Data extraction for systematic mapping study using a large language model – a proof-of-concept study in software engineering

Katia Romero Felizardo
K. R. Felizardo*
katiascannavino@utfpr.edu.br
Universidade Tecnológica Federal do Paraná – UTFPR
Cornélio Procópio, Paraná, Brazil

Igor Steinmacher
School of Informatics, Computing, and Cyber Systems at Northern Arizona University
Flagstaff, US
Igor.Steinmacher@nau.edu

Márcia Sampaio Lima
Universidade do Estado do Amazonas – UEA
Amazonas, Brazil
msllima@uea.edu.br

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

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

Monalessa Perini Barcellos
Universidade Federal do Espírito Santo – UFES
Espírito Santo, Brazil
monalessa@inf.ufes.br

ABSTRACT

Context: In Software Engineering (SE), systematic mapping study (SMS) is one of the methods adopted for evidence-based decision-making, selecting and synthesizing relevant literature on a specific research topic. Tool support is beneficial due to the time-intensive nature of the SMS process and its activities. **Gap:** Large Language Models (LLMs) such as ChatGPT–4.o can potentially accelerate repetitive activities, such as data extraction in the SMS process. Therefore, having a tool to assist this activity could save time and effort. This proof-of-concept study evaluates how ChatGPT–4.o can support SMS activities in SE, particularly data extraction. **Method:** We assessed the accuracy of utilizing ChatGPT–4.o for extracting data in one SMS, in contrast to the manual extraction. **Results:** The accuracy of ChatGPT–4.o was 87.83%. **Conclusions:** Our preliminary findings suggest that entirely replacing the human extraction process with ChatGPT–4.o is not recommended. However, employing ChatGPT for semi-automated data extraction for evidence syntheses in SMSs in SE is promising.

CCS CONCEPTS

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

KEYWORDS

ChatGPT, large language model, systematic mapping study, data extraction, software engineering

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

ACM Reference Format:
Katia Romero Felizardo, K. R. Felizardo, Igor Steinmacher, Márcia Sampaio Lima, Anderson Deizepe, Tayana Uchôa Conte, and Monalessa Perini Barcellos. 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

1 INTRODUCTION

Secondary studies, including systematic literature reviews (SLRs) and systematic mapping studies (SMSs), are crucial in advancing Software Engineering (SE) research and practice. However, the current approach to conducting them is laborious, so the resulting evidence synthesis may be up to date when published [8].

Formulating research questions, conducting comprehensive literature searches, critically selecting relevant studies, extracting data from included studies, and synthesizing/categorizing evidence are activities involved in secondary studies. Among these activities, extracting data from selected studies is one of the most crucial and also time-consuming and costly, since data are often manually extracted from the studies into standardized tables [3]. The large number of scientific literature further adds to these challenges [20]. 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.

Felizardo et al. [4] mention that data extraction is the activity with the least automation. In particular, fully automated data extraction is challenging due to the different ways SE researchers report results (e.g., tables or graphs), restricted full-text access, or because of the lack of information provided by the authors. Another challenge is obtaining high-quality, accurate extracted data.

In most cases, data are manually extracted 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.

Data extraction bias can undermine the validity of evidence syntheses 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].

Large language models (LLM) have emerged as a potential solution, with recent studies suggesting their capability to enhance the efficiency of secondary studies with superior performance when compared to older artificial intelligence (AI) methods [1, 2, 6, 12, 16, 18, 19, 21]. Given this context, our proof-of-concept paper aims to **assess the accuracy of ChatGPT–4.o in extracting pre-specified data elements from PDF versions of full-text primary studies compared to data extraction by humans.**

What is already known?

- Data extraction is a crucial procedure for evidence categorization in SMS, but it is labor-intensive and error-prone [4, 8, 10, 20].
- There has been a growing effort to automate data extraction based on natural language processing, language models, and recently, large language models (LLMs) [1, 2, 6, 12, 16, 18, 19, 21].

What is new?

- Our study is the first to demonstrate and compare the performance of the ChatGPT versus the manual extraction of data elements for SMS in SE. We demonstrate high accuracy for simple data element extraction, which degraded with more complex data types.
- ChatGPT exhibited an overall 87.83% accuracy in extracting data from the 25 primary studies selected in the considered SMS. The accuracy in extracting bibliometric data was 99.7%, and for data elements related to the RQs (more subjective data) was 65.11% (when compared to human-data extraction).
- Our study demonstrates the promising potential of employing ChatGPT–4.o to support SMS data extraction in SE.

What is the potential impact for SE researchers?

- ChatGPT is a promising tool to enhance accuracy in data extraction for SMS. However, future research needs to understand its capabilities and limitations in SE.
-

2 METHODOLOGY

Our study investigates the preliminary accuracy of ChatGPT data extraction capabilities for data elements commonly used for SMS evidence categorization. Therefore, in this section, we detail the method followed to assess the accuracy of ChatGPT to help with this activity.

2.1 Replicated SMS

We started our research by selecting a secondary study for replication. We chose a convenient data source [14] for two reasons: 1) the original data extraction was available, and 2) the original SMS was published in a reputational journal. The data for this SMS was extracted by the SMS authors and verified by our team researcher.

The replicated SMS 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 [14]. The data extraction was performed using a data extraction form to extract relevant data to answer the research questions (RQ) consistently. One SMS author initially extracted the data, and the other authors participated in the discussion meetings to solve doubt and double-check data, as suggested in [9]. The three RQs were:

- **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?
- **RQ3 –** How are the results evaluated using games and customized/personalized gamified systems?

2.2 Prompting ChatGPT

As illustrated in Table 1, for each data element, we created a prompt to extract data from the 25 included studies.

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. Because the API does not have the functionality to receive PDF files via upload, we initially extracted the text content of each study in Python. 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 extraction item (e.g., title, author, year, among others). As a result, the consumption of time and financial resources becomes significant, and 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 extraction item was previously assigned a key, where all values were obtained in a single iteration per analyzed study, reducing execution time in 97% and costs in 95.7%.

2.3 Accuracy assesssment

In this study, accuracy is the proportion of correctly extracted data items. It is calculated as (TP + TN)/(TP + FP + TN + FN), where true negative (TN) is the number of data items the ChatGPT correctly identified as unavailable in the full-text article; true positive (TP) is the number of data items correctly extracted; false negative (FN) is the number of data items missed or incorrectly extracted; and false positive (FP) is the number of hallucinated data, i.e., data items ChatGPT fabricated in cases where no data were available in the full-text article. One researcher classified the data types, and a second reviewed their correctness.

3 RESULTS

Our results are presented in Tables 2, 3, 4, and 5.

Table 1: Prompts used to automatically data extraction.

Research Question		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.
General data	Context of application	State the context of the application.
RQ1	What is the user classification strategy adopted?	State the name of the strategy(ies) used to classify users according to their preferences.
RQ2	What instrument(s) was (were) used to identify types of users?	State the instrument name used to identify user types (profiles).
RQ3	How are the results evaluated using games and customized/personalized gamified systems?	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.

Table 2 shows data manually extracted by the SMS authors and those extracted with the support of ChatGPT to answer the RQs.

Table 3 details the hits during data extraction, i.e., the number of data items correctly extracted (TP) and those correctly identified as unabailable (TN).

Table 4 specifics the errors during data extraction, i.e., the number of data items missed or incorrectly extracted and the number of data (FN) items fabricated by ChatGPT (FP).

Table 5 offers ChatGPT's accuracy in extracting data for SMS.

3.1 Discussions

The accuracy for the data elements extracted by the ChatGPT–4.o versus manual extraction is featured in Table 5. The accuracy of ChatGPT was 87.83%, versus manual extraction. In instances where data were available, ChatGPT successfully extracted the pertinent information of 321 cases (true positives) (see Table 3). However, ChatGPT fabricated 22 instances of data (FP) (see Table 4).

The accuracy was higher for simple extractions (e.g., bibliometric data, title, year – 99.7%) than those related to the RQs (65.11%) (see Table 5. Using ChatGPT to support data extraction is promising when compared to human performance. 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].

As outlined in Table 4, we identified 45 errors made by ChatGPT during the data extraction process. Data was incorrectly extracted in 15 instances, and ChatGPT missed the available data in eight (8) situations. Additionally, it seemed to have generated extra information in 22 cases. We could not locate the evidence in the study pdf, i.e., ChatGPT hallucinated about these data elements.

RQ1 – *What are user classification strategies commonly adopted in studies on the customization/personalization of games or gamified systems?* As summarized in Table 2—column 2 (RQ1), lines 3–6, considering human extraction, among the 25 included studies in the replicated SMS, 14 used strategies based on user interaction, as follows: 1) User types hexad framework: seven studies [S2, S3, S4, S5, S6, S7, S8]; 2) BrainHex gamer typology: five studies [S9, S10, S11, S12, S13]; 3) Player type: one study [S14]; and 4) Multidimensional approach: one study [S1]. Six studies used personality-based strategies, and the typologies cited in the articles were: 1) MBTI: three studies [S15, S16, S17], and 2) Five-factor model of personality: three studies [S18, S19, S20]. Five studies [S12, S21, S22, S23, S24] used taxonomy based on learning style, and the only learning style-based strategy found was the FSLSM. One study [S25] used the achievement goal questionnaire-revised (AGQ–R) to assess motivation for learning.

ChatGPT incorrectly extracted three (3) interaction-based strategies (user types hexad framework [S5], brainHex gamer typology [S11, S12], two (2) personality-based strategies (MBTI [S16] and five-factor model of personality [S18]), and one (1) learning style-based strategy (taxonomy [S22]) (see Table 2—column 3, RQ1).

RQ2 – *What and how are the instruments used to identify types of users?* Regarding human extraction (see Table 2—column 2, RQ2), four studies [S2, S6, S7, S8] adopted Hexad's questionnaire without any adaptation, and three studies [S3, S4, S5] did adopt it. Five studies used the BrainHex gamer questionnaire [S9, S10, S11, S12, S13] to identify users' profiles. One study [S17] used the MBTI instrument in its completeness, and two [S15, S16] used an adapted version. S19 used the Big Five Inventory (BFI), and the BFI–10, a simplified version of the Big Five, is being used by S18. S20 used the iGFP–5 questionnaire based on the Big Five Personality Factors model. Four studies [S14, S22, S23, S24] used the complete Index of Learning Styles Questionnaire to identify students learning styles. In [S25], the AGQ–R was used to classify each user according to their motivation to perform the activities.

Five was the number of errors made by ChatGPT concerning the data for RQ2 [S1, S3, S14, S16, S21] (see Table 2—column 3, RQ2). ChatGPT correctly extracted the Hexad framework as an

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

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

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 existing data; however, it does not affect the SMS's conclusions. Likewise, without details about adaptations, ChatGPT got 2 (two) MBTI extractions correct [S15 and S17] but unsuccessful in one of the instances [S16]. Relative to the FSLS model, one case was lost [S14], and one false data was created [S21]. Again, ChatGPT fabricated data that was unavailable in the S1 PDF (situational motivation scale – SIMS). These errors could lead to erroneous SMS conclusions.

RQ3 – *How are the results evaluated using games and customized/personalized gamified systems?* Concerning human extraction data (see Table 2—column 2, RQ3), most studies [S1, S6, S9, S21] used questionnaires as data sources to measure user motivation. Many studies [S2, S8, S13, S18] proposed questionnaires to analyze particular aspects of their research, such as the users' opinions regarding certain game elements and their perception of the application's effect. Another approach was to evaluate the effects of gameful applications

Table 3: ChatGPT hits during extraction per data element type.

	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	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 during extraction per data element type.

	Misidentified data (FN)	Missed data (FN)	Fabricated data (FP)
Bibliometric data	3 [S11,S17,S20]	0	0
General data	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 in supporting data extraction for SMS.

	Incorrect ChatGPT extractions		Correct ChatGPT extractions		
Data	**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%

Accuracy* = (TP + TN)/(TP + FP + TN + FN), where: TN – data correctly identified as unavailable in the full-text article; TP – data correctly extracted; FN – data missed + incorrectly extracted; and FP – fabricated data.

based on user behavior in these systems, usually according to log records [S3, S4, S5, S12, S14, S15, S16, S17, S20, S25]. In addition, questionnaires and log records were combined in eight studies [S6, S7, S10, S11, S14, S22, S23, S24].

Concerning psychological mediators, in five studies [S4, S5, S10, S12, S19], the effects were measured through user preferences; three studies presented motivation-based gamification results [S1, S6, S14], and seven [S1, S2, S10, S13, S22, S23, S24] deal with questionnaires on perceptions of enjoyment, usefulness, flow experience, and behavior intention. Behavioral mediators include studies that compute the number of times the user has used the application. Usually, this data comes from the tool's usage log or similar measures. Three studies measured performance to evaluate the application's effect on users [S6, S20, S24]. Ten studies [S3, S4, S8, S11, S14, S15, S16, S17, S20, S25] measure student performance, such as grades and concepts attributed to learning activities in educational applications.

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. The opinion of humans was that the data source in these studies was 'log records.' For S19, ChatGPT also stated the adoption of a 'questionnaire'; however, no data was manually extracted about this. Therefore, ChatGPT faked this data. Surprisingly, for situations in which two sources were combined [S6, S7, S10, S11, S22, S23, and S4], ChatGPT was unsuccessful for only two instances [S7, S10]. However, for both S7 and S10, one of the data sources was extracted.

Considering psychological mediators positively, in four (4) circumstances [S7, S8, S12, S23], like humans, ChatGPT did not extract data. Negativelly, 10 (ten) [S3, S9, S11, S15, S16, S17, S18, S20, S21, S25] data were fabricated by ChatGPT. Seven (7) data were available in the PDFs but were missed by the ChatGPT ('preferences' [S10, S12, S19]; 'motivation' [S1]; 'enjoyment and usefulness' [S1, S10, S23]). SMS authors classified the data from S10 as 'preferences'; however, ChatGPT extracted one data element and categorized

it as 'motivation.' We observed a recurrence of 'engagement' categorization by ChatGPT (n = 15). 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] was missed related to the 'learning outcomes' category; however, nine (9) data were fabricated.

3.2 Threats to Validity

Some limitations of our study should be noted. We used a small sample size to analyze data extraction accuracy from only 25 primary studies included in one SMS article published in a specific journal. Therefore, the representativeness is limited, and exploring more articles with different data types is a consideration for future research.

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, in our replication, we set temperature and top_p parameters to zero (0) to control the "creativity" (randomness) of the data extracted by ChatGPT–4.o.

Currently, LLMs are not training models on domain-specific data. Future research could create datasets representing SE knowledge to investigate whether training on SE corpora increases the accuracy of ChatGPT in data extraction.

4 RELATED WORK

Replacing manual data extraction from scientific articles with automated data extraction based on LLMs is the focus of recent studies [1, 5, 11, 15, 17, 21].

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. [21] point out that LLMs can extract data similar to those extracted by humans. Similarly, Gartlehner et al. [5] assessed using 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 (6) 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—especially in the medical field—, we also investigated using LLM to extract data. However, our study investigated ChatGPT adoption, especially in the SE field for SMS. Comparable to our results, 87.83%, the accuracy of other studies published in the literature concerning applying LLMs to support extraction data has varied between 90.0% [15] and 96.3% [5].

5 FINAL REMARKS

This proof-of-concept paper assessed the accuracy of an LLM (GhatGPT–4.o) in extracting data from SMS, comparing it with human data extraction. We selected distinct types of data, posing varying degrees of difficulty. GhatGPT demonstrated an accuracy of 90.2% and made 45 errors, with false positive (i.e., allucination) being the most common error (n = 22). Our preliminary findings demonstrate the promising potential of employing ChatGPT for semi-automated data extraction for evidence syntheses in SMS in SE.

APPENDIX

6 ONLINE RESOURCES

Supplementary materials are available on https://figshare.com/s/2e24b24ae3404a2e081

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. 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.

[3] D. S. Cruzes and T. Dybå. 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. *Resesearch 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* 3 (2014), 291–301. Issue 7.

[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

Table 6: List of includes studies in the replicated SMS

ID	Title	Year
S1	Personalization improves gamification: evidence from a mixed-methods study	2021
S2	Game element preferences and engagement of different hexad player types in a gamified physics course	2020
S3	An investigation of gamification typologies for enhancing learner motivation	2014
S4	Investigating the impact of a meaningful gamification-based intervention on novice programmers' achievement	2018
S5	Towards a motivational design? Connecting gamification user types and online learning activities	2020
S6	To tailor or not to tailor gamification? An analysis of the impact of tailored game elements on learners' behaviours and motivation	2020
S7	Exploring personalization of g amification in an introductory programming course	2021
S8	Gamifying teacher students' learning platform: Information and communication technology in teacher education courses	2020
S9	Adaptive gamification for learning environments	2018
S10	Adaptation of gaming features for motivating learners	2017
S11	Adapting gamified learning systems using educational data mining techniques	2020
S12	Reinforcement learning for new adaptive gamified LMS	2019
S13	Does tailoring gamified educational systems matter? The impact on students' flow experience	2020
S14	An adaptive feedback system to improve student performance based on collaborative behavior	2019
S15	Analyzing the effect of game-elements in e-learning environments through MBTI-based personalization	2016
S16	Toward a personalized game-based learning environment using personality type indicators	2017
S17	The relationship between gender and game dynamics in e-learning environment: An empirical investigation	2018
S18	Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors	2018
S19	Educational gamification based on personality	2017
S20	The impact of gamification on students' learning, engagement and behavior based on their personality traits	2020
S21	Adaptive gamification in e-learning based on students' learning styles	2019
S22	A fundamental study for gamification design: Exploring learning tendencies' effects	2020
S23	An experimental study: Personalized gamifed learning based on learning style	2020
S24	The empirical investigation of the gamified learning theory	2020
S25	Personalization of gamification-elements in an e-learning environment based on learners' motivation	2016

(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* 1 (2020), 1–18. Issue 105. 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://doi.org/10.1101/2024.02.20.24303083

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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