

Original Paper

Critical Assessment of Large Language Models' (ChatGPT) Performance in Data Extraction for Systematic Reviews: Exploratory Study

Hesam Mahmoudi¹, PhD; Doris Chang¹, BSc; Hannah Lee¹, BSc; Navid Ghaffarzadegan², PhD; Mohammad S Jalali¹, PhD

¹MGH Institute for Technology Assessment, Harvard Medical School, Boston, MA, United States

²Industrial and System Engineering Department, Virginia Tech, Falls Church, VA, United States

Corresponding Author:

Mohammad S Jalali, PhD
MGH Institute for Technology Assessment, Harvard Medical School
125 Nashua St
Boston, MA 02114
United States
Phone: 1 6177243738
Email: msjalali@mgm.harvard.edu

Abstract

Background: Systematic literature reviews (SLRs) are foundational for synthesizing evidence across diverse fields and are especially important in guiding research and practice in health and biomedical sciences. However, they are labor intensive due to manual data extraction from multiple studies. As large language models (LLMs) gain attention for their potential to automate research tasks and extract basic information, understanding their ability to accurately extract explicit data from academic papers is critical for advancing SLRs.

Objective: Our study aimed to explore the capability of LLMs to extract both explicitly outlined study characteristics and deeper, more contextual information requiring nuanced evaluations, using ChatGPT (GPT-4).

Methods: We screened the full text of a sample of COVID-19 modeling studies and analyzed three basic measures of study settings (ie, analysis location, modeling approach, and analyzed interventions) and three complex measures of behavioral components in models (ie, mobility, risk perception, and compliance). To extract data on these measures, two researchers independently extracted 60 data elements using manual coding and compared them with the responses from ChatGPT to 420 queries spanning 7 iterations.

Results: ChatGPT's accuracy improved as prompts were refined, showing improvements of 33% and 23% between the initial and final iterations for extracting study settings and behavioral components, respectively. In the initial prompts, 26 (43.3%) of 60 ChatGPT responses were correct. However, in the final iteration, ChatGPT extracted 43 (71.7%) of the 60 data elements, showing better performance in extracting explicitly stated study settings (28/30, 93.3%) than in extracting subjective behavioral components (15/30, 50%). Nonetheless, the varying accuracy across measures highlighted its limitations.

Conclusions: Our findings underscore LLMs' utility in extracting basic as well as explicit data in SLRs by using effective prompts. However, the results reveal significant limitations in handling nuanced, subjective criteria, emphasizing the necessity for human oversight.

JMIR AI 2025;4:e68097; doi: [10.2196/68097](https://doi.org/10.2196/68097)

Keywords: large language models; generative artificial intelligence; systematic reviews; evidence synthesis; human-AI collaboration

Introduction

Systematic literature reviews (SLRs) are indispensable across various fields, synthesizing evidence to inform

decision-making in areas as diverse as public health, policy, and biomedical sciences, where rigor and comprehensiveness are paramount. With the rapid expansion of the literature, SLRs are more important than ever to help not only

synthesize evidence but also identify areas in which the literature is robust or deficient [1]. However, conducting SLRs is resource intensive, involving manual and careful screening of potentially relevant studies [2]. In particular, SLRs that assess and report on analytical methods and key findings require more domain-specific expertise and multiple researchers for coding, making them more challenging.

Given the recent rapid advancement of large language models (LLMs), researchers have proposed their potential utility in conducting SLRs [3,4]. Several reviews have found that artificial intelligence (AI)-enabled methods exhibit reasonable performance and improved efficiency in literature screening [5-9], an integral component of SLRs. In particular, the role of SLRs in health and biomedical sciences underscores the need for reliable, accurate data extraction tools that maintain the rigorous standards expected in these fields. However, studies that have tested the capabilities of AI for data extraction have identified challenges necessitating human intervention for completion [6,8].

To understand whether recently developed LLMs can overcome this barrier, many studies have evaluated the performance of various LLMs in automating SLR tasks [10-12]. Their findings reveal LLMs' potential in extracting data that are relatively easily retrievable (ie, study design, participant characteristics, and primary outcomes). However, these studies have not explored the ability of LLMs to extract more complex data (eg, methods used to obtain study outcomes), which may pose greater challenges as extracting such complex information is often subject to individual researchers' perspectives [13,14].

LLMs currently demonstrate advantages in assisting researchers, such as having rapid response times and providing high-level summaries of results [15]; however, a recent study revealed the shortcomings in ChatGPT's depth of knowledge and contextual understanding for conducting SLRs in comparison to researchers [16]. Yet, there is growing potential for ChatGPT's utility as an assistant in complex qualitative content analysis, as shown by a recent study that assessed its ability to categorize strategies and behaviors in forum posts about reducing sugar intake [17]. Thus, although LLMs show potential in data extraction tasks for SLRs, there is more to be explored on the current LLMs' capability to undertake a comprehensive approach that involves extracting not only basic study characteristics but also information critical for interpreting the results of studies within SLRs.

In this study, we aimed to evaluate an LLM's (ChatGPT) ability to extract more complex, nuanced data from scientific studies, representing a novel approach that goes beyond the simpler tasks of extracting descriptive information, such as study design or participant characteristics, which have been the primary focus of prior research. Our study applied a structured series of prompts and validations to systematically gauge where LLMs excel and where human oversight remains essential.

We used ChatGPT (GPT-4) for its accessibility and ease of use, recognizing that most individuals conducting SLRs may not be proficient with more customizable, technical platforms,

such as the GPT application programming interface (API). Sophisticated LLM frameworks (eg, retrieval-augmented generation) offer technical capabilities that may better address needs for SLRs, but these features generally require AI-specific expertise. Consequently, our study prioritized a more common usage scenario in which a widely accessible GPT model is used for data extraction tasks.

Methods

Overview

We focused on COVID-19 simulation modeling studies as a case study, leveraging a large collection of study reports that our team recently assembled for a SLR that aimed to assess the incorporation of human behavior dynamics in COVID-19 simulation models [18]. For this study, the data elements selected for extraction were aligned with the specific objectives of our review: to evaluate ChatGPT's capacity to manage both explicit and nuanced data in SLRs. To determine whether LLMs can effectively screen papers and extract information, we randomly chose 10 of the papers and extracted data both manually and using ChatGPT. We selected a sample of 10 studies [19-28] as this number was manageable for an in-depth exploration of each paper, allowing us to conduct detailed comparisons of extraction accuracy and making it feasible to perform multiple iterations of prompt engineering and assessment throughout the process.

Our examination of COVID-19 modeling studies leveraged our team's expertise and enabled us to confidently determine correct answers for meaningful comparisons with ChatGPT's outputs. Although COVID-19 may no longer be at the forefront of global attention, these studies remain a *methodologically rich test case* rather than a focus of current clinical relevance. They combine straightforward features (eg, study design, interventions) with more abstract elements of human behavior (eg, compliance with public health measures, mobility changes), allowing us to systematically assess ChatGPT's performance across varying levels of complexity. Furthermore, the pandemic's urgency previously highlighted the value of accurate, rapid synthesis of research findings during health emergencies, making it an ideal context for evaluating the practical utility of LLMs in accelerating SLRs.

Many researchers who conduct SLRs may not necessarily have the expertise required to implement LLMs through advanced tools. Thus, we conducted our analysis using ChatGPT, given that it is one of the most widely adopted models, attracting 393 million users each month, as shown by October 2024 data [29]. We used the web browser interface for its user-friendly design, as opposed to an API. Although using the API might have offered more controllable responses, our study prioritized accessibility and usability, reflecting the real-world context in which many researchers engage with ChatGPT. We selected the GPT-4 model specifically, as at the time of our analysis, it was one of the few that directly analyzed full text as a PDF file. Although our approach could also be applied to studies available in other formats, our focus

on PDF files reflects their frequent use in academic publishing and SLRs.

Ethical Considerations

This study did not involve human participants, identifiable human data, or interaction with individuals. As such, it did not fall under the scope of research requiring review by an institutional review board, and ethical approval was not required.

Data Elements

We defined the sets of measures to be extracted as (1) study settings (ie, analysis location, modeling approach, and analyzed interventions) and (2) behavioral components (ie, changes in travel and mobility, perception of risk and severity, and compliance and resistance to public health measures). Therefore, for each of the 10 studies [19–28], we extracted 6 distinct data elements, resulting in 60 data elements.

We distinguished these measures to reflect their nature: study settings are straightforward, whereas extracting information on behavioral components in COVID-19 models is influenced by researchers' perspectives [13]. We confined our study settings to information explicitly stated in the text. For behavioral components, we categorized them into no mention (A), mentioned but not modeled (B), modeled exogenously but not analyzed (C), modeled exogenously and analyzed (D), modeled endogenously but not analyzed (E), and modeled endogenously and analyzed (F). Endogenous modeling incorporated human behavior as an internal part of the model, where it both influenced and was influenced by the spread of COVID-19. Exogenous modeling indicated behavioral changes were external factors impacting the spread of COVID-19 without being influenced by it. This classification ranged from minimal (A) to comprehensive (F) incorporation into COVID-19 models.

Data Extraction Process and Comparison

Two trained researchers independently extracted data from the 10 studies [19–28] and then reconciled discrepancies in their findings—6 related to study settings and 11 to behavioral components—for convergence. The researchers discussed any unresolved discrepancies with a third senior researcher to reach a consensus. Subsequently, we initiated a dedicated session for every study and prompt, uploading the individual

files (in PDF) into the GPT-4 model using ChatGPT's user interface (accessed in January–April 2024) and documented the responses. To gain confidence in our manual screening, wherever ChatGPT, consistently through iterations of prompts, provided answers that disagreed with our manual coding, we reassessed our original codings. Following this reassessment and after making necessary adjustments, we finalized the manual screening results and considered them as the correct responses. We calculated individual researchers' average accuracy rates (the percentage of their correct responses before any consensus was reached), allowing us to directly compare ChatGPT's performance against the individual researchers' average accuracy for each measure throughout prompt iterations [8].

Prompt Engineering

We started by providing ChatGPT with a general prompt to extract each desired data element. Due to initially unsatisfactory results, we iteratively engineered prompts based on the initial responses and our manual coding. This process involved altering the wording of prompts for clarity and concisely adding specific descriptions of our objective, approach, and definitions of key terms [30]. Particularly, we often guided ChatGPT to base any interpretations strictly on what was explicitly stated in the text, as it often made incorrect inferences. Furthermore, we followed up with ChatGPT regarding its incorrect responses by inquiring about potential improvements to the prompts after clarifying the desired answers [30].

We continued to refine the prompts until we achieved complete alignment with our manual screening results, reached saturation in improvement, or explored viable avenues for prompt enhancement to the best of our capabilities. Given the exploratory nature of this study and the fact that LLMs are designed to interact with users in real time, prompt refinement without a formal training phase reflects a common use case for this technology. Recent tutorials and case studies have demonstrated ChatGPT's feasibility in domain-specific and rapid literature reviews, reinforcing its relevance as a practical, real-world tool that benefits from iterative, user-guided prompt refinement [31,32]. Table 1 illustrates an example of the iterative process used to prompt ChatGPT for extracting data elements in this study. We applied a similar method to extract additional data elements, with the processes reported in Multimedia Appendix 1.

Table 1. Example of the iterative process of prompt engineering to extract data elements.

Version	Description of prompt modification	Prompt
1	Initial prompt	What is the simulation modeling approach used in this paper?
2	Improved clarity	What model is used?
3	Focused output	Specify the overall type of model used in the study. If the modeling approach is unspecified, please state so.
4	Avoid overreporting	Specify the overall type (as opposed to the name) of the model used in the study. If the paper does not explicitly introduce the type of model used, state so by returning "unspecified."
5	Exclude inferred information	Specify the overall type (as opposed to the name) of the model used in the study. If the PDF file does not explicitly introduce the type of model used, state so by returning "unspecified," and do not infer the type of model.

Version	Description of prompt modification	Prompt
6	Emphasis on explicit information	Read this PDF file line by line. Only specify the overall general type, as opposed to the name, of the model used in the study. If the author(s) of this PDF file do not explicitly introduce the type of model used, state so by returning “unspecified,” and do not infer the type of model. Be sure to only specify the type of foundational analytical model rather than any supplementary methods.
7	Step-by-step instructions	<p>Read the provided PDF document line by line, focusing on identifying the general category or type of model mentioned in the study. Your task is to:</p> <ul style="list-style-type: none"> Identify the type of model: Look for any mention of the foundational analytical model used in the research. Specify only the general category or type (eg, regression model) rather than the specific name or variant. Explicit mention required: If the document does not explicitly mention the type of model used in the analysis, respond with “unspecified.” Avoid making inferences based on the context or the data presented. Focus on the foundational model: Concentrate on identifying the primary analytical model that the study is based on. Disregard any supplementary methods, tools, or analytical techniques that are mentioned unless they are integral to the foundational model itself.

Results

Table 2 summarizes ChatGPT’s responses across 6 measures for the last iteration of prompts. These measures were divided into study settings and behavioral components.

Through the course of iterations, we identified 4 instances where ChatGPT consistently disagreed with our manual coding but was determined to have provided the correct answer upon reassessment. These instances were among the prompts related exclusively to study settings: 3 pertained to answers generated for prompts about interventions analyzed, and 1 addressed the correct location of analysis.

As a result of iterative prompt engineering, the average accuracy of ChatGPT’s responses showed a marked

improvement of 33% and 23% between the initial and final iterations for extracting study settings and behavioral components, respectively ([Figure 1](#)). Specifically, in our initial prompts, 26 (43.3%) of 60 ChatGPT responses were correct. However, the latest prompt version yielded 43 (71.6%) correct answers. The iterative responses from ChatGPT and a comparison with our manual screening are detailed in [Multimedia Appendix 1](#). By the fourth iteration, ChatGPT outperformed the individual screeners’ average accuracy in identifying study settings ([Figure 1](#)). However, ChatGPT consistently could not achieve a level of precision comparable to that of manual screeners ([Figure 1](#)) when extracting behavioral components.

Table 2. ChatGPT’s responses in the final version of the prompts.^a

Study	Measure group 1: study settings			Measure group 2: behavioral components in COVID-19 models ^b		
	Location of analysis	Type of model, as presented by authors	Interventions analyzed	Changes in travel and mobility	Perception of risk and severity	Compliance and resistance to public health measures
Giordano et al [19]	Italy	Compartmental model known as the SIDARTHE model	<ul style="list-style-type: none"> Mass vaccination campaigns Nonpharmaceutical interventions Intermittent open-close strategies Different transmission rates due to new variants 	A [B] ^c	B	B [A]
Tuomisto et al [20]	Helsinki University Hospital region in Finland	Agent-based model	<ul style="list-style-type: none"> Physical isolation Testing and tracing Mobility restrictions Health care capacity enhancement (Import of infections)^c 	D	B [A]	C [A]
Ashcroft et al [21]	Does not focus on a specific geographic region	Mathematical model	<ul style="list-style-type: none"> Quarantine Test-and-release strategies 	A	A	C [D]

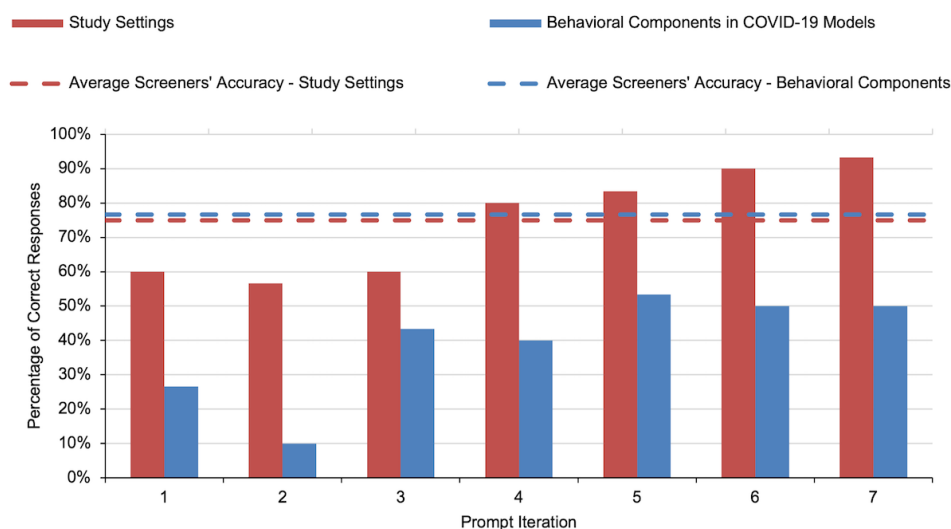
Study	Measure group 1: study settings			Measure group 2: behavioral components in COVID-19 models ^b		
	Location of analysis	Type of model, as presented by authors	Interventions analyzed	Changes in travel and mobility	Perception of risk and severity	Compliance and resistance to public health measures
Sneppen et al [22]	Sweden	Agent-based model	<ul style="list-style-type: none"> • Reinforced hygiene adherence • Limiting social contacts • Lockdown strategies 	A	A	B [A]
Wong et al [23]	Hong Kong	Unspecified	<ul style="list-style-type: none"> • Hygiene procedures • Aggressive escalation of border control • Implementing COVID-19 tests for overseas returners • Quarantine measures and social distancing 	A [B]	A	B [A]
Gostic et al [24]	Unknown	Mathematical model	<ul style="list-style-type: none"> • Active case finding • Symptom screening • Risk screening 	D [B]	A	A [D]
Kinoshita et al [25]	Unknown	Two-type branching process model	<ul style="list-style-type: none"> • Contact tracing • Case isolation 	A	B	C [A]
Paul et al [26]	Emphasis on South Asia, including India, Bangladesh, and Pakistan	SEIR epidemic model	<ul style="list-style-type: none"> • Lockdown • Social distancing • Individual-based precautionary measures 	C [D]	D [A]	B
Ebigbo et al [27]	Unknown	Model-based on theoretical assumptions	<ul style="list-style-type: none"> • Routine pre-endoscopy virus testing • High-risk personal protective equipment use • (Pre-endoscopy risk assessment questionnaire) 	A	A	B [A]
Kim and Paul [28]	Unknown	Unspecified	<ul style="list-style-type: none"> • Automated contact tracing • Use of personal protective equipment • Limited social distancing 	A	A	C [D]

^aAll responses are shortened for presentation.

^bResponse categories: A, no mention; B, mentioned but not modeled; C, modeled exogenously but not analyzed; D, modeled exogenously and analyzed; E, modeled endogenously but not analyzed; F, modeled endogenously and analyzed.

^cBrackets indicate manually screened responses, and parentheses flag ChatGPT's additional incorrect info; both are used only in cells with incorrect responses.

Figure 1. Average percentage of correct ChatGPT responses throughout iterations (bars) in comparison to the average accuracy of screeners before consensus (dashed lines) across study settings and behavioral components in COVID-19 models.

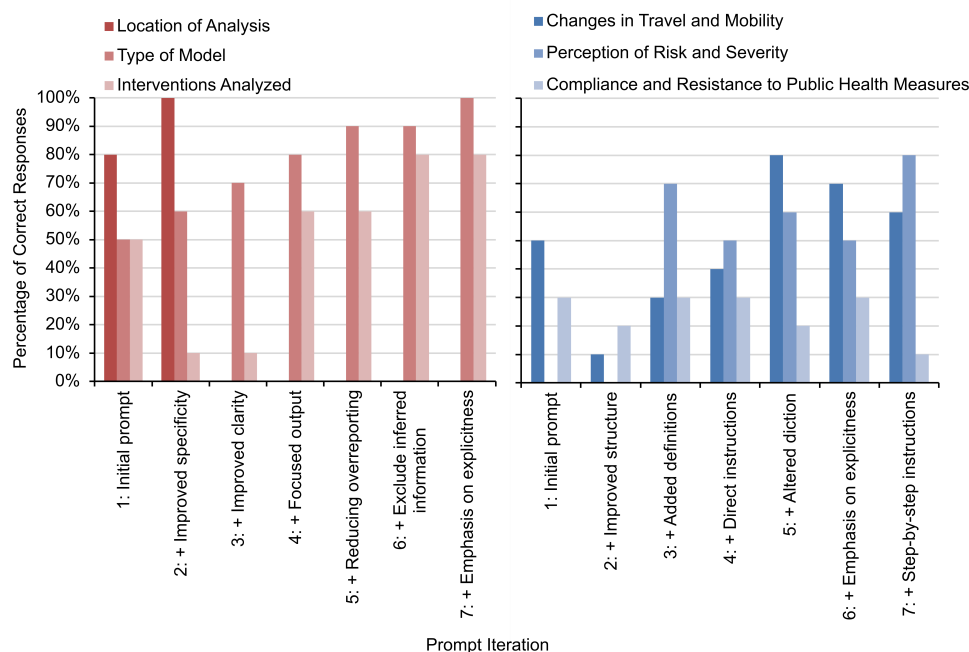


Specifically, ChatGPT provided correct responses for all 10 papers [19-28] when prompted to identify the analyzed location and model type used by the second and seventh iterations of prompts, respectively. At most, ChatGPT correctly identified 8 of 10 interventions analyzed within the studies, achieving a peak accuracy of 80% by the sixth iteration (Figure 2). Conversely, it took 6 and 7 iterations to reach a peak accuracy of 80% (ie, 8 of 10 correct classifications) for classifying how each study assessed changes in travel and mobility and the perception of risk and severity,

respectively. Our alignment with ChatGPT's responses for coding compliance and resistance to public health measures only achieved a maximum consistency rate of 3 (30%) correct answers across the 10 studies [19-28] (Figure 2).

Finally, the contrast between the distribution of the manual coding and ChatGPT is presented in Figure S1 in [Multimedia Appendix 1](#), and the distributions of ChatGPT's responses are presented in Figure S2 in [Multimedia Appendix 1](#).

Figure 2. Percentage of correct ChatGPT responses throughout iterations of each prompt. The red bars depict the progression of ChatGPT's accuracy for the three study setting elements, and the blue bars present the progression for the three behavioral component elements. Since the location of analysis achieved 100% by the second version, we concluded further iterations for this prompt.



Discussion

Principal Findings

Our analysis underscores the finding that ChatGPT's assistance in full-text screening of study reports is particularly useful when handling simple inquiries, specifically general study settings, for which details are typically explicit in the text. However, the task of assessing nuances that necessitates drawing inferences (eg, the integration of human behavior in COVID-19 models), where the study may or may not have an explicit discussion, presents a significant challenge for ChatGPT.

Rather than prolonging the iterative process to elicit correct responses, we adopted a strategy of structuring prompts based on ChatGPT's previous responses to ensure explicitness over the course of 7 iterations. Although it may have been possible to continue iterating to improve response accuracy, such a strategy is impractical in real-world scenarios. In addition, we acknowledge the potential risk of overfitting due to iterative prompt engineering on the same set of studies; however, our primary objective was to explore the utility of ChatGPT in real-world use cases where prompt refinement is common practice. This utility is especially relevant, given the broad reliance on SLRs to guide evidence-based practices across fields, with particular importance in health and biomedical sciences, where decisions impact health outcomes [31]. Hence, we limited this study to a manageable number of papers to better understand the limitations of prompt engineering itself. Importantly, the framework we used for evaluating ChatGPT (GPT-4) in COVID-19 modeling studies can be applied similarly to other SLRs regardless of the specific topic.

We highlighted that ChatGPT's performance is influenced by the explicitness of information within the text, not just by the clarity or objectivity of the prompts. This underscores a nuanced limitation: the technology's current dependency on explicit textual evidence for accurate data extraction. This limitation is notable, even with the use of straightforward prompts, underscoring a significant barrier in LLMs' application to literature analysis. For instance, despite clear prompts, ChatGPT often struggled to correctly identify the model type in studies—expected to be straightforward—unless explicitly mentioned. Even when instructed to label model types as “unspecified” in the absence of clear documentation, early iterations often resulted in incorrect answers rather than adherence to the “unspecified” directive. This illustrates that ChatGPT's accuracy is dependent not only on the prompt structure but also on the presence of explicitly detailed textual information. Hence, a central insight from this analysis is the significant obstacle that LLMs encounter when navigating ambiguity, further complicating the tasks of engineering effective prompts.

Furthermore, among ChatGPT's responses that did not correspond to our manual coding, we observed a tendency for ChatGPT to extrapolate beyond the presented data. Despite instructions to confine responses to the explicit content of each study, ChatGPT often listed additional interventions not

stated by the authors. This pattern of overreporting was also evident when categorizing studies based on the extent of their integration of behavioral components into the model. ChatGPT frequently assigned a higher integration level than that supported by the studies' text (Multimedia Appendix 1). These skewed errors align with ChatGPT's tendency to “hallucinate,” or provide confidently articulated yet factually unsupported responses [10].

Despite these challenges, LLM tools may be useful for SLRs. For 3 (30%) studies, ChatGPT correctly identified and analyzed interventions that we initially overlooked. For example, in Wong et al's study [23], ChatGPT identified COVID-19 testing as an intervention in 6 of 7 prompt iterations—a detail our original manual assessment missed. This led us to reevaluate the study, and upon confirming ChatGPT's accuracy, we modified our assessment accordingly. Conversely, we revisited our categorization of behavioral components in 6 incidences with which ChatGPT consistently disagreed, but we confirmed that our original manual coding was correct.

The main contribution of this study is to extend the understanding of current LLMs' capabilities and limitations in handling complex data, providing valuable insights for those who conduct SLRs and are exploring the use of LLM platforms. Our observation that ChatGPT outperformed the average accuracy of individual reviewers when identifying study settings underscores its utility as an assistant or a second reviewer in extracting basic measures for SLRs. These results support previous research, which indicates the potential of LLMs for handling basic data extraction tasks effectively [10,11] and their use as a collaborator [32] or a second rater [33] in SLRs. However, for ChatGPT to be effectively used in this role, there remains a need for researchers to provide clear and detailed prompts that provide the relevant context. This approach requires researchers to have a thorough understanding of the context relevant to their inquiries and have access to reliable, coded data elements to directly compare against ChatGPT's responses.

In terms of extracting complex components, ChatGPT failed to achieve comparable accuracy to that of individual screeners, highlighting the continued necessity of manual data extraction and additional research to overcome the limitations of this technology. Although further testing with a test sample may yield different results, the insufficient performance observed in our training data alone suggests that current LLMs remain unreliable for handling complex data extraction tasks. These findings align with other studies that discuss LLM performance, which similarly conclude that although automation is advancing, prevailing errors emphasize that structured oversight remains critical [15,34].

Limitations

This study is subject to several limitations. First, this study focused on one aspect of the SLR process, given that other steps (eg, writing Boolean query formulations [35] and screening titles and abstracts [10,36]) have already been examined in greater detail. Second, we evaluated ChatGPT with only one review topic, which limits the generalizability

of our results. Third, we often noticed inconsistencies in ChatGPT's responses for the same prompts, but we did not formally assess reproducibility. Fourth, although accurate data extraction highly depends on the LLMs' capabilities in accurately parsing PDF files [12], our paper does not quantify the impact of any errors in converting PDF files to text. Nonetheless, our results still highlight the stark differences in LLMs' capabilities in accurately extracting simple versus complex data elements. Fifth, we selected ChatGPT for this study due to its wide accessibility and usability. We recognize that our exclusive focus is a limitation, as other models may have provided different insights or comparative performance benchmarks. Future research should explore the capabilities of multiple models to provide a more comprehensive understanding of LLM performance in this context. For example, certain advanced approaches (eg, retrieval-augmented generation) exist to better handle evidence or references from external sources. Since our focus was on a common, practical scenario in which typical reviewers use a publicly accessible GPT model, these specialized methods fall outside

the scope of this study. Finally, due to the rapidly changing nature of LLMs, our findings may not hold over time.

Conclusion

Overall, LLMs, and ChatGPT in particular, show promising performance in assisting the extraction of explicitly stated information from the full text of study reports, particularly when limited scientific reasoning is required. However, ChatGPT currently exhibits limited potential for fully automating data extraction across more complex, subjective measures. Our findings emphasize the ongoing necessity of human oversight in handling complex, nuanced data extraction tasks, even as LLMs continue to improve. This position is consistent with broader calls in the literature to adopt a cautious, well-evaluated approach to integrating LLMs into evidence synthesis workflows [15]. We highlight an important contribution to human-AI collaboration research, demonstrating the need to integrate AI tools with human oversight in SLRs, particularly in areas where current models fall short.

Acknowledgments

The authors thank Ali Akhavan, Zeynep Hasgul, and Ning-Yuan Georgia Liu, who provided feedback on earlier versions of this paper, as well as the members of the US National Science Foundation (NSF) grant team, who shared feedback on the analysis design. This work was supported by the US NSF, Division of Mathematical Sciences and Division of Social and Economic Sciences (grant agreement 2229819).

Data Availability

This study did not involve analysis or acquisition of datasets outside what is reported; however, all data relevant to this paper and the process of research are shared in [Multimedia Appendix 1](#).

Authors' Contributions

HM and MSJ designed the experiment. HM, DC, and HL reviewed the papers to form the ground truth. HL, HM, and DC performed prompt engineering. HL and HM drafted the manuscript. HM and HL visualized results. MSJ and NG provided supervision and funding. All authors have reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The iterative process of prompt engineering and additional results.

[\[PDF File \(Adobe File\), 120 KB-Multimedia Appendix 1\]](#)

References

1. Owens JK. Systematic reviews: brief overview of methods, limitations, and resources. *Nurse Author Ed*. Dec 2021;31(3-4):69-72. [doi: [10.1111/nae2.28](#)]
2. Phillips V, Barker E. Systematic reviews: structure, form and content. *J Perioper Pract*. Sep 2021;31(9):349-353. [doi: [10.1177/1750458921994693](#)] [Medline: [34228554](#)]
3. Hossain MM. Using ChatGPT and other forms of generative AI in systematic reviews: challenges and opportunities. *J Med Imaging Radiat Sci*. Mar 2024;55(1):11-12. [doi: [10.1016/j.jmir.2023.11.005](#)] [Medline: [38040497](#)]
4. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. *Br Dent J*. Jul 2023;235(2):90-92. [doi: [10.1038/s41415-023-6132-y](#)] [Medline: [37500847](#)]
5. Feng Y, Liang S, Zhang Y, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. *J Am Med Inform Assoc*. Jul 12, 2022;29(8):1425-1432. [doi: [10.1093/jamia/ocac066](#)] [Medline: [35641139](#)]
6. Blaizot A, Veettil SK, Saidoun P, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Res Synth Methods*. May 2022;13(3):353-362. [doi: [10.1002/jrsm.1553](#)] [Medline: [35174972](#)]
7. de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. *Computing*. Oct 2023;105(10):2171-2194. [doi: [10.1007/s00607-023-01181-x](#)]

8. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev*. Jun 15, 2015;4(1):78. [doi: [10.1186/s13643-015-0066-7](https://doi.org/10.1186/s13643-015-0066-7)] [Medline: [26073888](#)]
9. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. *JCPP Adv*. Jun 2024;4(2):e12234. [doi: [10.1002/jcv2.12234](https://doi.org/10.1002/jcv2.12234)] [Medline: [38827982](#)]
10. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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. *Res Synth Methods*. Jul 2024;15(4):616-626. [doi: [10.1002/jrsm.1715](https://doi.org/10.1002/jrsm.1715)] [Medline: [38484744](#)]
11. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. Jul 2024;15(4):576-589. [doi: [10.1002/jrsm.1710](https://doi.org/10.1002/jrsm.1710)] [Medline: [38432227](#)]
12. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. *Res Synth Methods*. Sep 2024;15(5):818-824. [doi: [10.1002/jrsm.1732](https://doi.org/10.1002/jrsm.1732)] [Medline: [38895747](#)]
13. Achter S, Borit M, Cottineau C, Meyer M, Polhill JG, Radchuk V. How to conduct more systematic reviews of agent-based models and foster theory development - taking stock and looking ahead. *Environ Model Softw*. Feb 2024;173:105867. [doi: [10.1016/j.envsoft.2023.105867](https://doi.org/10.1016/j.envsoft.2023.105867)]
14. Li T, Higgins JP, Deeks JJ. Chapter 5: collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane; 2023. URL: <https://www.cochrane.org/authors/handbooks-and-manuals/handbook> [Accessed 2025-08-15]
15. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst Rev*. Apr 29, 2023;12(1):72. [doi: [10.1186/s13643-023-02243-z](https://doi.org/10.1186/s13643-023-02243-z)] [Medline: [37120563](#)]
16. Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G. Evaluating literature reviews conducted by humans versus ChatGPT: comparative study. *JMIR AI*. Aug 19, 2024;3(1):e56537. [doi: [10.2196/56537](https://doi.org/10.2196/56537)] [Medline: [39159446](#)]
17. Bijker R, Merkouris SS, Dowling NA, Rodda SN. ChatGPT for automated qualitative research: content analysis. *J Med Internet Res*. Jul 25, 2024;26(1):e59050. [doi: [10.2196/59050](https://doi.org/10.2196/59050)] [Medline: [39052327](#)]
18. Lee H, Mahmoudi H, Chang D, Jalali MS. Review of human behavior integration in COVID-19 modeling studies. *J Public Health (Oxf)*. Jul 12, 2025;fdaf082. [doi: [10.1093/pubmed/fdaf082](https://doi.org/10.1093/pubmed/fdaf082)] [Medline: [40650616](#)]
19. Giordano G, Colaneri M, Di Filippo A, et al. Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat Med*. Jun 2021;27(6):993-998. [doi: [10.1038/s41591-021-01334-5](https://doi.org/10.1038/s41591-021-01334-5)] [Medline: [33864052](#)]
20. Tuomisto JT, Yrjölä J, Kolehmainen M, Bonsdorff J, Pekkanen J, Tikkanen T. An agent-based epidemic model REINA for COVID-19 to identify destructive policies. *medRxiv*. Preprint posted online on Apr 17, 2020. [doi: [10.1101/2020.04.09.20047498](https://doi.org/10.1101/2020.04.09.20047498)]
21. Ashcroft P, Lehtinen S, Angst DC, Low N, Bonhoeffer S. Quantifying the impact of quarantine duration on COVID-19 transmission. *Elife*. Feb 5, 2021;10:1-33. [doi: [10.7554/eLife.63704](https://doi.org/10.7554/eLife.63704)] [Medline: [33543709](#)]
22. Sneppen K, Taylor RJ, Simonsen L. Impact of superspreaders on dissemination and mitigation of COVID-19. *medRxiv*. Preprint posted online on Jul 3, 2020. [doi: [10.1101/2020.05.17.20104745](https://doi.org/10.1101/2020.05.17.20104745)]
23. Wong MCS, Ng RWY, Chong KC, et al. Stringent containment measures without complete city lockdown to achieve low incidence and mortality across two waves of COVID-19 in Hong Kong. *BMJ Glob Health*. Oct 2020;5(10):e003573. [doi: [10.1136/bmjgh-2020-003573](https://doi.org/10.1136/bmjgh-2020-003573)] [Medline: [33028700](#)]
24. Gostic K, Gomez AC, Mummah RO, Kucharski AJ, Lloyd-Smith JO. Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *Elife*. Feb 24, 2020;9:e55570. [doi: [10.7554/eLife.55570](https://doi.org/10.7554/eLife.55570)] [Medline: [32091395](#)]
25. Kinoshita R, Anzai A, Jung SM, et al. Containment, contact tracing and asymptomatic transmission of novel coronavirus disease (COVID-19): a modelling study. *J Clin Med*. Sep 27, 2020;9(10):3125. [doi: [10.3390/jcm9103125](https://doi.org/10.3390/jcm9103125)] [Medline: [32992614](#)]
26. Paul A, Chatterjee S, Bairagi N. Prediction on COVID-19 epidemic for different countries: focusing on South Asia under various precautionary measures. *medRxiv*. Preprint posted online on Apr 11, 2020. [doi: [10.1101/2020.04.08.20055095](https://doi.org/10.1101/2020.04.08.20055095)]
27. Ebigbo A, Römmele C, Bartenschlager C, et al. Cost-effectiveness analysis of SARS-CoV-2 infection prevention strategies including pre-endoscopic virus testing and use of high risk personal protective equipment. *Endoscopy*. Feb 2021;53(2):156-161. [doi: [10.1055/a-1294-0427](https://doi.org/10.1055/a-1294-0427)] [Medline: [33080647](#)]
28. Kim H, Paul A. Automated contact tracing: a game of big numbers in the time of COVID-19. *J R Soc Interface*. Feb 2021;18(175):20200954. [doi: [10.1098/rsif.2020.0954](https://doi.org/10.1098/rsif.2020.0954)] [Medline: [33622147](#)]
29. Duarte F. Number of ChatGPT users. *Exploding Topics*. Dec 2024. URL: <https://explodingtopics.com/blog/chatgpt-users> [Accessed 2025-08-29]

30. Heston TF, Khun C. Prompt engineering in medical education. *International Medical Education*. Aug 2023;2(3):198-205. [doi: [10.3390/ime2030019](https://doi.org/10.3390/ime2030019)]
31. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: update of a living systematic review. *F1000Res*. Oct 2023;10:401. [doi: [10.12688/f1000research.51117.2](https://doi.org/10.12688/f1000research.51117.2)]
32. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models (LLM) as a tool in literature reviews: an LLM automated systematic review. *arXiv*. Preprint posted online on Sep 6, 2024. [doi: [10.48550/arXiv.2409.04600](https://doi.org/10.48550/arXiv.2409.04600)]
33. Motzfeldt Jensen M, Brix Danielsen M, Riis J, et al. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. *PLoS One*. Jan 7, 2025;20(1):e0313401. [doi: [10.1371/journal.pone.0313401](https://doi.org/10.1371/journal.pone.0313401)] [Medline: [39774443](https://pubmed.ncbi.nlm.nih.gov/39774443/)]
34. Chen X, Zhang X. Large language models streamline automated systematic review: a preliminary study. *arXiv*. Preprint posted online on Jan 9, 2025. [doi: [10.48550/arXiv.2502.15702](https://doi.org/10.48550/arXiv.2502.15702)]
35. Wang S, Scells H, Koopman B, Zucco G. Can ChatGPT write a good Boolean query for systematic review literature search? Presented at: SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 23-27, 2023:1426-1436; Taipei, Taiwan. [doi: [10.1145/3539618.3591703](https://doi.org/10.1145/3539618.3591703)]
36. Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. *J Comput Lang*. Aug 2024;80:101287. [doi: [10.1016/j.cola.2024.101287](https://doi.org/10.1016/j.cola.2024.101287)]

Abbreviations

AI: artificial intelligence

API: application programming interface

LLM: large language model

SLR: systematic literature review

Edited by Fida Dankar; peer-reviewed by Bill Thies, Ching Nam Hang; submitted 28.10.2024; final revised version received 27.03.2025; accepted 14.04.2025; published 11.09.2025

Please cite as:

Mahmoudi H, Chang D, Lee H, Ghaffarzadegan N, Jalali MS

Critical Assessment of Large Language Models' (ChatGPT) Performance in Data Extraction for Systematic Reviews: Exploratory Study

JMIR AI 2025;4:e68097

URL: <https://ai.jmir.org/2025/1/e68097>

doi: [10.2196/68097](https://doi.org/10.2196/68097)

© Hesam Mahmoudi, Doris Chang, Hannah Lee, Navid Ghaffarzadegan, Mohammad S Jalali. Originally published in JMIR AI (<https://ai.jmir.org>), 11.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.