

Data Extraction for Systematic Mapping Study using a Large Language Model

A Proof-of-Concept Study in Software Engineering

Katia Romero Felizardo*
katiascannavino@utfpr.edu.br
Universidade Tecnológica Federal do
Paraná – UTFPR/Brazil

Márcia Sampaio Lima
Universidade do Estado do Amazonas
– UEA/Brazil
mllima@uea.edu.br

Anderson Deizepe
Universidade Tecnológica Federal do
Paraná – UTFPR/Brazil
deizepeanderson@gmail.com

Tayana Uchôa Conte
Universidade Federal do Amazonas –
UFAM/Brazil
tayana@icomp.ufam.edu.br

Monalessa P. Barcellos
Federal University of Espírito Santo –
UFES/Brazil
monalessa@inf.ufes.br

Igor Steinmacher
Northern Arizona University/USA
Igor.Steinmacher@nau.edu

ABSTRACT

Context: Systematic mapping studies (SMS) are adopted in Software Engineering (SE) to select and synthesize relevant literature on a research topic and, thus, support evidence-based decision-making. Performing SMS is effort-demanding and time-consuming. Hence, using tools is beneficial. Large Language Models (LLMs) such as ChatGPT-4.0 can potentially accelerate repetitive activities, such as data extraction in SMS, saving time and effort. **Goal:** We conducted this work to evaluate and provide preliminary evidence on how ChatGPT-4.0 can support data extraction in SMS. **Method:** We performed a proof-of-concept study and assessed the results' accuracy of using ChatGPT 4.0 to extract data in one SMS compared to the results produced manually. **Results:** The accuracy of ChatGPT-4.0 was 87.83%. **Conclusions:** Our preliminary findings suggest that entirely replacing the manual data extraction with ChatGPT-4.0 is not recommended. However, employing ChatGPT for semi-automated data extraction to aid in evidence synthesis in SMS is promising.

CCS CONCEPTS

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

ACM Reference Format:

Katia Romero Felizardo, Márcia Sampaio Lima, Anderson Deizepe, Tayana Uchôa Conte, Monalessa P. Barcellos, and Igor Steinmacher. 2024. Data Extraction for Systematic Mapping Study using a Large Language Model: A Proof-of-Concept Study in Software Engineering. In *Proceedings of 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Evidence-Based Software Engineering (ESBE) provides knowledge about techniques, methods, and tools for developing quality software, which is the main objective of Software Engineering (SE). The construction of this knowledge is based on primary studies, such as surveys, case studies, and controlled experiments. These studies, although relevant, are not sufficient to generalize the results. Therefore, ESBE uses secondary studies, including systematic literature reviews (SLRs) and systematic mapping studies (SMSs), to identify, evaluate, and interpret all available research relevant to a particular research question, topic area, or phenomenon of interest [9]. Secondary studies are crucial to summarize evidence and provide a panorama of the state of the art on the researched topic. However, the current approach to conducting them is laborious, so the resulting evidence synthesis may be up to date when published [8].

Secondary studies involve several activities, such as formulating research questions (RQs), conducting comprehensive literature searches, critically selecting relevant studies, extracting data from the included studies, and synthesizing/categorizing evidence. Among them, extracting data is one of the most crucial, time-consuming, and costly since data are often manually extracted from the studies and recorded in tables [3]. The large scientific literature further adds to these challenges [21]. The number of articles included in secondary studies varies significantly, ranging from dozens to thousands. Extracting data from them requires considerable effort from the researchers. This activity has been the one with the least automation [4]. Fully automating data extraction is challenging due to the different ways SE researchers report results, restricted full-text access, or the lack of information provided by the authors. Another challenge is obtaining high-quality and accurate extracted data.

In secondary studies, data has been mostly extracted manually and through different strategies (e.g., a single researcher extracts the data, and a second one validates it) [9]. On average, a researcher spends 107 min per study, and dual-independent data extraction takes up to 172 min [10]. Moreover, data extraction bias can undermine the validity of evidence synthesis and the rate of data extraction errors up to 63%. Their causes are multifaceted, including missing available data, misclassifications, misinterpretations

stemming from ambiguous reporting in primary studies, or straightforward data entry mistakes. Time constraint is another factor that can enhance the risk of data extraction bias [10].

Recent studies suggest the large language models (LLMs) capability to enhance efficiency in this context, with superior performance when compared to older AI methods [1, 2, 6, 12, 16, 19, 20, 22]. LLMs can automate various tasks in natural language processing, understand context, and predict semantic relationships. Among the largest and most effective LLMs is the GPT-4o, the latest GPT model incorporated in ChatGPT. LLMs have emerged as a potential approach to aid in secondary studies.

In this work, we take a step towards this approach by investigating the use of ChatGPT-4.0 to support data extraction and providing preliminary evidence to seed questions that can be explored in future research. For that, we performed a proof-of-concept to *assess the accuracy of ChatGPT-4.0 (compared to data extracted by humans) in extracting data from the PDF full-text of primary studies selected in an SMS*.

Our study achieved high accuracy in extracting more simple and objective data. Accuracy degraded when referring to more complex data (non-straightforward information). ChatGPT exhibited an overall 87.83% accuracy in extracting data from 25 studies selected in the considered SMS. The accuracy in extracting bibliometric data (i.e., objective data) was 99.7%, and in data related to the RQs (i.e., non-straightforward information data) was 65.11%. Our study demonstrates the potential of employing ChatGPT-4.0 to support data extraction in SMS in SE. In summary:

What is already known? Data extraction is key for evidence categorization in SMS, but it is labor-intensive and error-prone [4, 8, 10, 21]. There has been a growing effort to automate data extraction based on natural language processing, language models, and recently, LLMs [1, 2, 6, 12, 16, 19, 20, 22].

What is new? Our study is the first to demonstrate the accuracy of ChatGPT in data extraction for SMS in ES.

What is the potential impact on SE researchers? This work offers preliminary evidence that ChatGPT is a promising tool to aid in data extraction in SMS in SE and shines a light on the need for future research to understand ChatGPT capabilities and limitations in this context.

2 METHODOLOGY

Our proof-of-concept involved three main steps: planning, data collection, and interpretation. In the first one, we established the study goal, selected the SMS to be considered, defined the data collection strategy and how to calculate accuracy. In the second step, we prepared the SMS data to be accessed by ChatGPT, defined the prompts, ran them, recorded data, and calculated accuracy. Lastly, we analyzed data, comparing the results produced by ChatGPT with the ones produced by humans. In the following, we present further information about the SMS used in the study, the prompts, and the accuracy calculation.

2.1 Replicated SMS

The SMS used in this study is [14]. It aimed to identify user profiles in games or gamified environments and evaluate the impact of game elements within these environments based on the users' profiles.

It answered three main RQs: **(RQ1)** What are user classification strategies commonly adopted in studies on the customization/personalization of games or gamified systems?; **(RQ2)** What and how are the instruments used to identify types of users? and **(RQ3)** How are the results evaluated using games and customized/personalized gamified systems?

In total, 448 papers were evaluated in the SMS, and, in the end, 25 passed the selection criteria and were selected for data extraction. We chose this SMS for convenience and two reasons: data extracted by the SMS authors was available, and the SMS was published in a high-quality journal. Data extraction was performed using a data extraction form to record relevant data to answer the RQs consistently. One SMS author extracted the data, verified by the other authors (quality assurance). Discussions occurred in meetings to solve doubts and double-check data (as suggested in [9]).

2.2 Prompting ChatGPT

To enable ChatGPT to extract data from the 25 included studies, we created a prompt to extract each data item, considering both bibliometric data and the RQs. Table 1 presents the defined prompts.

We created a Jupyter Notebook to extract data with ChatGPT and run it in the Google Colab environment. The Jupyter Notebook script receives `Top_p` as parameters, which determines the cumulative probability for choosing the most likely tokens, controlling diversity (zero (0) was used for more predictable and focused responses); Temperature adjusts the creativity of the response (zero (0) was used, for responses without creative variation); the ChatGPT model used (GPT-4o, most recent and with a context window of 128,000 tokens, the number necessary to support the full-text), and the OpenAI API key.

Given that the API does not have the functionality to receive PDF files via upload, we initially extracted the text content of each study in Python Tables and formulas, as long as they were described in text format (latex or word), were considered; however, OCR (Optical Character Recognition) was not used on images. Then, we sent the API to build the context window to extract the data. As the ChatGPT API is stateless, with each interaction, it is necessary to recreate the context (send the entire PDF content again) to respond to each data item (e.g., title, author, year, among others). As a result, the consumption of time and financial resources becomes significant. To mitigate the use of these resources, a strategy based on text formatted in pre-defined key and value (JSON, JavaScript Object Notation) – was used, in which each data item was previously assigned a key (e.g., “id”, ..., “psychological_mediators”) where all values (e.g., “1”, ..., “User Satisfaction”) were obtained in a single iteration per analyzed study (e.g. Paper S1.pdf), reducing execution time in 97% and costs in 95.7%. The JSON presented here:

```
{Paper S1.pdf
  "id": "1",
  "title": "User Experience Design Using ML: An SLR",
  "author1": "A.M.H. Abbas",
  "author2": "K.I. Ghauth",
  "author3": "C.-Y. Ting",
  "year": "2022",
  "source\_type": "Journal",
  "type\_application": "Machine Learning",
  "strategy": "Classification Algorithms",
  "instrument": "Surveys and Interviews",
  "data\_source": "User Interaction Data",
  "psychological\_mediators": "User Satisfaction"}
```

2.3 Accuracy Calculation

In this study, accuracy refers to the proportion of correctly extracted data. We considered the data extracted by SMS authors as our benchmark (oracle). Accuracy is calculated as $(TP + TN)/(TP + FP + TN + FN)$, where true positive (TP) is the number of data items ChatGPT correctly extracted; true negative (TN) is the number of data items ChatGPT correctly identified as unavailable in the full-text article; false positive (FP) is the number of hallucinated data (i.e., data items ChatGPT fabricated in cases where no data was available in the full-text article); and false negative (FN) is the number of data items missed or incorrectly extracted by ChatGPT. To calculate accuracy, one researcher classified data into the aforementioned types, and a second reviewed their correctness.

3 RESULTS

The results are presented in Tables 2, 3, 4, and 5. Table 2 shows data manually extracted by the SMS authors (i.e., oracle) and those extracted by ChatGPT to answer the RQs. Table 3 details the hits in data extraction, i.e., the number of data items correctly extracted (TP) and those correctly identified as unavailable (TN). Table 4 specifies the errors in data extraction, i.e., the number of data items missed or incorrectly extracted (FN) and the number of data items fabricated by ChatGPT (FP). Table 5 shows ChatGPT's accuracy in the study. In the tables, S1, S2, ..., S25 refer to the studies included in the SMS. Information about them is available in the SMS package.

3.1 Discussions

As shown in Table 5, the accuracy of ChatGPT considering all data items was 87.83%. In the papers where data was available, ChatGPT successfully extracted data in 321 occurrences (true positives) (see Table 3). However, ChatGPT fabricated 22 instances of data items (FP) (see Table 4). The accuracy was higher when ChatGPT extracted simple and objective data (98.6% when referring to bibliometric data) than when it extracted data related to the RQs (65.11%) (see Table 5). The accuracy of data extraction by humans is about 65% in single extraction (made by one researcher) and 75% in double extraction (made by one researcher and reviewed by another) [17]. Therefore, compared to human performance, the study results suggest that using ChatGPT to support data extraction in SMS is promising.

As outlined in Table 4, we identified 45 errors made by ChatGPT in data extraction. Data was incorrectly extracted on 15 occurrences, and ChatGPT missed available data in eight situations. Additionally, it seemed to have generated extra data in 22 cases (we did not find the referred data in the full text, i.e., ChatGPT hallucinated that data). In nine of these hallucination cases, data referring to human emotions were extracted, including “motivation” (4 occurrences), “personality type” (3), and “engagement” (3).

In the sequence, we detail and discuss the results considering the three RQs investigated in the SMS.

Concerning **RQ1** (*What are user classification strategies commonly adopted in studies on the customization/personalization of games or gamified systems?*), as summarized in Table 2, ChatGPT incorrectly extracted three interaction-based strategies (user types hexad framework [S5], brainHex gamer typology [S11, S12], two personality-based strategies (MBTI [S16] and five-factor model of

personality [S18]), and one learning style-based strategy (taxonomy [S22]) (see Table 2—column 3, RQ1).

We noticed variations in data extracted by ChatGPT referring to “User Types Hexad Framework” strategy, such as “Gamification User Type Hexad Model,” “Hexad User Types,” and “Hexad Model.” This occurs probably due to the lack of a unique term to designate that strategy. While some authors classify Hexad as a framework, others define it as a model. We argue that having a consensual and well-established terminology for specifying strategies and using such terminology in primary studies addressing the strategies could facilitate extracting related data in SMSs, mainly when performed by humans.

Strategies with acronyms were extracted in extensive or abbreviated form, such as “MBTI” or “Myers-Briggs type indicator.” Mixed forms (extensive and acronym in parentheses) were also observed by us, such as the “Folder-Silverman learning style model (FSLSM),” “learning tendencies (LTs),” and “Elliot achievement goal questionnaire-revised (AGQ-R).” The “five-factor model of personality” strategy was extracted from studies S19 and S20 as “personality traits.” The biggest mistake was the extraction of the “Naïve Bayes classifier” [S18] as a “user classification strategy.”

As for **RQ2** (*What and how are the instruments used to identify types of users?*), ChatGPT made five errors in data related to RQ2 [S1, S3, S14, S16, S21] (see Table 2—column 3, RQ2). It correctly extracted the “Hexad framework” as an instrument used to identify types of users in 6 out of 7 cases, failing in S3. Moreover, details about whether or not the framework was adapted were not considered. This “error” impacts the quality of the extracted data; however, it does not affect the SMS's conclusions. Likewise, without details about adaptations, ChatGPT got two MBTI extractions correct [S15 and S17] but was unsuccessful in one extraction [S16]. Relative to the FSLM model, one case was lost [S14], and one false data was created [S21]. Again, ChatGPT fabricated data unavailable in S1 (situational motivation scale – SIMS). These errors could lead to erroneous SMS conclusions. We observed that in most extractions (six), the type of instrument (e.g., questionnaire, scale) was extracted, as in S13, “BrainHex questionnaire.”

Regarding **RQ3** (*How are the results evaluated using games and customized/personalized gamified systems?*), unlike human classification (see Table 2—column 3, RQ3), ChatGPT considered that studies S5, S15, and S16 adopted a “questionnaire” as the data source to analyze the results. According to data extracted by humans, “log records” were the data source in these studies. For S19, ChatGPT also stated the adoption of a “questionnaire”; however, humans did not find such data. Therefore, ChatGPT faked this data. Surprisingly, in cases where two data sources were combined [S6, S7, S10, S11, S22, S23, and S4], ChatGPT was only unsuccessful in identifying the correct sources in two instances (S7 and S10). However, for both S7 and S10, ChatGPT did manage to identify one of the data sources correctly.

Considering psychological mediators, humans and ChatGPT agree in four circumstances [S7, S8, S12, S23] in which data was not extracted. On the other hand, ChatGPT fabricated data in 10 occurrences [S3, S9, S11, S15, S16, S17, S18, S20, S21, S25]. In seven cases, data was available in the PDFs but ChatGPT did not capture it (“preferences” [S10, S12, S19]; “motivation” [S1]; “enjoyment and usefulness” [S1, S10, S23]). SMS authors classified the data from

Table 1: Engineering prompts used to automate data item.

Information needs	Data item	Engineering prompt
Bibliometric data	Author’s last name	State the authors’ last name, styled as a proper noun with the first letter capitalized
	Articles’ year	State the article year
	Articles’ source	State the article source type
	Articles’ source name	State article source name, capitalize the first initials letters
	Articles’ source acronym	State the acronym source
SMS Context	Application domain	State the application domain
RQ1	Adopted user classification strategy adopted	State the name of the strategy(ies) used to classify users according to their preferences
RQ2	Instrument(s) used to identify types of users	State the instrument name used to identify user types (profiles)
RQ3	Data source used to measure the results	State the data source used to measure the results of customized or personalized games and gamified systems on students
	Physiological mediator	State the name of the physiological mediator used to evaluate the results of customized or personalized games and gamified systems on students
	Behavior mediator	State the name of the behavior mediator used to evaluate the results of customized or personalized games and gamified systems on students

S10 as “preferences”; however, ChatGPT extracted one data item and categorized it as “motivation.” We observed a recurrence of “engagement” categorization by ChatGPT (15 occurrences). A possible explanation is that although the “engagement” term is associated with investing physical, cognitive, and emotional energy into a specific task, in studies in Computer Science education, the term is generally not explicitly defined [7]. Furthermore, it is measured using indirect observation [13]. For behavioral mediators, only one data [S8] related to the learning outcomes category was missed; however, nine data were fabricated.

After analyzing results related to RQ1–RQ3, we observed that different factors can impact the adoption of ChatGPT in data extraction. For example, ChatGPT prompts are fragile, and they are sensitive to subtle changes in their formatting. Since previous research indicated that zero-shot learning tends to achieve better classification performance than few-shot learning in SLR [18], in the present study, we adopted a zero-shot prompt, i.e., we did not provide ChatGPT with examples of data that could be extracted. ChatGPT’s behavior when inputting one-shot or n-shot prompts (where n is the number of examples provided) is an open issue, and further research needs to be conducted to understand better the engineering prompt for extracting data for SMS. On one hand, not including examples reduces the researchers’ effort, on the other hand, extracting data from a few articles is not that difficult and could (or could not) improve how ChatGPT performs extractions (at least for specific data items).

Another challenge in using ChatGPT for data extraction is understanding and classifying ChatGPT errors and their consequences. For instance, in our view, missing available data and misclassification are errors that can affect the SMS’s conclusions. The source of ChatGPT errors also needs to be clarified, including whether the error was fabricated (ChatGPT hallucination) or a misinterpretation caused by imprecise reporting data in primary studies.

To minimize errors while a manual data extraction is conducted, two reviewers usually conduct independent processes. One of them extracts data, and the other verifies the extracted data. Disagreements are discussed between the two researchers or even involving

a third reviewer. In that respect, when performed by a single researcher, ChatGPT could be employed as the second review (e.g., differences in data extracted by the researcher and by ChatGPT are data to be reviewed by the single researcher), increasing the quality of extracted data.

3.2 Threats to the Validity

Some limitations of our study should be highlighted and considered with the results. First, the study considered a single SMS and a small sample of 25 primary studies (the ones included in the SMS). The SMS characteristics influence the results (e.g., research topic, RQs, selected papers). Therefore, representativeness is limited. Exploring other SMS and more articles with different data types and research questions is necessary and should be considered for future research.

Another important threat is that we considered data extracted by humans as our benchmark. However, manual extractions are error-prone. To minimize this threat, we selected an SMS that adopted good practices for data extraction. Data was extracted by one researcher and reviewed by others, and discussions were held until a consensus was reached.

Moreover, ChatGPT can generate different responses even when the exact prompt is presented multiple times. Although the inherent stochasticity of LLMs has a minimal impact on the accuracy of overall data extraction [5], further research is essential to investigate the impact of prompt rounds on data extraction accuracy in SE since prompt rounds may result in different responses. To mitigate this limitation, we set temperature and top_p parameters to zero to control the “creativity” (randomness) of the data extracted by ChatGPT-4.o.

4 RELATED WORK

Replacing manual data extraction from scientific articles with automated data extraction based on LLMs has been the focus of recent studies [1, 5, 11, 15, 17, 22]. Many researchers are optimistic that LLMs will soon become powerful data extraction tools.

Mahuli et al. [11] state that AI can assist by sharing the complete text and specifying the wanted information or data to be extracted. Hoai et al. [22] point out that LLMs can extract data similar to those

Table 2: Comparing data extracted manually with those automatically extracted by ChatGPT-4.o.

RQ	Human extraction	ChatGPT extraction
RQ1	Interaction-based strategies User types hexad framework [S2,S3,S4, S5 ,S6,S7,S8] BrainHex gamer typology [S9,S10, S11 , S12 ,S13] Multidimensional approach [S1] Player type [S14] Personality-based strategies MBTI [S15, S16 ,S17] Five-factor model of personality [S18 ,S19,S20] Learning style-based strategies Taxonomies [S12,S21, S22 ,S23,S24] Motivation-based strategies Achievement goal questionnaire-revised (AGQ-R) [S25]	Interaction-based strategies User types hexad framework [S2,S3,S4,S6,S7,S8] BrainHex gamer typology [S9,S10,S13] Multidimensional approach [S1] Player type [S14] Personality-based strategies MBTI [S15,S17] Five-factor model of personality [S19,S20] Learning style-based strategies Taxonomies [S12,S21,S23,S24] Motivation-based strategies AGQ-R [S25]
RQ2	User types Hexad framework User types Hexad framework, without any adaptation [S2,S6,S7,S8] User types Hexad framework, with adaptations [S3,S4,S5] BrainHex Gamer Typology BrainHex gamer typology [S9,S10,S11,S12,S13] Myers-Briggs Type Indicator (MBTI) MBTI in its completeness [S17] An adapted version of MBTI [S15, S16] Five Factor Model of Personality BFI-10 [S18] Big Five Inventory (BFI) [S19] iGFP-5 questionnaire [S20] Felder-Silverman Learning Style Model (FSLSM) FSLSM [S14,S22,S23,S24] Achievement Goal Questionnaire-Revised (AGQ-R) AGQ-R [S25]	User types Hexad framework User types Hexad framework [S2,S4,S5,S6,S7,S8] BrainHex Gamer Typology BrainHex gamer typology [S9,S10,S11,S12,S13] MBTI MBTI [S15,S17] Five Factor Model of Personality BFI-10 [S18] BFI [S19] iGFP-5 questionnaire [S20] FSLSM FSLSM [S21 ,S22,S23,S24] AGQ-R AGQ-R [S25] Situational Motivation Scale (SIMS) SIMS [S1]
RQ3	Data source used to analyze the results Consolidated instruments to measure user's motivation [S1,S6,S9,S21] User-specific questionnaires [S2,S8,S13,S18] Log records [S3,S4, S5 ,S12,S14, S15 , S16 ,S17,S20,S25] Questionnaires and log records [S6, S7 , S10 ,S11,S22,S23,S24] Psychological mediators Preferences [S4,S5, S10 , S12 , S19] Motivation [S1 ,S6,S14] Enjoyment and Usefulness [S1 ,S2, S10 ,S13,S22, S23 ,S24] Behavioral mediators – Distal Outcomes Performance [S6,S20,S24] Learning outcomes [S3,S4, S8 ,S11,S14,S15,S16,S17,S20,S25]	Data source used to analyze the results Questionnaires [S1,S2, S5 ,S6,S8,S9,S13, S15 , S16 ,S18, S19 ,S21] Log records [S3,S4,S12,S14,S17,S20,S25] Questionnaires and log records [S6,S11,S22,S23,S24] Psychological mediators Preferences [S4,S5] Motivation [S6, S10 ,S14] Enjoyment and Usefulness [S2,S13,S22,S24] Behavioral mediators – Distal Outcomes Performance [S6,S20,S24] Learning outcomes [S3,S4,S11,S14,S15,S16,S17,S20,S25]

Studies in bold, located in column 2, represent data extracted exclusively by humans.

Studies in bold and italic, located in column 3, represent data incorrectly extracted by ChatGPT.

Studies in bold and underlined, located in column 3, represent data created by ChatGPT, which did not exist in the PDF.

Table 3: ChatGPT hits in data extraction.

	Data correctly identified (TP)		Data correctly identified as unavailable (TN)	
	Human extraction	ChatGPT extraction	Human extraction	ChatGPT extraction
Bibliometric data	216	213	0	0
General data (SMS context)	25	22	0	0
RQ1	26	20	0	0
RQ2	23	20	0	0
RQ3: data source	25	20	2	0
RQ3: psychol. mediators	15	8	12	0
RQ3: behav. mediators	13	12	13	4
Total	343	321	27	4

Table 4: ChatGPT errors in data extraction.

	Misidentified data (FN)	Missed data (FN)	Fabricated data (FP)
Bibliometric data	3 [S11,S17,S20]	0	0
General data (SMS context)	3 [S10,S21,S23]	0	0
RQ1	6 [S5,S11,S12,S16,S18,S22]	0	0
RQ2	3 [S3,S14,S16]	0	2 [S1,S21]
RQ3: data source	5 [S5,S7,S10,S15,S16]	0	1 [S19]
RQ3: psychol. mediators	1 [S10]	7 [S1(twice), S10(twice),S12,S19,S23]	10 [S3,S9,S11,S15,S16,S17,S18,S20,S21,S25]
RQ3: behav. mediators	0	1 [S8]	9 [S1,S2,S5,S9,S10,S18,S19,S21,S22]
Total	15	8	22

Table 5: ChatGPT accuracy.

Data	Incorrect ChatGPT extractions		Correct ChatGPT extractions		
	False Positive	False Negative	True Positive	True Negative	Accuracy
All data elements	22	23 (15 + 8)	321	4	87.83%
Generic data + RQ1–RQ3 data elements	22	26 (18 + 8)	102	4	68.83%
Only RQ1–RQ3 data elements	22	23 (15 + 8)	80	4	65.11%
Only bibliometric data elements	0	3 (3 + 0)	213	0	98.6%

extracted by humans. Similarly, Gartlehner et al. [5] assessed the use of an LLM (Claude-2) to extract 16 distinct data types, posing varying degrees of difficulty (160 data elements across ten studies). Across 160 data elements, Claude 2 demonstrated an overall accuracy of 96.3% and made six errors.

Polak et al. [15] proposed the ChatExtract method to fully and accurately automate data extraction with minimal initial effort and background. Through prompts, the method identifies sentences with data, extracts them, and assures the data's correctness through a series of follow-up questions. They found accuracy to be nearly 90.0% using GhatGPT-4.o.

Sun et al. [17] evaluated the performance of ChatPDF and Claude for use in automated data extraction. Their results highlight the potential of these LLM-based AI tools for automated data extraction. Alshami et al. [1] agree that although LLMs can help generate research questions and suggest boolean research terms, they are restricted to data extraction. In common, Sun et al. [17] and Alshami et al. [1] alert that while promising, the percentage of correct responses is still unsatisfactory. Therefore, improvements are needed to adopt them in research practice.

Similar to the previously mentioned studies, we also investigated using LLM to extract data. However, to our knowledge, this is the first study evaluating the accuracy of using an LLM to extract data in SMS in SE. Moreover, our study also includes the definition of prompts for ChatGPT. The results obtained from our study are consistent with the ones from studies performed in other areas (accuracy has varied between 90.0% [15] and 96.3% [5]). However, considering the particularities of the SE area, it is necessary to investigate further how LLMs perform in this context, how to build or improve solutions using LLMs, and how to associate them with SE activities. With this work, we take a step towards this direction.

5 FINAL REMARKS

This paper presented a proof-of-concept that investigated the accuracy of an LLM (GhatGPT-4.o) in extracting data in an SMS in SE. Different data items (bibliometrics, context, related to RQs) were

considered, representing different degrees of difficulty. GhatGPT demonstrated an accuracy of 87.83% and made 45 errors, with false positives (i.e., hallucination) being the most common error (22 occurrences). The preliminary findings demonstrate the promising potential of employing ChatGPT for semi-automated data extraction to evidence synthesis in SMS in SE.

While the findings may have future implications for the application of ChatGPT in supporting SMS, they are initial insights. In future work, we plan to investigate other LLM tools' accuracy in different SE contexts. We also intend to explore the use of ChatGPT in all secondary study types (SMS, SLR, rapid reviews) and their activities, including, for example, defining search strings, selecting studies, and writing final reports based on specific stakeholders (e.g., academia, practitioners). Finally, we hope this study also inspires other SE researchers to investigate ChatGPT's advantages and limitations before widespread adoption.

6 ONLINE RESOURCES

Supplementary materials are publicly available in <https://figshare.com/s/2e24b24ae3404a2e0815>.

ACKNOWLEDGMENTS

This study was financed by the 1) National Science Foundation (NSF) grants #2236198, #2247929, and #2303042; 2) Brazilian National Council for Scientific and Technological Development (CNPq) grant, processes 302339/2022 – 1, 314797/2023–8 and 443934/2023–1; 3) Samsung-UFAM Project for Education and Research (SUPER – Article 48 of Decree number 6.008/2006, SUFRAMA); 4) FAPEAM (through the POSGRAD 22–23 project), CAPES (Finance Code 001), and 5) Espírito Santo Research and Innovation Support Foundation (FAPES) - processes 2023–5L1FC, 2021–GL60J, 2022–NGKM5, and T.O. 1022/2022.

REFERENCES

- [1] 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.

- [2] A. Angheliescu, 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.
- [3] D. S. Cruzes and T. Dybá. 2010. Synthesizing evidence in software engineering research. In *ACM-IEEE Symposium on Empirical Software Engineering and Measurement (ESEM'10)*. ACM, Bolzano-Bozen, Italy, 1–10.
- [4] K. R. Felizardo and J. C. Carver. 2020. *Automating Systematic Literature Review*. Springer International Publishing, New York, US, Chapter 11, 327–355.
- [5] G. Gartlehner, L. Kahwati, R. Hilscher, I. Thomas, S. Kugley, K. Crotty, M. Viswanathan, B. Nussbaumer-Streit, G. Booth, N. Erskine, A. Konet, and R. Chew. 2024. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods* 1, March 2024 (2024), 1–14. <https://doi.org/10.1002/jrsm.1710>
- [6] 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.
- [7] M. Ibanez, A. Di-Serio, and D. Delgado-Kloos. 2014. Gamification for engaging computer science students in learning activities: A case study. *IEEE Transactions on learning technologies* 7, 3 (2014), 291–301.
- [8] 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.
- [9] B.A. Kitchenham, D. Budgen, and P. Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC, USA.
- [10] T. Li, I. J. Saldanha, J. Jap, B. T. Smith, J. Canner, S. M. Hutfless, V. Branch, S. Carini, W. Chan, B. de Bruijn, B. C. Wallace, S. A. Walsh, E. J. Whamond, M. H. Murad, I. Sim, J. A. Berlin, J. Lau, K. Dickersin, and C. H. Schmid. 2019. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. *Journal of Clinical Epidemiology* 115 (2019), 77–89. <https://doi.org/10.1016/j.jclinepi.2019.07.005>
- [11] 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.
- [12] 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.
- [13] D. W. Newton, J.A. LePine, J.K. Kim, N. Wellman, and J.T. Bush. 2020. Taking engagement to task: The nature and functioning of task engagement across transitions. *Journal of Applied Psychology* 105, 1 (2020), 1–18. <https://doi.org/10.1037/apl0000428>.
- [14] 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.
- [15] M.P. Polak and D. Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* 15, 1 (2024), 1569. <https://doi.org/10.1038/s41467-024-45914-8>
- [16] 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.
- [17] Z. Sun, R. Zhang, S.A. Doi, L. Furuya-Kanamori, T. Yu, L. Lin, and C. Xu. 2024. How good are large language models for automated data extraction from randomized trials? <https://www.medrxiv.org/content/10.1101/2024.02.20.24303083v1>
- [18] E. Syriani, I. David, and G. Kumar. 2024. Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages* 80 (2024), 101287. <https://doi.org/10.1016/j.cola.2024.101287>
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.