGPT classifications as a unique source to classify studies, it may add a new perspective, like an "external reviewer". When only one (1) researcher performs the selection, he or she should consider discussing his or her decisions with other colleagues. In that respect, SE researchers could use the divergent ChatGPT classifications as clues about what studies should be doubly reviewed for inclusion or exclusion, replacing the external researcher review. Furthermore,

Table 4: Comparative of ChatGPT classifications regarding our benchmarks.

		Actual values (SLR1)				Actual values (SLR2)			
		Likert >= 4		Likert >=5		Likert >= 4		Likert>=5	
		Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Predicted values	Positive	50	35	48	17	128	68	113	27
	Negative	14	35	16	53	20	232	35	273
	Accuracy	63.4%		75.3%		80.3%		86.1%	

LLM could be employed to compare and analyze the different reviewers' decisions, supporting the group in a common sense about the inclusions and exclusions.

• **Loss of evidence (FN) –** Considering the articles incorrectly classified in SLR1, ChatGPT had FN 16 (relevant articles excluded). For SLR2, the number of FNs was 35. FN decisions have more negative impact on the results of an SLR than FP, since we may lose evidence.

Analyzing the loss of evidence at the end of the selection process, i.e., studies excluded by ChatGPT-4 when reading the title-abstract-keyword but included by the researchers (FN) and that remained in the final selection, reading the full text; for SLR1, there were two (2) losses of evidence, and for SLR2, the loss was four (4).

The "conservative" inclusion perspective, i.e., articles with rates 4 (neither agree nor disagree) in one of the criteria are included, changed the classification of 20 studies in SLR1. Eighteen correctly excluded studies (TN) were classified as included (FP), generating reading overhead in the second selection stage. However, two (2) studies incorrectly excluded (TN) were included (TP). From these two studies, one remains included in the second stage of selection activity, meaning that one (1) evidence that had been lost was recovered. About SLR2, 56 studies changed their classifications considering "conservative" inclusion. In total, 15 TP emerged, meaning that these pieces of evidence previously lost using the Likert-5 were recovered through Likert-4. However, 41 new FPs have emerged, and reviewers would have to read their full texts. Of these new FPs, only two (2) were retained by the reviewers after the second stage, reading the full text.

We believe that to improve the accuracy of ChatGPT responses, we could supplement its 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.

• **ChatGPT x Novices researchers –** Our study has two SLRs conducted by researchers of different expertise levels. SLR1 was conducted solely by experts, while SLR2 was the first SLR conducted by a PhD candidate and her supervisor. Despite having received training in conducting SLRs and having the support of an expert in the method, the researchers in SLR2, due to their lack of experience, may have been overly cautious during the selection process.

When teaching the procedure for conducting an SLR, it is emphasized that if there are doubts in the selection stage, the article should be accepted and moved on to the next stage. This cautious approach could explain the higher number of exclusions by ChatGPT compared with the researchers in SLR2. Specifically, 35 FN (studies excluded by ChatGPT that were included by the researchers) were included in SLR2. However, the impact of these FNs on the results of the SLR2 is relatively low. This is because out of the 35 studies excluded by ChatGPT in the selection stage, 31 were subsequently excluded by the researchers in the next stage.

These results suggest that ChatGPT could be a valuable tool for novice researchers during the selection process, helping them when in doubt. It would be even more beneficial if the LLM explained each inclusion or exclusion.

• **Sensibility of prompts and convergence of responses –** When using LLMs, it is essential to have a deeper understanding of how different factors interplay with their adoption. Based on our preliminary results, we conjecture that various factors may influence the outcomes. For example, since SE is inherently context-dependent, adopting LLMs in the area could be enhanced by training models on domain-specific data. Moreover, LLMs used to be sensitive to subtle changes in prompt formatting. Given the prompt's fragility, we intend to understand better the factors that impact prompts on LLMs (order sensitivity, context length) and find techniques that make them more accurate and reliable.

Moreover, LLMs are non-deterministic and provide different answers to the same 'question' when asked multiple times. Therefore, several rounds may help SE researchers identify the point of convergence. However, given the current limitation of LLM-based tools in helping navigate the solution space, researchers often limit themselves to selecting the first-round solution, missing the opportunity to have a more comprehensive view of the accuracy. We recognize these gaps and will investigate factors that impact using LLMs to support SE research and empower them to navigate the LLM space better.

To address these points, we set the ChatGPT temperature parameter to 0 and top_p to 0.1 to prioritize a moderate level of randomness and diversity while maintaining a degree of control and accuracy. However, more studies are needed to understand better configuration settings of LLMs to explain the advantages and disadvantages of changing these settings and how they impact LLM output in selecting studies.

LLM performance and data processing capabilities are also directly related to other parameters, such as the context window and the max tokens. The context window equals the number of words an LLM can process simultaneously. In our research, the context window was not a barrier since we prompted one study at a time. However, this parameter cannot be ignored if we consider the final selection, in which the full text is evaluated.

The max tokens parameter sets the limit for the total number of tokens, including the input fed to an LLM as a prompt and the output tokens produced by an LLM in response to such a prompt. Again, we did not observe any limitations regarding our study since only

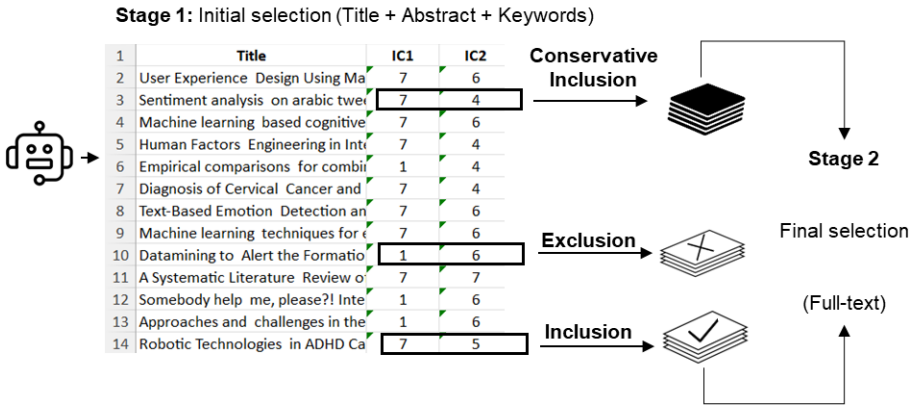

Figure 2: In a "Conservative" inclusion perspective, articles with rates 4 (neither agree nor disagree) in one of the criteria are included.

the title, abstract, and keywords were provided as input. However, this parameter should be investigated for other SLR activities.

• **LLMs will improve over time** – We expect our research to improve as the area evolves. The SE community needs to engage in a dialogue about the emerging capabilities of LLMs. This discussion should focus on the current state and anticipate future developments. By doing so, we can better understand how to integrate these advanced tools into our research methodologies, including SLRs, ensuring that they complement rather than replace SE researchers.

• **Classification activity** – To better elaborate on the capabilities of ChatGPT, we assigned it the responsibility of classifying articles into specific predefined categories using a 1–7 Likert scale. Articles with a rate greater than or equal to five (5 – somewhat agree) for each inclusion criterion were included. However, other Likert scales or a boolean choice (included or excluded) could be adopted. The accuracy of these variations of categories was not analyzed and merited future attention.

• **SLR reproducibility** – Since we defined a prompt to input ChatGPT, the same procedure can be replicated multiple times by following the same guidelines, enhancing the reproducibility of the selection method. Moreover, Huotala et al. [41] recommend future SLRs publish replication packages with selection data to enable more conclusive experimenting regarding ChatGPT to support the selection process.

• **Ethical considerations** – Ethical considerations need to regulate using ChatGPT as a reviewer in SLRs in SE to safeguard the interests of the research community and the population using the SLR results.

4.2 Future work

In future work, we plan to i) identify critical factors influencing researchers adopting LLMs to support their SLRs; ii) identify features that impact the LLMS prompts regarding order sensitivity and context length; iii) determine the optimal number of rounds to convergence of LLMs responses; iv) assess the accuracy of LLM responses to support the SLR process as a whole; and v) investigate how the LLM accuracy varies across different SE domains.

Efforts also could be made to develop visual or graphical representations to comprehend ChatGPT's output. The visualizations, which are easier for researchers to understand, can improve the interpretability of ChatGPT's responses. For example, Felizardo et al. [19, 21, 25], researched analytical techniques that employ visualization (e.g., Visual Text Mining – VTM) to support different activities of the SLR process, particularly the selection of studies.

Moreover, although LLM may facilitate and expedite SLRs, how to integrate it into the review process is unclear. In prospective work, we plan to propose a framework for integrating LLMs into the SLR process, including the rationale for LLM use, activities definition, prompt templates, data entries, human role, and metrics.

Furthermore, for SLRs, transparency of method and assessment rigor are essential, and these features are absent from ChatGPT classifications. The use of ChatGPT to eliminate studies automatically has not yet been fully proven. It may be used with care in clinical areas, but more developmental and evaluative research is needed in SE. As future work, other LLMs, e.g., Gemini, that show the reasoning for each inclusion or exclusion could be adopted to mitigate the low level of confidence issue.

By recognizing the strengths and limitations of LLMs and taking appropriate measures to enhance their accuracy, SE researchers can maximize the benefits of AI in selecting studies. We invite the SE community to discuss the challenges and potential research opportunities in leveraging LLMs to support their SLRs.

4.3 Threats to Validity

Our goal was to investigate the adoption of LLMs to aid the selection activity using evidence from two existing SLRs and corresponding replication rather than carrying out a formal experiment with a simulated scenario or re-doing a few SLRs ourselves. Therefore, our analysis of selecting studies with ChatGPT support was demonstrated by evidence gathered from two replications, which can be seen as a threat to the conclusion's validity. These two particular SLRs were selected by criteria directed toward the data (list of included and excluded studies and selection criteria) necessary to run the replications. However, the limited number of validations means that results may not represent the full spectrum of the SE domain.

Furthermore, there is a risk that the individual selection becomes biased. Even using selection criteria, researchers have decided whether to select an article based on their interpretation. Although we considered the experts' opinions to be benchmarks, we cannot say that all of them are correct. Overall, the articles were judged according to the SLR's design and the predefined criteria used in the evaluation to help minimize the validity threats to the conclusion.

We considered only inclusion criteria. In the future, we will investigate how ChatGPT behaves concerning the exclusion criteria as "a conference article has a more recent journal version, then the journal article is included, and the conference article is excluded", or ' 'only the most recent version of the article is included if multiple updates for studies are found". The entire collection of documents must be provided to ChatGPT to consider these exclusion criteria.

Even though ChatGPT was the state-of-the-art model in this writing, other LLMs may show different accuracy in selecting studies. However, as a first-cut assessment of the use of ChatGPT to support the selection of studies in SLRs in SE, we believe our study met its goal.

5 FINAL REMARKS

The answer to the RQ "*To what extent is ChatGPT able to support the SLR process, especially concerning the selection activity?*" is that ChatGPT's support is limited as of now, but it can filter with an accuracy greater than 75%.

The advancement of fundamental AI methods, particularly the latest LLMs, provides opportunities for SLR-AI. However, significant concerns exist about the accuracy of the studies selected by AI methods. This study provides the first comprehensive overview of the accuracy of adopting LLMs to aid SLR. Research on LLMs' accuracy in SE is still in its early stages, and there are challenges in adapting prompt, defining metrics, and evaluating its use in the SE domain. Future research efforts should focus on tackling these challenges.

Therefore, the contributions of this study encompass multiple areas, including:

(1) We presented an empirical investigation of selecting studies by leveraging the power of LLMs. We defend that by combining the strengths of human expertise and AI capabilities, we could streamline the traditional SLR process, maintaining its accuracy.
(2) We compared the accuracy of utilizing ChatGPT for selecting studies to be included in an SLR, in contrast to conventional methods.
(3) We applied LLM in two SLRs, and our findings highlight the potential application of LLM in selecting articles, revealing valuable insights into the current research landscape and shedding light on future directions in the domains of SE. There are promising avenues for future research to fully explore LLMs' capabilities to aid other SLR activities, investigate its limitations in diverse SE contexts, and support other SLR activities and secondary study types.

Finally, it is essential to remember that ChatGPT cannot replace the experience of researchers but can support them by selecting studies. SE researchers should consider the ChatGPT a complementary "opinion" to avoid biases and inaccuracies in the included evidence, resulting in fair results.

6 ONLINE RESOURCES

Supplementary materials are available on https://figshare.com/s/f783c33c13e7cdb84

ACKNOWLEDGMENTS

Professor Katia Romero Felizardo is funded by a research grant from the Brazilian National Council for Scientific and Technological Development (CNPq), Grant 302339/2022 – 1.

This research was carried out within the Samsung-UFAM Project for Education and Research (SUPER) scope, according to Article 48 of Decree number 6.008/2006 (SUFRAMA). We also thank the financial support granted by CNPq through processes 314797/2023–8 and 443934/2023–1. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) – Finance Code 001. In addition, this work was partially supported by FAPEAM through the POSGRAD 22-23 project.

Finally, the present work is the result of the Research and Development (R&D) project 001/2020, signed with UFAM and FAEPI, Brazil, which has funding from Samsung, using resources from the Informatics Law for the Western Amazon (Federal Law nº 8.387/1991). Its disclosure is under article 39 of Decree No. 10.521/2020.

All authors have contributed to all activities related to the submission.

REFERENCES

[1] R. Abilio, G. Vale, D. Pereira, C. Oliveira, F. Morais, and H. Costa. 2014. Systematic literature review supported by information retrieval techniques: A case study. In *40th Latin American Computing Conference (CLEI'14)*. IEEE, Montevideo, Uruguay, 1–11.
[2] J. J. G. Adeva, J. M. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa. 2014. Automatic text classification to support systematic reviews in medicine. *Expert System Application* 4, 41 (2014), 1498–1508.
[3] A. Al-Zubidy and J. C. Carver. 2014. *Review of Systematic Literature Review Tools – Technical Report SERG–2014-03.* Technical Report. University of Alabama.
[4] A. Al-Zubidy and J. C. Carver. 2019. Identification and prioritization of SLR search tool requirements: an SLR and a survey. *Empirical Software Engineering* 1, 24 (2019), 139–169.
[5] A. Al-Zubidy, J. C. Carver, D. P. Hale, and E. E. Hassler. 2017. Vision for SLR tooling infrastructure: Prioritizing value-added requirements. *Information and Software Technology* 2017, 91 (2017), 72–81.
[6] R. Alchokr, M. Borkar, S. Thotadarya, G. Saake, and T. Leich. 2023. Supporting Systematic Literature Reviews Using Deep-Learning-Based Language Models. In *1st Workshop on Natural Language-based Software Engineering (NLBSE'22)*. ACM, Pittsburgh, Pennsylvania, 67—-74.
[7] N. Ali and B. Tanveer. 2022. A Comparison of Citation Sources for Reference and Citation-Based Search in Systematic Literature Reviews. *e-Informatica Software Engineering Journal* 16, 1 (2022), 220106.
[8] N. Ali and M. Usman. 2019. Automatic text classification to support systematic reviews in medicine. *Information and Software Technology* 12, 1 (2019), 48–50.
[9] M. B. Aliyu, R. Iqbal, and A. James. 2018. The Canonical Model of Structure for Data Extraction in Systematic Reviews of Scientific Research Articles. In *15th Conference on Social Networks Analysis, Management and Security (SNAMS'18)*. IEEE, Valencia, Spain, 264–271.
[10] A. Alshami, M. Elsayed, E. Ali, A. E. E. Eltoukhy, and T. Zayed. 2023. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* 11, 7 (2023), 1–7.
[11] A. Anghelescu, F. C. Firan, G. Onose, C. Munteanu, A. Trandafir, I. Ciobanu, S. Gheorghita, and V. Ciobanu. 2023. PRISMA Systematic Literature Review, including with Meta-Analysis vs. ChatbotGPT (AI) regarding Current Scientific Data on the Main Effects of the Calf Blood Deproteinized Hemoderivative Medicine (Actovegin) in Ischemic Stroke. *Biomedicines* 6, 11 (2023), 1–13.
[12] A. Bannach-Brown, P. Przybyla, J. Thomas, A. S. C. Rice, S. Ananiadou, J. Liao, and M. Macleod. 2019. Machine learning algorithms for systematic review:

reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews* 1, 8 (2019), 23.

[13] D. Bowes, T. Hall, and S. Beecham. 2012. SLuRp: a tool to help large complex systematic literature reviews deliver valid and rigorous results. In *2nd Workshop on Evidential Assessment of Software Technologies (EAST'12)*. ACM, Lund, Sweden, 33–36.

[14] J. C. Carver, E. Hassler, E. Hernandes, and N. A. Kraft. 2013. Identifying barriers to the systematic literature review process. In *7th Symposium on Empirical Software Engineering and Measurement (ESEM'13)*. ACM, Baltimore, Maryland, US, 203–213.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[16] V. dos Santos, A.Y. Iwazaki, K.R. Felizardo, E.F. de Souza, and E. Y. Nakagawa. 2021. Towards Sustainability of Systematic Literature Reviews. In *15th Symposium on Empirical Software Engineering and Measurement (ESEM'21)*. ACM, Bari, Italy, 1–6.

[17] S.C.P.F. Fabbri, E. Hernandes, A. Di Thommazo, A. Belgamo, A. Zamboni, and C. Silva. 2012. Using Information Visualization and Text Mining to Facilitate the Conduction of Systematic Literature Reviews. In *14th Conference on Enterprise Information Systems (ICEIS'12)*. Springer, Wroclaw, Poland, 243–256.

[18] S.C.P.F. Fabbri, C. Silva, E. Hernandes, F. Octaviano, A. Di Thommazo, and A. Belgamo. 2016. Improvements in the StArt Tool to Better Support the Systematic Review Process. In *20th International Conference on Evaluation and Assessment in Software Engineering (EASE'16)*. ACM, Limerick, Ireland, 1–5.

[19] K. R. Felizardo, G. F. Andery, F. V. Paulovich, R. Minghim, and J. C. Maldonado. 2014. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology* 10, 54 (2014), 1079–1091.

[20] K. R. Felizardo and J. C. Carver. 2020. *Automating Systematic Literature Review*. Springer International Publishing, New York, US, Chapter 11, 327–355.

[21] K. R. Felizardo, S. G. MacDonell, E. Mendes, and J. C. Maldonado. 2012. A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews. *journal of Software* 2, 7 (2012), 450–461.

[22] K. R. Felizardo, E. Y. Nakagawa, S. C. P. F. Fabbri, and F. C. Ferrari. 2017. *Systematic literature review in software engineering: theory and practice (in portuguese)* (1 ed.). Elsevier, Brazil.

[23] K. R. Felizardo, E. Y. Nakagawa, S. G. MacDonell, and J. C. Maldonado. 2014. A visual analysis approach to update systematic reviews. In *18th Conference on Evaluation and Assessment in Software Engineering (EASE'14)*. ACM, London, England, UK, 1–4.

[24] K. R. Felizardo, E. Y. Nakwgawa, D. Feitosa, R. Minghim, and J. C. Maldonado. 2010. An approach based on visual text mining to support categorization and classification in the systematic mapping. In *14th Conference on Evaluation and Assessment in Software Engineering (EASE'10)*. ACM, UK, 1–10.

[25] K. R. Felizardo, M. Riaz, M. Sulayman, E. Mendes, S. G. MacDonell, and J. C. Maldonado. 2011. Analysing the use of graphs to represent the results of systematic reviews in software engineering. In *25th Brazilian Symposium on Software Engineering (SBES'11)*. IEEE, Brazil, 174–183.

[26] K. R. Felizardo, N. Salleh, R. M. Martins, E. Mendes, S. G. MacDonell, and J. C. Maldonado. 2012. Using visual text mining to support the study selection activity in systematic literature reviews. In *5th Software Engineering and Measurement (ESEM'12)*. ACM, Lund, Sweden, 1–10.

[27] K. R. Felizardo, S. H. Takemiya, and E. F. Souza. 2017. Analyzing the use of graphical abstracts to support study selection in secondary studies. In *Experimental Software Engineering (ESELAW'17)*. Springer, Buenos Aires, Argentina, 1–10.

[28] L. Feng, Y. Chiam, E. R. M. F. Abdullah, and U. Obaidellah. 2017. Using suffix tree clustering method to support the planning phase of systematic literature review. *Malaysian journal of Computer Science* 4, 30 (2017), 311–332.

[29] A. M. Fernández-Sáez, M. Genero, and F. P. Romero. 2010. SLR-Tool – a tool for performing systematic literature reviews. In *5th Conference on Software and Data Technologies (ICSOFT'10)*. Springer, Athens, Greece, 157–166.

[30] M. Ghafari, M. Saleh, and T. Ebrahimi. 2012. A federated search approach to facilitate systematic literature review in software engineering. *International journal of Software Engineering & Applications* 2, 3 (2012), 1–13.

[31] S. González-Toral, R. Freire, R. Gualán, and V. Saquicela. 2019. A ranking-based approach for supporting the initial selection of primary studies in a Systematic Literature Review. In *XLV Latin American Computing Conference (CLEI'19)*. IEEE, Panama City, Panama, 1–10.

[32] S. Götz. 2018. Supporting Systematic Literature Reviews in Computer Science: The Systematic Literature Review Toolkit. In *21st Conference on Model Driven Engineering Languages and Systems: Companion Proceedings (MODELS'18)*. ACM, Copenhagen, Denmark, 22–26.

[33] S. Groppe and L. Hartung. 2020. ReViz: A Tool for Automatically Generating Citation Graphs and Variants. In *Digital Libraries at Times of Massive Societal Transition*, E. Ishita, N.L.S. Pang, and L. Zhou (Eds.). Springer, Kyoto, Japan, 107–121.

[34] E. Guo, M. Gupta, J. Deng, Y. J. Park, M. Paget, and C. Naugler. 2024. Automated Paper Screening for Clinical Reviews Using Large Language Models – Data Analysis Study. *Journal of Medical Internet Research* 1, 26 (2024), e48996.

[35] R. Gupta, J. B. Park, C. Bisht, I. Herzog, J. Weisberger, J. Chao, K. Chaiyasate, and E. S. Lee. 2023. Expanding Cosmetic Plastic Surgery Research With ChatGPT. *Aesthetic Surgery Journal* 8, 43 (2023), 930–937.

[36] E. Hassler, Carver, J. C., D. Hale, and A. Al-Zubidy. 2016. Identification of SLR tool needs -– results of a community workshop. *Information and Software Technology* 70 (2016), 122–129.

[37] E. Hassler, J.C. Carver, N. A. Kraft, and D. Hale. 2014. Outcomes of a community workshop to identify and rank barriers to the systematic literature review process. In *18th Conference on Evaluation and Assessment in Software Engineering (EASE'14)*. ACM, London, England, UK, 1–10.

[38] M. A. Hearst. 1993. *ExtTiling: A Quantitative Approach to Discourse Segmentation – Technical Report 93/24.* Technical Report. University of California.

[39] E. Hernandes, A. Zamboni, A. Thommazo, and S. C. P. F. Fabbri. 2012. Using gqm and tam to evaluate start – a tool that supports systematic review. *CLEI Eletronic journal* 1–2012, 15 (2012), 1–13.

[40] A. Hinderks, F.J.D. Mayo, J. Thomaschewski, and M.J. Escalona. 2020. An SLR-Tool: Search Process in Practice: A Tool to Conduct and Manage Systematic Literature Review (SLR). In *42nd Conference on Software Engineering: Companion Proceedings (ICSE-Companion – ICSE'20)*. IEEE, Seoul, Korea (South), 81—-84.

[41] A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä. 2024. The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews. In *28th International Conference on Evaluation and Assessment in Software Engineering (EASE'24)*. ACM, Salerno, Italy, 1–10.

[42] H. Jiang, X. Zhang, X. Cao, J. Kabbara, and D. Roy. 2023. PersonaLLM: Investigating the ability of GPT-3.5 to express personality traits and gender differences.

[43] S. Jonnalagadda, P. Goyal, and M. Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews* 4:78, 4 (2015), 1–16.

[44] T. Karakose. 2023. The Utility of ChatGPT in Educational Research-Potential Opportunities and Pitfalls. *Educational Process: International Journal* 2, 12 (2023), 7–13.

[45] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield. 2024. 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* 1, 1 (2024), 1–11.

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

[47] C. Kohl, E.J. McIntosh, S. Unger, N.R. Haddaway, S. Kecke, J. Schiemann, and R. Wilhelm. 2018. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environmental Evidence* 8, 7 (2018), 7–15.

[48] S. Laghrabli, L. Benabbou, and A. Berrado. 2015. A new methodology for literature review analysis using association rules mining. In *10th Conference on Intelligent Systems: Theories and Applications (SITA'15)*. IEEE, Rabat, Marocco, 1–6.

[49] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[50] C. Lausberger. 2017. *Konzeption von suchprozessen und suchstrategien für systematische literatur reviews (in German).* Master's thesis. Otto-von-Guericke-University Magdeburg.

[51] C. Lausberger. 2017. *Supporting study selection of systematic literature reviews in software engineering with text mining.* Master's thesis. University of Oulu.

[52] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Survey* 55, 9, Article 195 (2023), 35 pages. https://doi.org/10.1145/3560815

[53] S. A. Mahuli, A. Rai, A. V. Mahuli, and A. Kumar. 2023. Application ChatGPT in conducting systematic reviews and meta-analyses. *British Dental Journal* 235, 2 (2023), 90–92.

[54] V. Malheiros, E. Hohn, R. Pinho, M. Mendonça, and J.C. Maldonado. 2007. A visual text mining approach for systematic reviews. In *1st Symposium on Empirical Software Engineering and Measurement (ESEM'07)*. ACM, Madrid, Spain, 245–254.

[55] C. Marshall, O. P. Brereton, and B. A. Kitchenham. 2014. Tools to support systematic reviews in software engineering: A feature analysis. In *18th Conference on Evaluation and Assessment in Software Engineering (EASE'14)*. ACM, London, England, UK, 13:1–13:10.

[56] C. Marshall, B. A. Kitchenham, and O. P. Brereton. 2018. Tool features to support systematic reviews in software engineering – A Cross Domain Study. *e-Informatica Software Engineering journal* 1, 12 (2018), 79–115.

[57] G. D. Mergel, M. S. Silveira, and T. S. da Silva. 2015. A method to support search string building in systematic literature reviews through visual text mining. In *30th Annual ACM Symposium on Applied Computing (SAC'15)*. ACM, Salamanca, Spain, 1594–160.

[58] J. S. Molléri and F. B. V. Benitti. 2015. SESRA: a web-based automated tool to support the systematic literature review process. In *19th Conference on Evaluation and Assessment in Software Engineering (EASE'15)*. ACM, Nanjing, China, 1–10.

[59] C. Muñoz Caro, A. Niño, and S. Reyes. 2017. A bibliometric approach to systematic mapping studies: The case of the evolution and perspectives of community detection in complex networks. arXiv preprint arXiv:1702.02381.

[60] D. Najafali, J. M. Camacho, E. Reiche, L. G. Galbraith, S. D. Morrison, and A. H. Dorafshar. 2023. Truth or Lies? The Pitfalls and Limitations of ChatGPT in Systematic Review Creation. *Aesthetic Surgery Journal* 43, 8 (2023), NP654–NP655.

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

[62] J. L. Neto, A. D. Santos, C. A. A. Kaestner, and A. Freitas. 2000. Generating text summaries through the relative importance of topics. *Advances in Artificial Intelligence, IBERAMIA 2000* 1952, Lecture Notes in Computer Science (2000), 300–309.

[63] L.J. Neto, A.D. Santos, C.A.A. Kaestner, and A.A. Freitas. 2000. Generating Text Summaries through the Relative Importance of Topics. In *Advances in Artificial Intelligence*, M.C. Monard and J.S. Sichman (Eds.). Springer Berlin Heidelberg, Berlin, Germany, 300–309.

[64] F. Octaviano, C. Silva, and S.C.P.F. Fabbri. 2016. Using the SCAS Strategy to Perform the Initial Selection of Studies in Systematic Reviews: An Experimental Study. In *20th International Conference on Evaluation and Assessment in Software Engineering (EASE'16)*. ACM, Limerick, Ireland, Article 25, 10 pages.

[65] F. R. Octaviano, K. R. Felizardo, J. C. Maldonado, and S. C. P. F. Fabbri. 2016. Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? *Empirical Software Engineering* 6, 20 (2016), 1898–1917.

[66] B. K. Olorisade, E. de Quincey, P. Brereton, and P. Andras. 2016. A Critical Analysis of Studies That Address the Use of Text Mining for Citation Screening in Systematic Reviews. In *20th International Conference on Evaluation and Assessment in Software Engineering (EASE'16)*. ACM, Limerick, Ireland, 1–11.

[67] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 1, 4 (2015), 1–5.

[68] F. Osborne, H. Muccini, P. Lago, and E. Motta. 2018. Reducing the effort for systematic reviews in software engineering. https://research.vu.nl/en/publications/reducing-the-effort-for-systematic-reviews-in-software-engineering.

[69] M. Pessoa, M. Lima, F. Pires, G. Haydar, R. Melo, L. Rodrigues, D. Oliveira, E. Oliveira, L. Galvão, B. Gadelha, et al. 2023. A Journey to Identify Users' Classification Strategies to Customize Game-Based and Gamified Learning Environments. *IEEE Transactions on Learning Technologies* 1, 17 (2023), 527–541.

[70] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64 (2015), 1–18.

[71] N. Pulsiri and R. Vatananan-Thesenvitz. 2018. Improving Systematic Literature Review with Automation and Bibliometrics. In *Portland International Conference on Management of Engineering and Technology (PICMET'18)*. IEEE, Honolulu, Hawaii, USA, 1–8.

[72] R. Qureshi, D. Shaughnessy, K. A. R. Gill, K. A. Robinson, T. Li, and E. Agai. 2023. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Review* 12, 72 (2023), 1–4.

[73] H. Ramampiaro, D. Cruzes, R. Conradi, and R. Mendona. 2010. Supporting evidence-based Software Engineering with collaborative information retrieval. In *6th Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'10)*. IEEE, Chicago, IL, USA, 1–5.

[74] R. Rasmus, E. Bjarnason, and P. Runeson. 2017. A machine learning approach for semi-automated search and selection in literature studies. In *21st Conference on Evaluation and Assessment in Software Engineering (EASE'17)*. ACM, Karlskrona, Sweden, 1–10.

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

[76] R. Ros, E. Bjarnason, and P. Runeson. 2017. A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies. In *21st International Conference on Evaluation and Assessment in Software Engineering (EASE'17)*. ACM, Karlskrona, Sweden, 118–127.

[77] V. Santos. 2018. Concept maps construction using natural language processing to support studies selection. In *33rd Annual ACM Symposium on Applied Computing (SAC'18)*. ACM, Pau, France, 926–927.

[78] V. Santos, A. Iwazaki, E.F. Souza, K.R. Felizardo, and N.L. Vijaykumar. 2021. CrowdSLR: A Tool to Support the Use of Crowdsourcing in Systematic Literature Reviews. In *XXXV Brazilian Symposium on Software Engineering (SBES'21)*. ACM, Joinville, Brazil, 341—-346.

[79] Y. Shakeel, J. Krüger, I.V. Nostitz-Wallwitz, G. Saake, and T. Leich. 2019. Automated Selection and Quality Assessment of Primary Studies: A Systematic Literature Review. *Journal of Data and Information Quality* 12, 1, Article 4 (2019), 26 pages.

[80] Y. Shakeel, J. Krüger, I. v. Nostitz-Wallwitz, C. Lausberger, G. C. Durand, G. Saake, and T. Leich. 2018. (Automated) Literature Analysis – Threats and Experiences. In *13th Workshop on Software Engineering for Science (SE4Science'18)*. ACM, Gothenburg, Sweden, 20–27.

[81] G. Silva, P. Santos Neto, R. Santos Moura, A.C. Araújo, O. Cury da Costa Castro, and I. Ibiapina. 2019. An Approach to Support the Selection of Relevant Studies in Systematic Review and Systematic Mappings. In *8th Brazilian Conference on Intelligent Systems (BRACIS'19)*. IEEE, Salvador, BA, Brazil, 824–829.

[82] M. C. R. Silva. 2009. *Contextextractor: uma ferramenta de apoio para a extração de informações de contexto de artigos de engenharia de software experimental (in portuguese)*. Master's thesis. Universidade Salvador.

[83] E.M. Silva Junior and M.L. Dutra. 2021. A roadmap toward the automatic composition of systematic literature reviews. *Iberoamerican Journal of Science Measurement and Communication* 1, 2 (2021), 1–22.

[84] P. Singh, M. Galster, and K. Singh. 2018. How do secondary studies in software engineering report automated searches?. In *22nd Conference on Evaluation and Assessment in Software Engineering (EASE'18)*. ACM, Christchurch, New Zealand, 145–150.

[85] Paramvir Singh and Karanpreet Singh. 2017. Exploring Automatic Search in Digital Libraries: A Caution Guide for Systematic Reviewers. In *21st Conference on Evaluation and Assessment in Software Engineering (EASE'17)*. ACM, Karlskrona, Sweden, 236–241.

[86] M. Soundefinednicki and L. Madeyski. 2021. ASH: A New Tool for Automated and Full-Text Search in Systematic Literature Reviews. In *21nd Conference Computational Science (ICCS'21)*. Springer, Kraków, Poland, 362—-369.

[87] F. C. Souza, A. Santos, S. Andrade, R. Durelli, V. Durelli, and R. Oliveira. 2017. Automating search strings for secondary studies. Information Technology – New Generations. In *Part of the Advances in Intelligent Systems and Computing book series (AISC'17)*. Springer, Lviv, Ukraine, 839–848.

[88] D. Stefanovic, S. Havzi, D. Nikolic, D. Dakic, and T. Lolic. 2021. Analysis of the Tools to Support Systematic Literature Review in Software Engineering. *IOP Conference Series: Materials Science and Engineering* 1163, 1 (2021), 012013.

[89] Y. Sun, Y. Yang, H. Zhang, W. Zhang, and Q. Wang. 2012. Towards evidence-based ontology for supporting Systematic Literature Review. In *16th Conference on Evaluation Assessment in Software Engineering (EASE'12)*. ACM, Ciudad Real, Spain, 171–175.

[90] P. Tell, J. B. Cholewa, P. Nellemann, and M. Kuhrmann. 2016. Beyond the Spreadsheet: Reflections on Tool Support for Literature Studies. In *20th International Conference on Evaluation and Assessment in Software Engineering (EASE'16)*. ACM, Limerick, Ireland, 5 pages.

[91] F. Tomassetti, G. Rizzo, A. Vetro, L. Ardito, and M. Torchiano, M. Morisio. 2011. Linked data approach for selection process automation in systematic reviews. In *15th Conference on Evaluation and Assessment in Software Engineering (EASE'11)*. IET-IEEE, Durham, UK, 31–35.

[92] J. A. S. Torres, D. S. Cruzes, and L. Salvador. 2012. Automatic results identification in software engineering papers. Is it possible?. In *12th Conference on Computer Science and Its Applications*. Springer, Cheju Island, Korea, 108–112.

[93] G. H. Travassos, K. R. Felizardo, M Morandini, and C. Kolski. 2017. *A Tertiary Study on the Convergence of Human-Computer Interaction and Artificial Intelligence* (1 ed.). A Springer book series Learning and Analytics in Intelligent Systems, New York, NY, USA.

[94] R. van Dinter, B. Tekinerdogan, and C. Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology* 136 (2021), 106589.

[95] D. Voigt, O. Kopp, and K. Wild. 2021. Systematic Literature Tools: Are we there yet?. In *Central-European Workshop on Services and their Composition*. CEUR-WS, Kyiv, Ukraine, 1–10.

[96] S. Wang, H. Scells, B. Koopman, and G. Zuccon. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*. ACM, Taipei, Taiwan, 1426—-1436.

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

[98] M. Waseem, A. Ahmady, P. Liangz, M. Fehmidehx, P. Abrahamsson, and T. Mikkonen. 2023. Conducting Systematic Literature Reviews with ChatGPT. In *17th International Symposium on Empirical Software Engineering and Measurement (ESEM'23)*. ACM, New Orleans, Louisiana, USA, 1–10.

[99] W.M. Watanabe, K.R. Felizardo, A. Candido, E.F. de Souza, J.E.C. Neto, and N.L. Vijaykumar. 2020. Reducing efforts of software engineering systematic literature reviews updates using text classification. *Information and Software Technology* 128 (2020), 106395.

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

[101] L. H. Xuân-Lan and S. Thierry. 2023. Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in

Moderate-to-Severe Plaque Psoriasis. *Journal of Clinical Medicine* 12, 16 (2023), 5410.

[102] Z. Yu, N. A. Kraft, and T. Menzies. 2018. Finding better active learners for faster literature reviews. *Empirical Software Engineering* 6, 23 (2018), 3161–3186.

[103] Z. Yu and T. Menzies. 2019. Fast2: an intelligent assistant for finding relevant papers. *Expert Systems with Applications* 15, 120 (2019), 57–71.

[104] R. Zumaeta, E. Cuayla, and D. Mauricio. 2019. Automation Tool for the Planning Phase of the Systematic Review. In 7^{th} *International Engineering, Sciences and Technology Conference (IESTEC'19).* IEEE, Panama City, Panama, 280–285.

Received 20 February 2024; revised 12 March 2024; accepted 5 June 2024