future internet

MDPI

Article

A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis

Anjia Ye [1,*], Ananda Maiti [2,*], Matthew Schmidt [3] and Scott J. Pedersen [1]

[1] School of Education, University of Tasmania, Launceston, TAS 7248, Australia; scott.pedersen@utas.edu.au
[2] School of Information Technology, Deakin University, Geelong, VIC 3221, Australia
[3] School of Health Sciences, University of Tasmania, Sandy Bay, TAS 7005, Australia; matthew.schmidt@utas.edu.au
* Correspondence: anjia.ye@utas.edu.au (A.Y.); anandamaiti@ieee.org (A.M.)

Abstract: Systematic reviews (SRs) are a rigorous method for synthesizing empirical evidence to answer specific research questions. However, they are labor-intensive because of their collaborative nature, strict protocols, and typically large number of documents. Large language models (LLMs) and their applications such as gpt-4/ChatGPT have the potential to reduce the human workload of the SR process while maintaining accuracy. We propose a new hybrid methodology that combines the strengths of LLMs and humans using the ability of LLMs to summarize large bodies of text autonomously and extract key information. This is then used by a researcher to make inclusion/exclusion decisions quickly. This process replaces the typical manually performed title/abstract screening, full-text screening, and data extraction steps in an SR while keeping a human in the loop for quality control. We developed a semi-automated LLM-assisted (Gemini-Pro) workflow with a novel innovative prompt development strategy. This involves extracting three categories of information including *identifier*, *verifier*, and *data field* (IVD) from the formatted documents. We present a case study where our hybrid approach reduced errors compared with a human-only SR. The hybrid workflow improved the accuracy of the case study by identifying 6/390 (1.53%) articles that were misclassified by the human-only process. It also matched the human-only decisions completely regarding the rest of the 384 articles. Given the rapid advances in LLM technology, these results will undoubtedly improve over time.

Keywords: artificial intelligence; large language model; systematic review; methodology; workflow

check for updates

Citation: Ye, A.; Maiti, A.; Schmidt, M.; Pedersen, S.J. A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet* **2024**, *16*, 167. https://doi.org/10.3390/fi16050167

Academic Editor: Gianluigi Ferrari

Received: 20 March 2024
Revised: 3 May 2024
Accepted: 5 May 2024
Published: 12 May 2024

1. Introduction

Systematic reviews (SRs) provide a rigorous and comprehensive approach to synthesizing the evidence relevant to well-defined research questions. Such reviews facilitate the assessment of intervention efficacy, underscore extant research gaps, and underpin informed decision-making processes across many research areas [1]. Identifying research gaps is a principal utility of SRs. Through the systematic examination of the literature, SRs reveal domains where evidence is either insufficient or equivocal. Such insights are instrumental in shaping the target of subsequent research [2]. This methodology has contributed to ensuring that the most robust and reliable evidence is available to inform practice and policy decisions.

SRs aggregate evidence from disparate sources and present an unbiased and balanced synthesis of the available evidence. This is achieved through a methodical process comprising the identification, selection, and critical appraisal of relevant studies [3]. The resultant synthesis provides an evidence base upon which stakeholders can base their decisions.

The Preferred Reporting Items for SRs and Meta-Analyses (PRISMA) is a widely endorsed framework that delineates the essential elements and their reporting in SRs [1,4,5].

These elements include the development of the research question, the reporting of the search strategy, criteria for study selection, methods for data extraction, and the synthesis of results [6]. However, such rigorous procedures lead to a labor-intensive and time-consuming process that can generate a temporal lag and scope limitation [7,8].

The integration of artificial intelligence (AI) and large language models (LLMs) specifically presents a promising solution to these challenges by automating the study selection and data extraction tasks [9,10], as shown in Figure 1. However, research gaps in the application of LLMs within SRs need to be addressed to leverage this technology fully. These include concerns about the transparency of LLM processing and the reproducibility of the AI-assisted process.

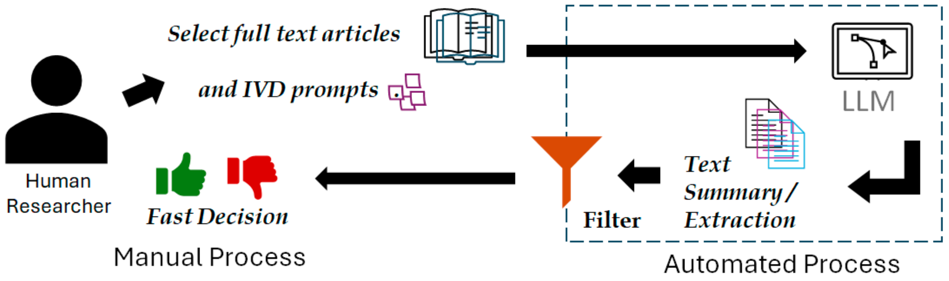


Figure 1. An overview of typical human–AI teaming for the proposed LLM-assisted applications.

In the proposed approach, the human-oriented tasks are simplified in terms of time and effort, although they may have to be performed iteratively sometimes. Based on this, we propose a hybrid human–LLM workflow to address these concerns. This approach uses an LLM to extract or summarize textual information into keywords, phrases, or succinct sentences, while human expertise is leveraged to make quick decisions based on the short extracted keywords and phrases. It allows for the initiation of clear and reproducible selection criteria and the rapid human validation of exclusion decisions. This can also reduce the cognitive burden and improve the accuracy and time efficiency of article screening and data extraction. We can use this workflow for other similar review methods as well, such as scoping reviews, which follow a systematic approach to map evidence on a topic and share the same processes for screening and extracting data in a wider scope [11].

This article is organized into the following sections: Section 1 states the research aims and questions. Section 2 discusses the SR process and LLM-related approaches. Section 3 follows with our proposed methodology and a case study of developing and testing an LLM application for the proposed methodology. Finally, Section 4 contains a brief report on our results, and Section 5 summarizes our conclusions.

1.1. Research Aims and Contributions in This Work

This article outlines and demonstrates a methodology for conducting an SR using LLM assistance. The primary research contributions of this work are as follows:

1. A human-oriented workflow that effectively blends with an existing LLM and reduces the human workload for performing SRs. The aim is to lessen the cognitive burden of a human-only approach to increase accuracy while maintaining the transparency and reproducibility of the review process.
2. A novel strategy of prompt engineering, categorizing extracted data/information into *identifiers*, *verifiers*, and data fields (IVD), which is particularly efficient for extracting information from a large volume of research articles for performing SRs. The IVD strategy may be applied in other applications involving fixed-format documents using LLMs.
3. A novel approach to study selection in SRs, leveraging the capabilities of LLMs for efficient data extraction. This strategy employs an LLM to extract short, meaningful information from full-text articles. Humans can make quick, manual decisions based on pre-determined exclusion criteria with these extractions. This process is not only

faster but may also reduce the human bias that has been linked to conventional SRs [12]. These techniques could set the foundation for future advancements in the field.

1.2. Focus and Limitations of This Work

In this work, we propose a workflow assuming a functional online LLM service is available. We do not propose any specific algorithm. LLMs are an evolving technology with no specific, efficient way to use them to obtain desired open-ended results. Also, this workflow's efficiency depends on the field of study, the specific LLM used, the research aims, and the expertise level of the human researchers. As such, the computational complexity of this method is out of the scope of this paper. We have, however, developed new metrics, such as *completeness*, to measure the effectiveness of the input prompts.

Our goal is to create a workflow suitable for non-programmers. The strategies used in the workflow and prompt engineering are suitable for such a human with no technical background. It should also be noted that the proposed methods may be less helpful for humans with better cognitive capabilities. Such people may already be able to read many articles and process them manually with relative ease. However, the impacts of such human factors are out of the scope of this work. We aim to show that, as a proof of concept, the proposed hybrid workflow will complete the following:

- Reduce the workload of a researcher;
- Give the correct classifications regarding inclusion and exclusion.

2. Related Works

2.1. The Systematic Review Process

SRs are essential for evidence-based research. They use strict design and protocols to ensure objectivity, reduce bias, and enable replication of valuable evidence synthesis in various domains [5,11,13,14]. Protocols like PRISMA provide frameworks to uphold reporting standards and quality control [1]. Within an SR process, several key steps facilitate objective knowledge synthesis. For instance, a well-defined search strategy employs reputable databases, precise Boolean operators, and meticulously chosen keywords [14]. In conjunction, well-articulated inclusion and exclusion criteria govern the selection of relevant studies [15]. Furthermore, study selection is typically conducted independently by at least two reviewers working in parallel, with a third reviewer resolving any conflicts. This collaborative process helps maintain transparency and objectivity throughout [5]. These carefully structured procedures underscore the fundamental commitment of SRs to rigorous methodology, ensuring the reliability of evidence for informed decision-making.

SR methodologies that encompass literature searching, screening, data extraction, and synthesis are notoriously time-consuming [7]. The breadth of a review topic or a vast literature base can further increase the time burden. Given the dynamic nature of research fields, the time taken to publish an SR may lead to the omission of new, relevant findings, leading to a temporal lag that potentially diminishes the utility of the review's recommendations [8]. The optimal frequency for updating SRs remains a contentious issue, with a balance needed between practicality and currency [4]. Further research is needed to determine optimal strategies for mitigating this temporal lag and scope limitation, exploring technological tools, collaborative models, and methodologies designed for rapid knowledge synthesis.

2.2. General AI Application

In response to these challenges, various AI techniques have been used to enhance the process's efficiency and accuracy. A support vector machine (SVM) is a supervised machine learning model used to select studies. For example, Bannach-Brown et al. [16] automatically identified relevant records with a high accuracy of 98.7% with a training set (n = 5749) consisting of manually screened records. However, a major practical limitation of SVM frameworks is the reliance on extensive training through manual title and abstract screening [17,18]. It remains unclear when it is “safe” for reviewers to stop this manual process [19].

2.2.1. LLM-Based SR Approach

Recent breakthroughs in LLMs offer enticing possibilities for tasks like question-answering, summarization, and data extraction [9,20–22]. Researchers have investigated LLMs’ potential to fully automate screening and decision-making within SRs with several pre-trained models. For example, Guo’s team found that an LLM (GPT-3.5-Turbo) could achieve an overall accuracy of 90% when used for automatic title/abstract screening [22]. However, when Syriani, David, and Kumar [10] extended this approach with a larger set of more recent studies, they found that the application’s (ChatGPT) performance only correctly classified articles for inclusion and exclusion above 70% and 60%, respectively. Others noted that this drop in accuracy was due to a high reliance on pre-existing data, which can become less accurate when dealing with new knowledge [9,20]. Despite these promising findings, certain limitations of LLMs still pose challenges for their widespread adoption in SR methodology. One significant concern that has been raised is the lack of transparency in the automatic study selection process [23]. Studies have highlighted that the opaque nature of fully automatic approaches makes it challenging to the transparency and validity of the results, limiting these methods in the development stage [19,24].

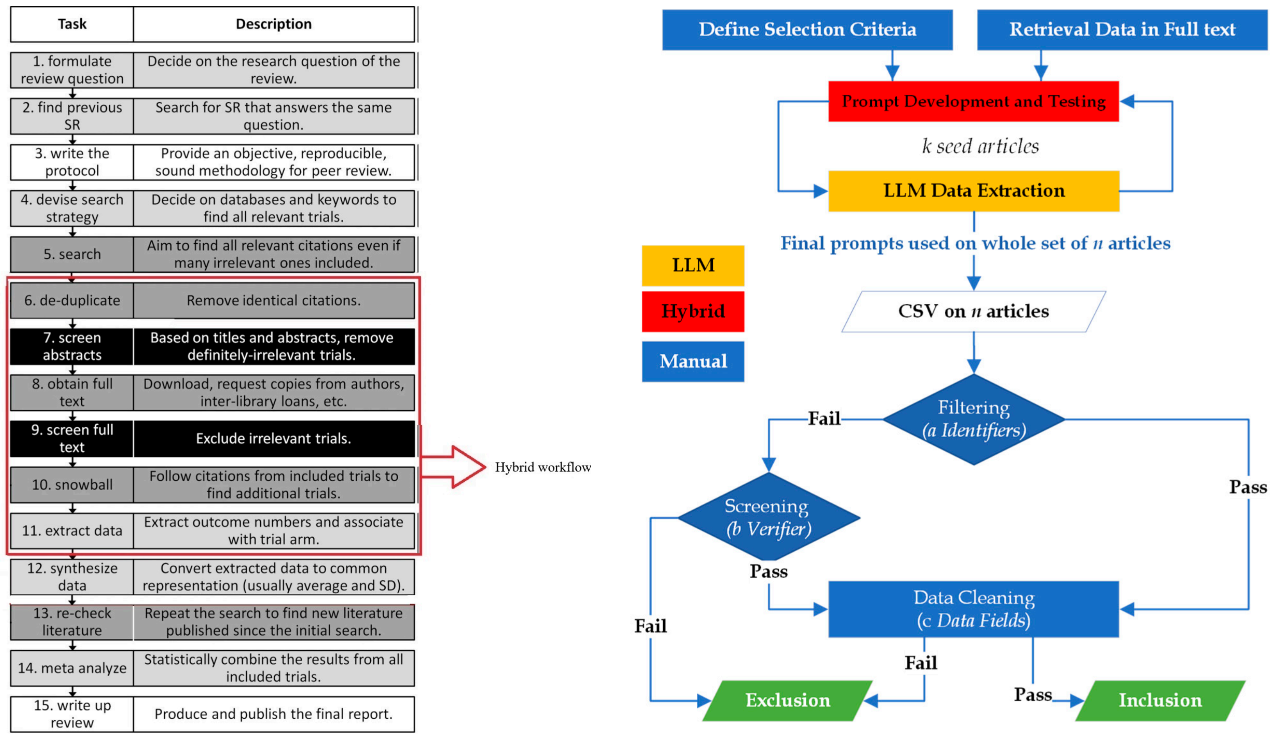
2.2.2. LLM-Assisted Human-in-the-Loop Approach

The black-box nature of automatic approaches can create a lack of academic acceptance. Thus, hybrid workflows are a potential method to ensure rigorous, trustworthy SR processes using AI assistance. Most practical tools used in SRs are designed as semi-automated processes. For example, Alshami, Elsayed, Ali, Eltoukhy, and Zayed [20] explored human–ChatGPT collaboration, with humans verifying ChatGPT’s title screening decisions (APA style citations) and selectively providing abstracts for additional context. However, context-length or token limitations hinder this method from being used for full article extraction. To address this, Khraisha, Put, Kappenberg, Warraitch, and Hadfield [9] tested the screening from full-text articles using GPT-4 and achieved an inclusion/exclusion decision accuracy of 88% with English peer-reviewed articles, while data extraction achieved an accuracy of 84%. This demonstrates the potential for LLMs in data extraction procedures, a task previously considered difficult [7,24]. As these LLM approaches continue to improve workflow transparency and performance accuracy, researchers may be provided with powerful tools to conduct more efficient and effective reviews.

Overall, developing novel LLM-assisted workflows that include rapid human verification and validation has the potential to cope with existing SR processes. Our proposed hybrid workflow is presented in the next section.

3. The Proposed Hybrid Workflow

We propose a new hybrid workflow that combines human reviewers and an LLM for screening and data extraction (Figure 2b) based on an existing SR workflow (Figure 2a) [13]. Our approach capitalizes on the LLM’s efficiency in extracting information from full-text articles, coupled with the human capability for reasoning and logical thinking when dealing with keywords and short phrases. To ensure accuracy and reliability, we integrate human decision-making at key stages of the workflow. Human researchers are particularly adept at scanning and filtering concise and meaningful keywords extracted from full-text articles, enabling them to make effective decisions based on the extracted information. By combining the strengths of both humans and LLMs, our hybrid workflow aims to improve the productivity and accuracy of the SR process.



Task	Description
1. formulate review question	Decide on the research question of the review.
2. find previous SR	Search for SR that answers the same question.
3. write the protocol	Provide an objective, reproducible, sound methodology for peer review.
4. devise search strategy	Decide on databases and keywords to find all relevant trials.
5. search	Aim to find all relevant citations even if many irrelevant ones included.
6. de-duplicate	Remove identical citations.
7. screen abstracts	Based on titles and abstracts, remove definitely-irrelevant trials.
8. obtain full text	Download, request copies from authors, inter-library loans, etc.
9. screen full text	Exclude irrelevant trials.
10. snowball	Follow citations from included trials to find additional trials.
11. extract data	Extract outcome numbers and associate with trial arm.
12. synthesize data	Convert extracted data to common representation (usually average and SD).
13. re-check literature	Repeat the search to find new literature published since the initial search.
14. meta analyze	Statistically combine the results from all included trials.
15. write up review	Produce and publish the final report.

(**a**) (**b**)

Figure 2. (**a**) Existing methods for SRs [13]. (**b**) New AI-assisted hybrid workflow stages.

Our modified workflow replaces the title and abstract screening, full-text screening, and data extraction steps of a typical SR, shown in Figure 2b. The workflow consists of using the full-text articles and the SR selection criteria to inform the development of prompts to extract the necessary IVD information. These outputs are defined below. Several iterations of IVD prompt engineering are performed to refine the prompts. The prompts are then used to enable the LLM to extract data from each study, which are converted and stored in a spreadsheet file. The filtering, screening, and data cleaning steps are then performed by a person using the spreadsheet file in CSV format to determine, through filtering, a study's inclusion or exclusion and to perform data extraction procedures. These steps are fully described below and include descriptive examples taken from our own SR case study.

3.1. Retrieve Data in Full Text

The LLM extracts the prompted information from full-text articles to enable humans to perform the initial inclusion/exclusion screening. This is different from the traditional SR methodology, where only the title and abstract are used in the initial screening and full-text articles are obtained later in the process [13]. To proceed, researchers must retrieve the full-text articles after identifying potential studies via the literature search strategy.

3.2. Define Selection Criteria

Prior to extracting data from full-text files, *inclusion* and *exclusion* criteria from the SR are collated by the research team. For example, in our case study below, we chose desk-based workers as our population of interest. At the end of this stage, we had a list of exclusion reasons, such as "*not desk-based workers*" to help eliminate unwanted articles.

3.3. IVD Prompt Development and Testing

As mentioned above, decision-making requires logical foundations and clear reasoning, areas where the inner workings of LLMs remain unclear. Two of the areas where LLMs excel are in *summarizing information* and *data extraction*. Data extraction accuracy depends on the engineering of appropriate *prompts*, which are inputs given to an LLM. For this application of LLM-supported SRs, we propose a set of prompts to guide LLMs in extracting data from full-text articles. These prompts were developed based on the SR selection criteria. This stage is an iterative phase and can be performed on an initial subset of all the articles chosen in the previous steps. It involves creating/updating prompts and testing them for refined outputs.

This initial set of articles can be termed *seed* articles. Every researcher would start their SR with at least a set of these initial articles. The human must first read them well enough to be confident that these *seeds* are relevant to the SR being performed and will be selected for inclusion. Once the prompts have been developed satisfactorily with the *k* seed articles, they can be used on the entire set of *n* articles.

3.3.1. Development of the Prompt Engineering Methodology

The goal of the IVD prompts is to accurately extract or summarize information from the full-text articles to allow humans to make inclusion/exclusion decisions. An example of a prompt is "*summarize the method in one or two sentences*". In the proposed hybrid workflow, the prompted outputs of the LLM are used by humans to make inclusion/exclusion decisions and to extract data from the studies. In our proposed methodology, the prompts are designed to output three types of information including *identifiers*, *verifiers*, and *data fields* (IVD). Identifiers and verifiers are used to make inclusion/exclusion decisions, while data fields are the extracted data from the studies. They are defined more specifically as follows:

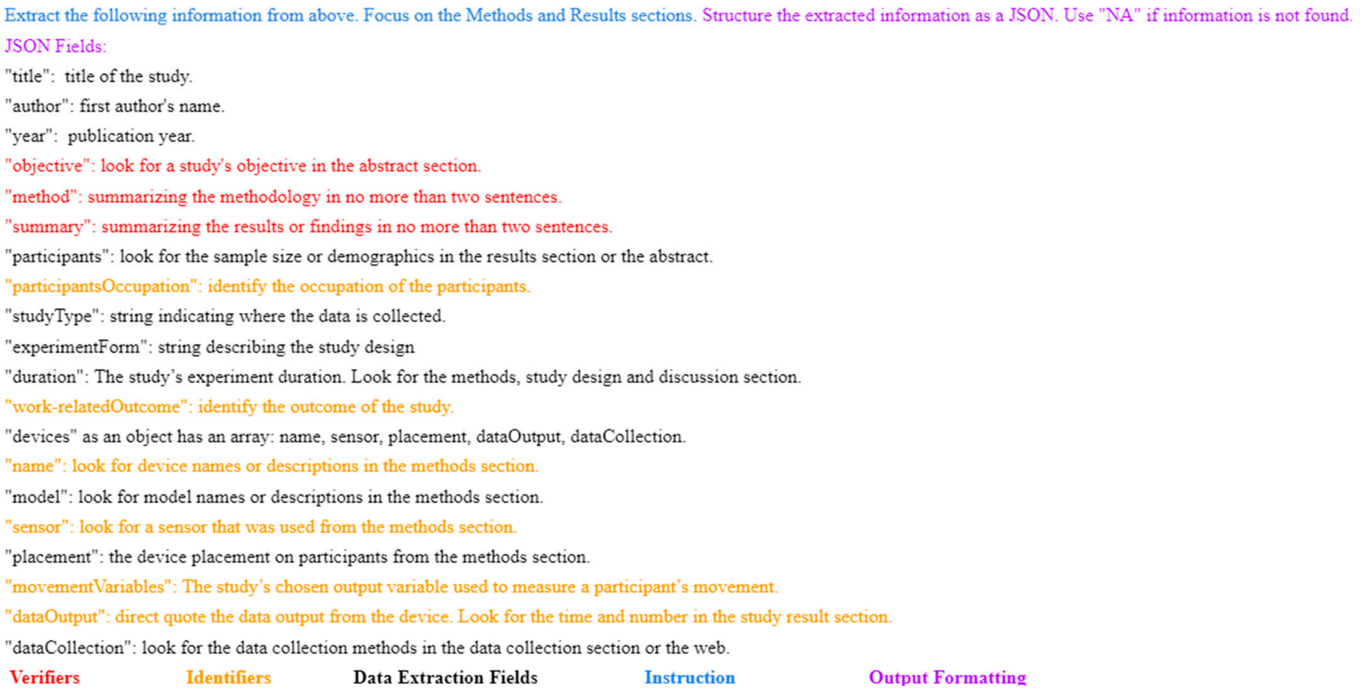
(a) **Identifiers** are the prompted outputs of the LLM, which are used by humans to make inclusion/exclusion decisions. The identifier prompts are engineered to reflect the *exclusion* criteria of a review and are typically one or a couple of words in length. Strategically worded prompts can extract identifying keywords or phrases in the output to aid human reviewers in making swift and accurate exclusion decisions. For example, in our case study, we focused on desk-based workers, so we developed a prompt to output an identifier that indicated the occupation of participants in the study to aid in decision-making about inclusion and exclusion.

(b) **Verifiers** are prompted outputs, which are concise phrases or sentences that summarize certain aspects of the study. Verifiers can be used in a similar way to conventional abstract screening used in the SR process. The output should include clear and concise sentences reflecting a summary of pertinent information, enabling human reviewers to validate the exclusion decisions quickly. This method aims to replace conventional title/abstract screening with a more efficient mechanism. For example, the LLM can be prompted to output a summary of the study's objective, method, or results to enable the human reviewer to use the verifier to assess the relevance of the study.

(c) **Data Fields** are the prompted outputs that direct the LLM to focus on predetermined fields that align with the data extraction employed in SRs. For example, we gathered data on the population and context of the study design, such as "*number of participants*" and "*study duration*". Identifiers may be treated as data fields, as well.

The LLM prompts also need to be tailored to maximize the accuracy of the extracted information and to output this information in a usable format. This is accomplished by providing additional instructions in the prompts to the LLM to refine how it should extract data and output the extracted information [25]. These instructions should be clear and specific to minimize the risk of incorrect or irrelevant data extraction, adhering closely to the SR's protocol. For example, "*Act as a researcher, extract the following information from a research article above*". Then, the researcher can provide a list of required information.

Figure 3 shows an IVD prompt design to output in JavaScript Object Notation (JSON) format. A number of keys specify the identifiers, verifiers, and data fields with a short description of each. There are non-JSON entries as well, which are instructions for the LLM on how the output should be constructed or specific sections of the article to focus on.



```
Extract the following information from above. Focus on the Methods and Results sections. Structure the extracted information as a JSON. Use "NA" if information is not found.
JSON Fields:
"title": title of the study.
"author": first author's name.
"year": publication year.
"objective": look for a study's objective in the abstract section.
"method": summarizing the methodology in no more than two sentences.
"summary": summarizing the results or findings in no more than two sentences.
"participants": look for the sample size or demographics in the results section or the abstract.
"participantsOccupation": identify the occupation of the participants.
"studyType": string indicating where the data is collected.
"experimentForm": string describing the study design
"duration": The study's experiment duration. Look for the methods, study design and discussion section.
"work-relatedOutcome": identify the outcome of the study.
"devices" as an object has an array: name, sensor, placement, dataOutput, dataCollection.
"name": look for device names or descriptions in the methods section.
"model": look for model names or descriptions in the methods section.
"sensor": look for a sensor that was used from the methods section.
"placement": the device placement on participants from the methods section.
"movementVariables": The study's chosen output variable used to measure a participant's movement.
"dataOutput": direct quote the data output from the device. Look for the time and number in the study result section.
"dataCollection": look for the data collection methods in the data collection section or the web.
```

Figure 3. Initial testing of the IVD prompt input.

Formulating appropriate instructions for the verifier, identifier, and data fields is the key to the success of LLM extraction. Common prompting techniques, such as defining the scope of the task, might achieve higher accuracy. While the exact values may vary across studies, the relatively fixed structure of research articles offers opportunities to target keywords within specific sections (see Table 1). This table serves as a guideline for researchers to improve their prompts by defining the zones for data extraction from research articles.

Table 1. Potential article section for locating appropriate prompt shots.

Section of a Candidate Article	*Verifier*	*Identifier*	*Data Fields*
Abstract	High	Possible	Possible
Introduction	High	Possible	Unlikely
Literature review	Unlikely	Unlikely	Unlikely
Methodology	Unlikely	High	High
Experimental setup	Unlikely	High	High
Results	Unlikely	Unlikely	High
Discussions	Possible	Unlikely	Unlikely

For example, one of the prompts used in our case study as a verifier was "*objective*". We defined it as "*Look for the study's objective in the abstract section*" to instruct the LLM to look into the abstract section instead of other sections.

To facilitate the easy conversion of the extracted data into spreadsheets or other formats suitable for analysis, we also instructed the LLM to format outputs using JavaScript Object Notation (JSON). This offers a structured, machine-readable format, allowing for the transition into the data synthesis stages. For example, "*Structure the extracted information as a JSON. Use "NA" if information is not found*".

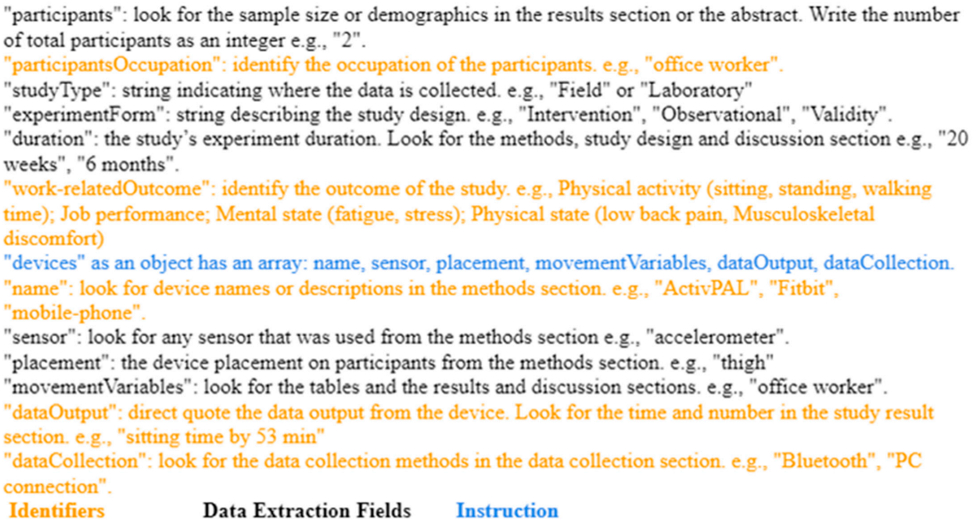
The prompts and instructions described above serve as the basis for prompt engineering, which effectively harnesses the capabilities of LLMs in the SR workflow. This ensures that the integration of AI tools complements and enhances the human component of the review process. At the end of this stage, we have an initial set of prompts for the LLM to extract data from a subset of articles (see Figure 3).

3.3.2. Prompt Testing and Refinement

There are several methods to refine the prompts to obtain more accurate data extraction. One method is few-shot prompting, which is a technique where a small number of examples are provided to the LLM to guide it toward the desired output [26]. A basic prompt can extract *verifiers* and *identifiers* and exclude many irrelevant studies. A refined prompt can further reduce labor time in manual data extraction by better handling synonyms and formats. The procedure for few-shot prompting is as follows:

1. Prepare initial few-shot prompts: Examples improve performance regardless of their correctness. This is not solely because of the accuracy of each individual example but also because they assist in formatting the outputs. Furthermore, we demonstrate that using "examples" explicitly leads to better results compared with not using examples [27]. Therefore, random examples could be added based on user experience.
2. Design initial few-shot prompt based on a relevant SR example: Develop prompts that concisely instruct the LLM on the task, incorporating the few-shot examples to illustrate the expected output. An LLM application could be utilized by providing the instruction, "You need to improve the following prompt to extract information from a research article," followed by the initial draft prompt.
3. Test the initial few-shot prompt: Input an initial few-shot prompt on a limited number of full-text articles. Review the outputs to improve the prompt with more relevant examples from the extraction.
4. Implement more broadly: Once the few-shot prompting is optimized, implement them for the broader data extraction task across the full dataset. Monitor the outputs to verify that the LLM maintains performance consistency as it processes varying articles.

Prompt refinement through few-shot prompting can result in more useful and accurate prompt outputs, especially identifier fields. Thus, reducing errors and inconsistencies in the data output, and, consequently, the amount of data cleaning required. For example, in Figure 4 our refined prompt contains the field "*experiment location*"; with few-shot prompting, we were able to add the examples of "*laboratory*", "*field*", and "*laboratory and field*", as these frequently appeared after our initial few-shot prompt. At the end of this stage, we had a prompt consisting of headings, their definitions, and a preview of data output to execute and collect outputs for further synthesis and catalog (see Figure 4).



"participants": look for the sample size or demographics in the results section or the abstract. Write the number of total participants as an integer e.g., "2".
"participantsOccupation": identify the occupation of the participants. e.g., "office worker".
"studyType": string indicating where the data is collected. e.g., "Field" or "Laboratory"
"experimentForm": string describing the study design. e.g., "Intervention", "Observational", "Validity".
"duration": the study's experiment duration. Look for the methods, study design and discussion section e.g., "20 weeks", "6 months".
"work-relatedOutcome": identify the outcome of the study. e.g., Physical activity (sitting, standing, walking time); Job performance; Mental state (fatigue, stress); Physical state (low back pain, Musculoskeletal discomfort)
"devices" as an object has an array: name, sensor, placement, movementVariables, dataOutput, dataCollection.
"name": look for device names or descriptions in the methods section. e.g., "ActivPAL", "Fitbit", "mobile-phone".
"sensor": look for any sensor that was used from the methods section e.g., "accelerometer".
"placement": the device placement on participants from the methods section. e.g., "thigh"
"movementVariables": look for the tables and the results and discussion sections. e.g., "office worker".
"dataOutput": direct quote the data output from the device. Look for the time and number in the study result section. e.g., "sitting time by 53 min"
"dataCollection": look for the data collection methods in the data collection section. e.g., "Bluetooth", "PC connection".

Figure 4. Refined IVD prompts with few-shot examples.

In each iteration, the output is in the form of JSON with prompts as keys and extracted information as their values. All these keys are then aggregated into a CSV with the keys as column headers. This CSV will have at most k rows for k seed articles. Based on this CSV, the quality of the prompt is determined, and it is decided whether to stop this prompt development. If the prompt development is stopped, then the final prompts are used on the whole set of n articles.

The proportion of identifiers, verifiers, and data field prompts could vary across different cases. However, the proposed IVD prompt techniques can be generalized as they can be tailored to various fields of study. Regardless of the research area, the human user would look for short, precise, directly extracted information (identifiers and data fields) along with short, summarized texts for other factors (verifiers).

3.3.3. Assessment of the Prompt Development and LLM Outputs

There are multiple methods to assess the success of prompt development, which can be quickly implemented and inform the iterative improvement of the prompts. One method is to measure the prompt output completeness, which is defined as the percentage of articles for which the LLM was able to find meaningful information for all the prompts that were asked for in the input JSON. Typically, a good set of prompts will return a very high level of completeness for the identifiers and data fields. A second measure of prompt success is to quantify the percentage of responses that align with the intent of a prompt. For example, a response describing a location aligns with a prompt asking for an experiment's location. There are diminishing returns to multiple prompt development cycles. However, they do not have to be 100% *complete* or *aligned* before proceeding to the next stage.

This process of refinement with more field-specific examples is small in terms of time consumed and can be performed after reviewing the previous LLM outputs. While processing thousands of articles may take several hours, the advantage lies in the autonomy of the LLMs, which operate without the need for continuous human input.

3.4. LLM Data Extraction and Formatting: JSON and CSV

Once the IVD prompts are finalized, they are used to extract information for all the articles (not just the seed articles). Then, the research has to consolidate all the JSON outputs for each article retrieved from an LLM into a single CSV file with n rows.

The next step is to format the headings. This step requires particular attention to detail as it involves maintaining consistency across all data fields. The process is critical as it allows for easy filtering, ensuring the accuracy of the analysis, especially in meta-analysis. It was noted that the LLM may give an output keyword that is slightly different from the desired keywords mentioned in the input prompt. For example, output headings may contain a mix of plural and singular terms, e.g., "*devices*" and "*device*", as multiple devices may be found in one article. In this case, all device details should be ensured under the same header to simplify the spreadsheet for the next stage.

The JSON data were imported into Microsoft Excel to sort and filter the studies based on the identifiers and verifiers. At the end of this stage, we produced a CSV file that was ready for filtering and screening. The headers (columns) of the CSV file are the individual *identifiers*, *verifiers*, and *data fields* asked for in the prompt. Each row represents the extracted information for each article input to the LLM.

At this stage, the identifier columns will have a high level of completeness, assuming a proper set of prompts were given, and the articles are relevant to the SR. It is also expected that the verifier columns will have near 100% or 100% completeness. It is possible the data fields may have a lower level of completeness as not all articles report all the data field variables. The final LLM output's completeness and accuracy are dependent on the quality of the refined prompt.

3.5. Selecting the Articles: Decision-Making

Now, with the assistance of the LLM output, humans are prepared to make the decision to select or reject an article by filtering, screening, and cleaning the data. These steps can be viewed in the hybrid workflow in Figure 2b.

Identifier Filtering: Rows can be filtered out by one of the identifiers that do not match the selection criteria. For example, we excluded rows that contained keywords such as “bus drivers” or “factory workers” because they were not relevant to our study of desk-based workers. Articles that pass this stage go through the data-cleaning step.

Verifier Screening: Those articles that were excluded by the identifier filtering process are then screened based on the verifiers to check the validity of this decision. Humans manually screen verifiers, representing the traditional abstract screening process. This step helps catch any articles where identifiers were missed by the LLM during the extraction process. Articles that are included at this stage are also moved to the data-cleaning stage. By following these steps, researchers can filter and verify the extracted data with a high degree of efficiency. After this screening, an inclusion dataset is ready for data cleaning.

Data Cleaning: Following the preliminary steps of filtering and screening, a data cleaning procedure is undertaken. Data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in the output to improve data quality, ensuring the accuracy and validity of the analysis and insights derived from the LLM assistance [28]. Like the conventional SR process, human experts screen and extract data from full-text articles. With the LLM-assisted workflow, the full-text screening and data extraction task is simplified to human search, verification, and correction of extracted data. A key aspect of the data-cleaning process is handling missing values. This step involves manual data search and human input. Minimizing missing data is crucial to reducing workload, and this can be achieved by refining the prompt and using techniques such as few-shot prompting to refine the prompt further.

In terms of human time consumption, humans can save time reading through hundreds or possibly thousands of full-text articles to determine inclusion or exclusion for an SR. The time complexity can be defined as follows:

- The identifier filtering only looks through small phrases consisting of one to three words, which would typically be repeated in multiple rows. For example, identifiers that have the same values across multiple articles can be automated using standard spreadsheet software.
- Next, the verifier screening is performed on the articles that failed the identifier filtering steps. The verifiers summarize key aspects of the articles into a few sentences, replacing the need to read larger excerpts of the article to make inclusion/exclusion decisions faster. The amount of repetition for these fields is less likely than for identifiers, as each article presents a different approach and solution.
- Lastly, during data cleaning, the human has fewer articles to evaluate when the information was not found by the LLM. Assuming the prompts were good, the human already had a good idea of what to look for in a full-text article for data cleaning. Our experience demonstrated that any article with less than 50% completeness in the CSV row is unlikely to be related to the research topic and may be excluded.

4. Case Study

In this section, we will evaluate the effectiveness of the proposed methodology with the human-only method guided by PRISMA guidelines [29]. We tested our methodology with an ongoing traditional systematic review on *Tracking the Sedentary Behavior of Office Workers*.

4.1. Data Resources

We conducted a systematic search of four electronic databases (Scopus, PubMed, Web of Science, and IEEE) for articles published between January 2000 and February 2023 that used technology to objectively measure desk-based workers’ sedentary behavior such as sitting, standing time, and the sit–stand transition. We followed PRISMA guidelines for initial title and abstract screening and then performed full-text screening on the articles included based on the title and abstract screening. This process resulted in a list of articles included in the final SR and a list of excluded articles. A set of 390 articles were used to test the proposed hybrid methodology.

4.2. Data Preparation

We used Endnote and Zotero to download all available PDF files for each article. We manually retrieved the PDF files of the articles that were not found by the two software programs that met the inclusion criteria and marked the rest as excluded. This study used n = 390 articles. We predefined selection criteria based on the scope of the review in order to make inclusion/exclusion decisions.

4.3. Prompt Engineering

To develop an IVD prompt, we initially identified keywords and phrases from the research question and the inclusion/exclusion criteria for our SR. These identifiers were used to create a search string for relevant studies. The initial prompt included examples for identifying work-related outcomes "*work-relatedoutcome*" based on data from another SR [30].

Our initial prompt to "*identify the outcome of the study, e.g., work disability, work performance, mood state, cognitive function, disability at work, work ability, sick leave*" was tested on a set of 50 seed articles, i.e., known PDF files that already met our selection criteria and were highly likely to be included in the SR. The output files from the initial prompt were used to optimize the prompt using a few-shot prompt strategy.

Our first output produced terms such as "*prolonged sitting and physical activity in the workplace*", "*occupational sitting time*", "*sedentary behavior, physical activity*", "*time-in-bed, sedentary time, physical activity*", "*reduced sedentary time*", and "*physical fitness*". We then refined the "*work-relatedOutcome*" field into "*identify the outcome of the study, e.g., physical activity (sitting, standing, walking time); job performance; mental state (fatigue, stress); physical state (low back pain, musculoskeletal discomfort)*".

The LLM's efficiency in processing articles was notable (Table 2): it took approximately 7.5 min to process a batch of 50 articles, with shorter batches requiring proportionately less time. All attempts with the LLM, using Gemini-Pro 1.0, handled articles ranging from 5000 to 12,000 words without exceeding a 9-min processing time.

Table 2. Time consumption for generating the LLM outputs.

k	Time to Process
50 articles	7.5 min
25 articles	3.5 min
10 articles	1.5 min

During our trials, we encountered some inconsistencies because of technical issues with the LLM on Google Cloud, resulting in unprocessed articles. However, we collated the outputs for all the different executions to ensure that we considered the prompt outputs for a given set of articles consistently. At the end of this extraction, there were 45 articles with consistent output, the actual seed articles. Following on from Figures 3 and 4, we report the outcomes for the two levels of prompts in Table 3 as follows:

- Initial prompt: where no examples were provided (see Figure 3).
- Refined prompt (few-shot): have examples directly from seed articles (see Figure 4).

In the table, we recorded the instances as follows:

- x is the total number of distinct terms (values) for a prompted output, e.g., identifier. This number may be high if the articles used a range of terms to describe a certain aspect of the study or if the prompt results in many incorrect responses.

- y is the number of unaligned, i.e., incorrect, or unwanted values from x distinct values. A smaller value reflects fewer erroneous responses and, thus, a better prompt.
- z is the number of rows in the spreadsheet, i.e., articles with unwanted values from y. A smaller value results from a better prompt.

Table 3. Performance of the LLM prompt development strategy, with the few-shot strategy.

(k = 45)	***Identifiers***									***Data Fields***					
Prompts	***"Occupation"***			***"Study Type"***			***"Work-Related Outcome"***			***"Device.Name"***			***"Device.Placement"***		
	x	y	z	x	y	z	x	y	z	x	y	z	x	y	z
Initial prompt	21	2	11	27	12	22	24	2	22	25	1	20	25	7	21
Refined prompt	18	0	0	4	2	4	18	0	0	34	0	0	23	1	7

Our observations indicated that the repeated use of identical prompts and prompt settings resulted in functionally identical CSV outputs, suggesting that the LLM generated stable outputs.

The data illustrated a decrease in variation and resultant uncertainty in human decision-making with the implementation of a more refined prompt design for the majority of the IVD prompts. This is because the x and y (and consequently z) consistently decrease in every iteration of prompt development, which means less filtering and reading for the human researcher. This improvement aligns with findings from the previous literature on prompt engineering [27,31]. With better prompt designs, the values in the columns turn more decisive.

Note that some data fields, such as the device name, saw an increase in x, while y and z became 0. This happened as the refined prompt was able to extract more individual names of devices properly without errors or unwanted values. In the initial prompt, the $y = 1$ was due to $z = 25$ rows with *blanks* for device names when the LLM was not able to find any suitable values. Thus, a better prompt should result in a decrease in y and z, but x may increase or decrease depending on the nature of the data.

We only used 13% of the total articles (50/390) for our case study. For any other hybrid SR, the researcher only needs to choose a smaller subset of the total set. This approach can work better for thousands of articles, whereas a much smaller number of articles can still work to generate a good prompt.

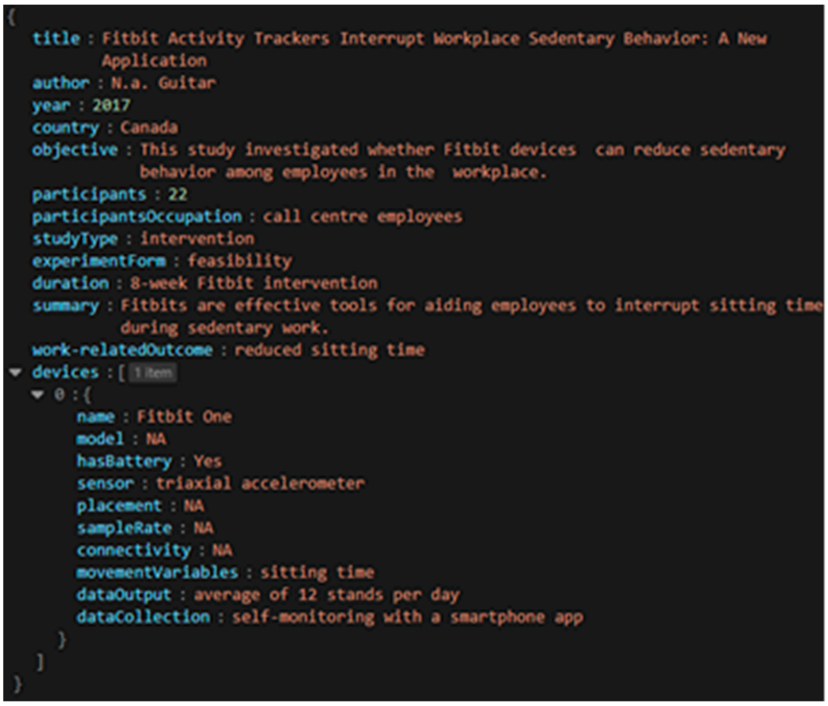
4.4. Evaluation Design

We designed three phases to evaluate the effectiveness of the LLM application as follows:

- Phase 1A (human-only): We followed the PRISMA flow diagram to screen articles for the SR. We applied predefined inclusion and exclusion criteria to select relevant articles. Each article was screened independently by two reviewers, and any conflict was resolved by team discussion. The results were divided into two groups as follows: inclusion (n = 170) and exclusion (n = 220). We used predefined data extraction templates to collect information from each article.
- Phase 1B (hybrid): We used a plugin with Gemini-Pro API (Top-p = 0.1, Temperature = 0.2) to extract data from the same articles. The Gemini-Pro model is an LLM created by DeepMind from Google. This LLM (version 1.0) has an input token limit of 30,720, which corresponds to about 18,432–24,576 English words [32]. Several prompts were engineered iteratively for articles based on the inclusion criteria and relevant keywords from the case study. Each prompt was sent to Gemini-Pro via the plugin and received the responses in JSON format and ran repeat three times because of the uncertainty in the AI output [10]. The data collection period was between 13 January 2024 and 2 February 2024. We developed an LLM application using an API to send the prompts to an LLM and collect the JSON outputs. The API was designed to handle

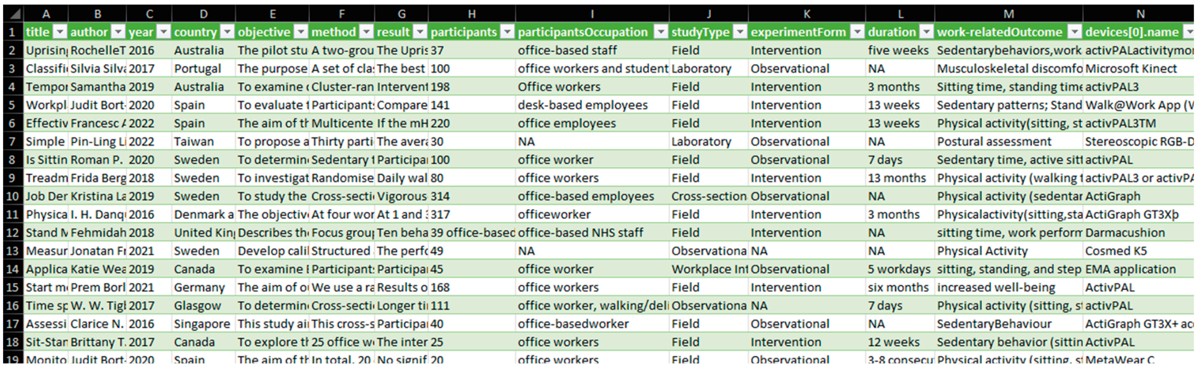
multiple requests simultaneously, which allowed us to collect the data quickly and efficiently. Once we had collected all JSON outputs, we stored them in a file (see Figure 5a). The data were pre-processed before the filter. After converting the JSON file to a spreadsheet using *"jsoneditoronline"*, one author filtered the data and removed any records according to the identifiers using Microsoft Excel (see Figure 5b). All records (n = 390) were filtered and screened by one author.

- Phase 2 (evaluation): One author validated the records from the hybrid workflow with the full-text articles of the included articles. Any conflicts between the hybrid and human-only methods were resolved by team discussion against the full-text articles to ensure that humans screened each study twice independently.



```
{
  title : Fitbit Activity Trackers Interrupt Workplace Sedentary Behavior: A New
          Application
  author : N.a. Guitar
  year : 2017
  country : Canada
  objective : This study investigated whether Fitbit devices  can reduce sedentary
              behavior among employees in the  workplace.
  participants : 22
  participantsOccupation : call centre employees
  studyType : intervention
  experimentForm : feasibility
  duration : 8-week Fitbit intervention
  summary : Fitbits are effective tools for aiding employees to interrupt sitting time
            during sedentary work.
  work-relatedOutcome : reduced sitting time
▼ devices : [ 1 item
  ▼ 0 : {
        name : Fitbit One
        model : NA
        hasBattery : Yes
        sensor : triaxial accelerometer
        placement : NA
        sampleRate : NA
        connectivity : NA
        movementVariables : sitting time
        dataOutput : average of 12 stands per day
        dataCollection : self-monitoring with a smartphone app
      }
    ]
}
```

(**a**)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	title	author	year	country	objective	method	result	participants	participantsOccupation	studyType	experimentForm	duration	work-relatedOutcome	devices[0].name
2	Uprisin	RochelleT	2016	Australia	The pilot stu	A two-grou	The Upris	37	office-based staff	Field	Intervention	five weeks	Sedentarybehaviors,work	activPALactivitymo
3	Classifi	Silvia Silva	2017	Portugal	The purpose	A set of cla	The best	100	office workers and student	Laboratory	Observational	NA	Musculoskeletal discomfo	Microsoft Kinect
4	Tempor	Samantha	2019	Australia	To examine	Cluster-ran	Intervent	198	Office workers	Field	Intervention	3 months	Sitting time, standing time	activPAL3
5	Workpl	Judit Bort-	2020	Spain	To evaluate t	Participant:	Compare	141	desk-based employees	Field	Intervention	13 weeks	Sedentary patterns; Stand	Walk@Work App (V
6	Effectiv	Francesc A	2022	Spain	The aim of th	Multicente	If the mH	220	office employees	Field	Intervention	13 weeks	Physical activity(sitting, st	activPAL3TM
7	Simple	Pin-Ling Li	2022	Taiwan	To propose a	Thirty parti	The avera	30	NA	Laboratory	Observational	NA	Postural assessment	Stereoscopic RGB-D
8	Is Sittin	Roman P.	2020	Sweden	To determin	Sedentary t	Participa	100	office worker	Field	Observational	7 days	Sedentary time, active sitt	activPAL
9	Treadm	Frida Berg	2018	Sweden	To investigat	Randomise	Daily wal	80	office workers	Field	Intervention	13 months	Physical activity (walking t	activPAL3 or activPA
10	Job Der	Kristina La	2019	Sweden	To study the	Cross-secti	Vigorous	314	office-based employees	Cross-section	Observational	NA	Physical activity (sedentar	ActiGraph
11	Physica	I. H. Danq	2016	Denmark a	The objectiv	At four wor	At 1 and 3	317	officeworker	Field	Intervention	3 months	Physicalactivity(sitting,sta	ActiGraph GT3Xþ
12	Stand M	Fehmidah	2018	United Kin	Describes th	Focus grou	Ten beha	39 office-basec	office-based NHS staff	Field	Intervention	NA	sitting time, work perform	Darmacushion
13	Measur	Jonatan Fr	2021	Sweden	Develop calil	Structured	The perfc	49	NA	Observationa	NA	NA	Physical Activity	Cosmed K5
14	Applica	Katie Wea	2019	Canada	To examine I	Participant:	Participa	45	office worker	Workplace Int	Observational	5 workdays	sitting, standing, and step	EMA application
15	Start m	Prem Borl	2021	Germany	The aim of o	We use a ra	Results o	168	office workers	Field	Intervention	six months	increased well-being	ActivPAL
16	Time sp	W. W. Tigl	2017	Glasgow	To determin	Cross-secti	Longer ti	111	office worker, walking/deli	Observationa	NA	7 days	Physical activity (sitting, s	activPAL
17	Assessi	Clarice N.	2016	Singapore	This study ai	This cross-s	Participa	40	office-basedworker	Field	Observational	NA	SedentaryBehaviour	ActiGraph GT3X+ ac
18	Sit-Star	Brittany T.	2017	Canada	To explore th	25 office w	The inter	25	office workers	Field	Intervention	12 weeks	Sedentary behavior (sittin	ActivPAL
19	Monito	Judit Bort-	2020	Spain	The aim of th	In total, 20	No signif	20	office workers	Field	Observational	3-8 consecu	Physical activity (sitting, s	MetaWear C

(**b**)

Figure 5. (**a**) Example output from Gemini-Pro. (**b**) Converted data ready for filtering.

4.5. Results

This section includes a brief inspection of the results from our case study and an analysis of the performance of the proposed workflow.

4.5.1. Observations from Filtering and Screening

Identifiers were developed using specific selection criteria to ensure consistency and accuracy in identifying relevant studies. In this context, the "*Field of Occupation*" was specifically designed to filter out studies that did not involve "*desk-based workers*". Furthermore, the "*Fields of Device Name*" and "*Sensor*" were designed to filter out studies that did not use "*objective activity measurement devices*" or that only used "*pedometers*". This research-specific filtering allowed humans to more efficiently move through this time-consuming process. Additionally, we prompted "*Field of Movement Variables*" to filter out studies that only measured changes in posture or that focused on an ergonomic assessment of posture. To exclude a subset of research studies considered "protocol articles" from our SR, we used verifiers to remove all articles that used future tense in the method section. Figure 6 illustrates the implementation of the proposed method.

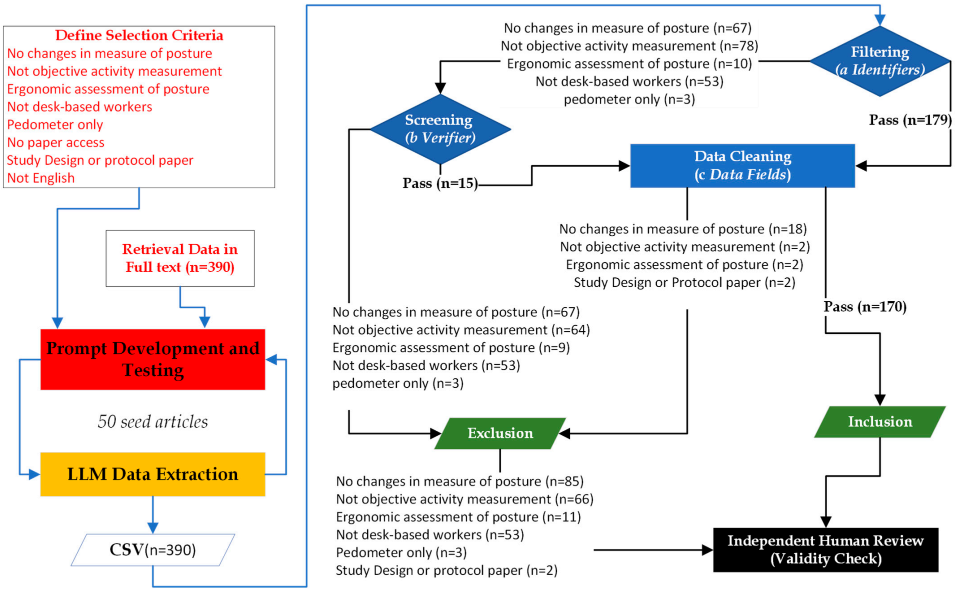
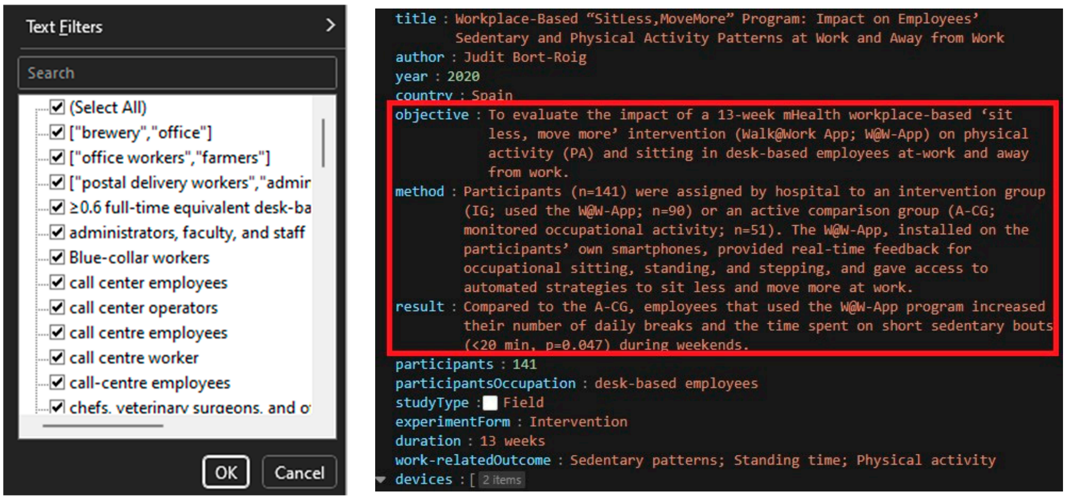

Define Selection Criteria
No changes in measure of posture
Not objective activity measurement
Ergonomic assessment of posture
Not desk-based workers
Pedometer only
No paper access
Study Design or protocol paper
Not English
Retrieval Data in Full text (n=390)
Prompt Development and Testing
50 seed articles
LLM Data Extraction
CSV(n=390)
No changes in measure of posture (n=67)
Not objective activity measurement (n=78)
Ergonomic assessment of posture (n=10)
Not desk-based workers (n=53)
pedometer only (n=3)
Filtering (a Identifiers)
Pass (n=179)
Screening (b Verifier)
Pass (n=15)
Data Cleaning (c Data Fields)
No changes in measure of posture (n=18)
Not objective activity measurement (n=2)
Ergonomic assessment of posture (n=2)
Study Design or Protocol paper (n=2)
Pass (n=170)
No changes in measure of posture (n=67)
Not objective activity measurement (n=64)
Ergonomic assessment of posture (n=9)
Not desk-based workers (n=53)
pedometer only (n=3)
Exclusion
Inclusion
No changes in measure of posture (n=85)
Not objective activity measurement (n=66)
Ergonomic assessment of posture (n=11)
Not desk-based workers (n=53)
Pedometer only (n=3)
Study Design or protocol paper (n=2)
Independent Human Review (Validity Check)

Figure 6. Flowchart showing the actual process of the SR in the case study.

Figure 7 illustrates that filtering identifiers (a) are fast compared with screen verifiers (b). We had a very high success number (*n* = 179) of articles in filtering identifiers to identify exclusion; only a small number of articles (*n* = 15) were missed and changed to inclusion during verifier screening. This passed only 194 articles to the data-cleaning phase. The CSV generated by this hybrid process is equivalent to forms filled out for conventional human-only SR. Here the data are in JSON format, while in the human-only SR, the same data are in tabular format.



☑ (Select All)
☑ ["brewery","office"]
☑ ["office workers","farmers"]
☑ ["postal delivery workers","admir
☑ ≥0.6 full-time equivalent desk-ba
☑ administrators, faculty, and staff
☑ Blue-collar workers
☑ call center employees
☑ call center operators
☑ call centre employees
☑ call centre worker
☑ call-centre employees
☑ chefs. veterinarv suraeons. and o

```
title : Workplace-Based “SitLess,MoveMore” Program: Impact on Employees’
        Sedentary and Physical Activity Patterns at Work and Away from Work
author : Judit Bort-Roig
year : 2020
country : Spain
objective : To evaluate the impact of a 13-week mHealth workplace-based ‘sit
            less, move more’ intervention (Walk@Work App; W@W-App) on physical
            activity (PA) and sitting in desk-based employees at-work and away
            from work.
method : Participants (n=141) were assigned by hospital to an intervention group
         (IG; used the W@W-App; n=90) or an active comparison group (A-CG;
         monitored occupational activity; n=51). The W@W-App, installed on the
         participants’ own smartphones, provided real-time feedback for
         occupational sitting, standing, and stepping, and gave access to
         automated strategies to sit less and move more at work.
result : Compared to the A-CG, employees that used the W@W-App program increased
         their number of daily breaks and the time spent on short sedentary bouts
         (<20 min, p=0.047) during weekends.
participants : 141
participantsOccupation : desk-based employees
studyType : □ Field
experimentForm : Intervention
duration : 13 weeks
work-relatedOutcome : Sedentary patterns; Standing time; Physical activity
devices : [ 2 items
```

(**a**) (**b**)

Figure 7. (**a**) Example Field of Occupation. (**b**) Example Field of Verifiers.

4.5.2. Observations from Data Cleaning

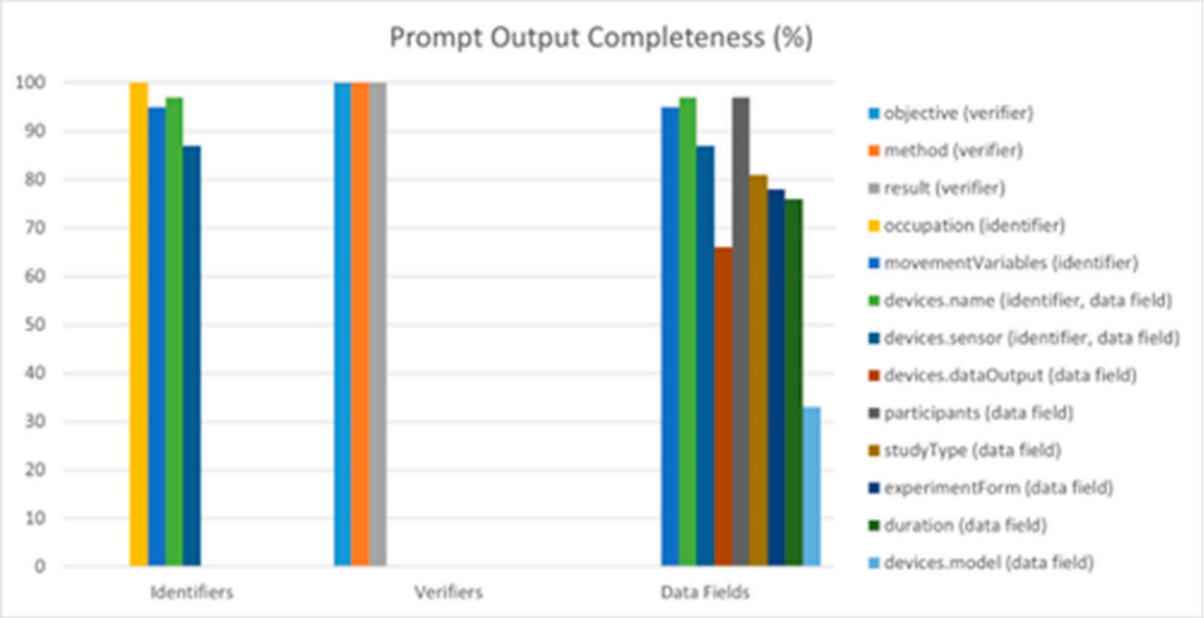
Cleaning the data after data extraction can be a time-consuming process. We applied a few strategies to refine our IVD prompts to reduce these time constraints. Our data-cleaning process incorporated the following steps:

- Standardizing formats: This involved applying uniform names to terms that may have several variations. For example, we standardized "activPAL" from "activPAL3", "ActivPAL3", "activPAL3c", "activPAL3", and "activPAL 3 micro" to reduce the sensitivity of the search.
- Handling missing values: Certain headings, such as "*devicemodel*" had a high prevalence of missing data. The LLM could not differentiate between a device's name and its model. Extracting such missing values involved identifying articles that met the selection criteria, particularly those with missing values in the identifiers, and searching for the specific missing information.
- Error removal: This involved identifying and rectifying errors in the output. Specific fields that required attention included the subjective measurements and the device model. These fields were reviewed for any inaccuracies or inconsistencies and corrected with information from the original article.

Although LLM output accuracy to assist with SR heavily relies on specific prompt engineering and may vary across discipline areas, we obtained a high efficiency for identifying the inclusion (n = 170) from full-text review articles (n = 194) simply by filtering and screening keywords and concise sentences from the LLM output.

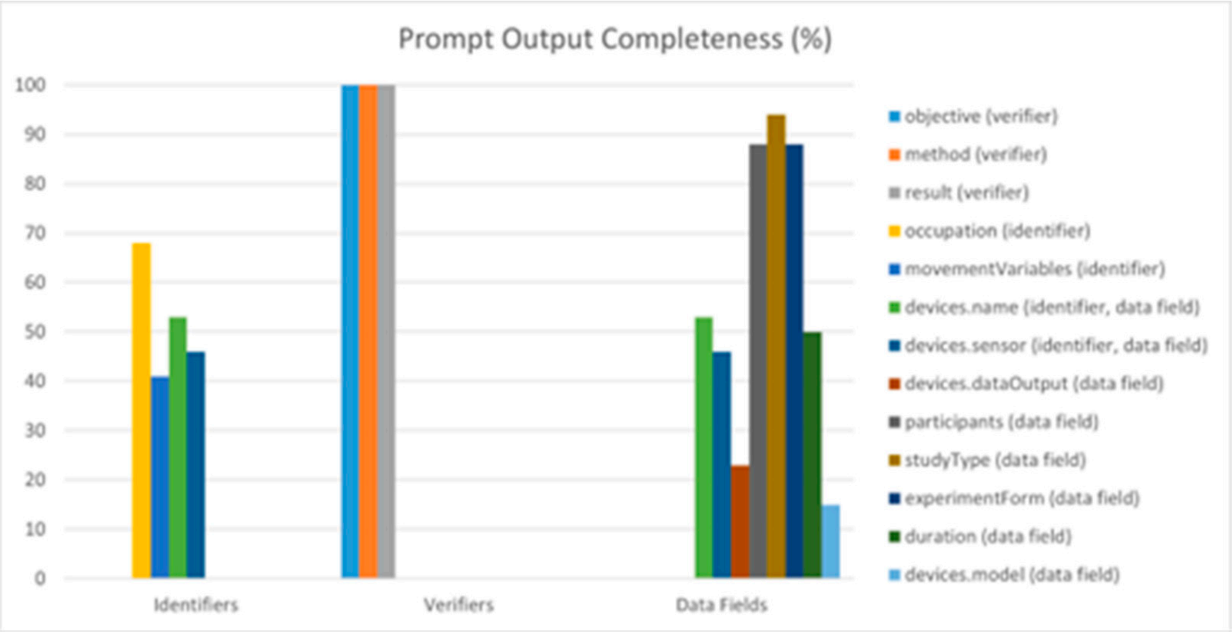
Figure 8 illustrates the efficiency of prompt output for the inclusion sector (a) and exclusion sector (b) once the prompt development and LLM executions finished. We calculated completeness as the number of rows, i.e., the article for which a value was found by the LLM in the full-text article for the specific prompt/CSV headers, compared to the total number of rows. So, for a prompt p:

$$completeness\ (p) = \frac{Number\ of\ articles\ with\ a\ valid\ extraction\ for\ p}{Total\ number\ of\ articles} \times 100 \quad (1)$$



(**a**)

Figure 8. *Cont.*



(**b**)

Figure 8. Efficiency of data extraction for the (**a**) inclusion sector and (**b**) exclusion sector.

Identifiers have a high rate of completeness, meaning the process of identifier filtering works with high accuracy, thus reducing the human workload during data cleaning. It is not surprising that all verifiers used in this case study (see Figure 3) had perfect completeness (100%) when asking for article summaries. The data fields (some of which also acted as identifiers) are relatively less filled in the CSV but still have over 50% completeness for most fields of the *included* articles. On the other hand, the *identifier* and *data field* completeness are considerably lower for the same prompts for the articles that were ultimately excluded. This shows that the LLM largely succeeds in extracting the desired data from the full-text articles, largely reducing the human's responsibility to screen the full-text articles.

Prompt development and testing are an iterative process during which humans do not have a clear idea of which article is included or not. However, as we started with a set of 50 articles that strongly met the selection criteria, we expected that at the end of the prompt development and testing, the prompts would have a very high level of completeness.

To provide an additional estimate of the time-saving capacity of our hybrid workflow, we measured the character length of the identifiers and verifiers (Table 4), comparing the speed at which human readers may be expected to perform the filtering steps using the LLM output compared to performing a full-text review to determine article inclusion or exclusion for an SR. Identifiers had a range of 10–25 characters, which were much smaller than the verifiers (150–300 characters). Both of these could be considered negligible compared with reading a full-text article with thousands of words.

Table 4. Lengths of data extractions for some prompts (in terms of characters).

Prompt	Type	Average	Std. Deviation
objective	verifier	184.90	61.29
method	verifier	294.81	108.76
result	verifier	291.91	121.40
occupation	identifier	18.49	11.88
movementVariables	identifier	15.53	9.45
devices.name	identifier, data field	15.52	9.47
devices.sensor	identifier, data field	17.29	20.47

4.5.3. Revised Human-Only Decision

The performance outcome of the hybrid workflow is shown in Table 5.

Table 5. Comparison of the two methods for inclusion/exclusion classification.

(n = 390)	***LLM-Assisted Hybrid Workflow***		***Traditional Workflow***	
	Correct	***Incorrect***	***Correct***	***Incorrect***
Inclusion	170	0	167	3
Exclusion	220	0	217	3

The hybrid workflow inclusion/exclusion decisions agreed with the human-only process in 384 out of 390 articles.

The hybrid workflow also improved the accuracy of the study selection process by overturning six decisions that were initially erroneous in the human-only SR. In the inclusion scenario, three (3/170 articles) papers were originally excluded by human reviewers, but the hybrid workflow identified that they did meet the inclusion criteria, which was incorrectly missed by the human-only SR. Upon further discussion and re-evaluation by the review team, it was agreed that these papers were indeed relevant and should have been included in the SR.

In the exclusion scenario, there was a total of 220 exclusions in the hybrid method. There was 100% (114/114 articles) agreement with the articles that the human process excluded during the *title* and *abstract* screening in the human-only SR. For the rest, the human reviewers initially included several papers, but the hybrid workflow flagged three (3/106 articles) of these as not meeting the necessary criteria. The review team's deliberation confirmed that these papers were not appropriate for inclusion and should be excluded.

Ultimately, for the LLM-assisted workflow, there was a total of 6/390, i.e., a 1.53% mismatch, compared with the fully human-performed SR. However, in each of the six cases, the proposed workflow improved the results.

5. Discussion

The hybrid workflow described in this paper helped our research team improve the accuracy and efficiency of our study selection by enabling us to identify relevant studies with greater precision. Nonetheless, the LLM application was not free from challenges in terms of efficiency and dependence on effective prompt engineering. Effective prompt design has the potential to speed human decision-making and reduce labor time during processes such as data cleaning.

5.1. How Can Prompt Engineering Be Optimized to Leverage the Capabilities of LLMs in Extracting Relevant Information from Full-Text Articles in SRs?

Our hybrid workflow, which integrated an LLM into the SR process to validate human-only SR results, demonstrated an increase in efficiency and accuracy by combining the strengths of LLMs in efficient and accurate information extraction with the critical thinking and decision-making capabilities of human reviewers. This integration capitalizes on the computational power of LLMs to swiftly process vast quantities of text, thereby minimizing the extensive reading typically required of human reviewers. Consequently, the human element of the workflow is refined to focus on higher-level evaluative and interpretive tasks.

For instance, in the case study, we demonstrated some potential for using LLM to extract occupation information about study participants. Tasks such as identifying sensor names and their placement on the body, which once demanded considerable time and meticulous work from human reviewers, were executed with precision by the LLM. This illustrates the LLM's capacity for reducing the manual burden of data extraction, allowing researchers to allocate more time to critical analysis. However, the case study also revealed limitations in the LLM's logical interpretation. When the LLM classified survey-based studies as employing "objective measurement" devices, it highlighted the model's difficulty with some complex reasoning and context-understanding tasks. This example underscores the necessity for human oversight to correct such logical discrepancies and ensure the review's criteria are met accurately. After data cleaning, any records lacking value in objective measurement can be excluded.

LLMs excel in rapidly sifting through and extracting pertinent details from full-text articles, a process that would otherwise be laborious and time-consuming for human reviewers. By leveraging the advanced natural language processing abilities of LLMs, researchers can swiftly gather data points that are relevant to their specific inclusion and exclusion criteria. Even with a poorly defined prompt, it is effective in saving time, and humans can filter out a large number of irrelevant studies. For instance, in the case study, we used the prompt "*participantsOccupation*" to quickly filter records unrelated to desk-based workers, excluding those with occupations such as bus drivers, blue-collar workers, factory workers, etc. This was achieved by quickly sorting and filtering the CSV file using basic sorting tools in Microsoft Excel. This streamlined the process by quickly identifying irrelevant studies, allowing us to focus on our population of interest. This efficiency improvement is not achieved at the expense of accuracy. Human reviewers' oversight ensures that the extracted information is verified and validated, maintaining the integrity of the review process.

5.2. What Is the Impact of the Hybrid Workflow in Conducting SRs?

The effectiveness of a hybrid workflow that combines the computational prowess of LLMs with the critical oversight of human reviewers can be substantial in reducing the workload of human reviewers while ensuring comprehensive and accurate data extraction. LLMs can quickly scan through thousands of articles, identifying and extracting relevant data based on predefined criteria. This can drastically reduce the time the first author spends on the initial screening of articles, which is typically the most time-consuming part of the SR process.

By rapidly processing and narrowing down the selection of articles, LLMs enable a much faster review cycle. For example, the dramatic contrast between the months it took for a human-only method to complete full-text screening and the two weeks required for the first author to double-check the results is a testament to the efficiency of LLMs. By implementing a hybrid workflow, articles (n = 169) identified for exclusion had a perfect agreement with the human-only screening results. Furthermore, conflicts that were not originally identified by the human-only screening were reviewed by our team, resulting in altered decisions (n = 6). The hybrid workflow, with LLMs handling the initial data extraction, alleviates this burden by ensuring that human reviewers are not overwhelmed with information, allowing them to focus on more complex and ambiguous cases that require human interpretation.

Furthermore, this effectively addressed the temporal lag that often undermines the timeliness of conventional SRs. With this accelerated process, the most recent and relevant research can be incorporated into the SR, ensuring that the findings reflect the latest developments in the field. This rapid inclusion of up-to-date research is crucial for maintaining the relevance and utility of SRs in informing current practice and policy decisions.

The aim of this hybrid workflow was to improve the paper classification procedure while extracting all the meaningful information needed to write a full SR article. The evaluation showed that the information extraction and inclusion/exclusion classification for writing an SR paper was at least equivalent and possibly better for the hybrid workflow SR compared with the traditional SR, as shown in Table 5, i.e., an improvement in classification accuracy. The human decision on what to do with the information extracted and the included articles, as part of the hybrid SR (or traditional SR), was out of the scope of this paper.

5.3. Comparison with the Human-Only Approach

There are certain differences in the way humans operate in the workflow compared with a completely human-performed SR.

i Conventional SR methodology might include a *snowballing* strategy. It is a technique used to identify additional relevant studies that may not have been captured by the initial search strategy [13]. However, this manual process can be time-consuming and labor-intensive, especially for SRs that deal with a large volume of studies [33].

In the case study, the snowballing process was not used as much. Instead, the researchers collected a large number of articles based on matching keywords and then passed them to the LLM for summarization or extraction. Because LLMs can handle many articles in a short period of time, researchers do not have to worry about the possible chain of connections between older and newer articles.

An advanced LLM could be deployed to re-extract the subset of studies that passed the previous screening. An advanced LLM is more adept at recognizing the targeted context within the studies, leading to more accurate and relevant data extraction. Consequently, this reduces the variables that researchers need to handle and minimizes the need for extensive data cleaning.

ii The hybrid workflow also reduces the chances of bias in selecting or reading articles. If the process is completed manually, researchers could pay less attention to certain articles based on meta-information. In the proposed hybrid workflow, this is not possible as all the articles are handed over to the LLM, which treats all equally.

5.4. Limitations and Future Work

Future work can improve certain aspects or provide more insight with further analysis. One limitation of this case study is that the accuracy of data extraction was not quantified. A common methodology in SR is to have one author extract the data fields from each study and then have a second author verify the extracted data. In the case study presented here, the LLM performed the initial data extraction, but the accuracy of this extraction was not verified. As commonly expected for LLMs in these types of applications, it was observed that the LLM did not manufacture content outside of what was in the articles for data fields and identifiers. However, this accuracy needs to be assessed in future research to determine if the human verification steps can be further minimized.

Current SR rating scales such as AMSTAR [34] are designed to assess traditional SR methodology. In the case of the hybrid workflow, two questions from the AMSTAR metric (5 and 6) that require researchers to duplicate study selection and data extraction may not be satisfied. Future work may demonstrate that the hybrid workflow is equivalent or superior to the duplicate procedures suggested in AMSTAR.

In addition, the choice of LLMs affects the accuracy and completeness of the extracted data. Different LLM models may have different strengths and weaknesses in terms of domain knowledge, reasoning ability, and natural language understanding. Therefore, selecting the most suitable model for a given task is not trivial and may require extensive evaluation.

Designing effective prompts requires trial and error. The quality of prompt design and the performance of LLMs can impact the efficiency of this method and the amount of manual labor needed for data cleaning. To overcome these limitations, the following directions can be explored in future work. Firstly, applying a multi-agent conversation framework where LLMs collaborate to provide improved quality of prompt writing and output completeness. Secondly, a snowballing strategy can be developed utilizing prompts and LLMs. Thirdly, our contribution includes the development of a swift human validation technique. More fine-tuned use of identifier and verifier terms can be used to verify the accuracy of automated classification approaches.

One technical limitation of the online LLM services was the smaller batches it worked in and the technical failures to produce the outputs for all articles in a batch because of timeouts. This required additional time to collate the prompt developments on the seed articles.

Another limitation encountered during the case study was related to pdf pre-processing. Certain device-related details were absent in the extraction process either because they were not explicitly mentioned in the text or were presented exclusively in a non-textual format (e.g., graphs, charts). We can enhance the extraction tools by exploring PDF pre-processing solutions using multimodal like Gemini-Pro-Vision. This model's image description capabilities may allow for the extraction of valuable information from visual elements within articles, potentially streamlining the data-cleaning process. To assess the external validity of our approach, one can further conduct tests with LLMs such as GPT-4 or GPT-4-Turbo. These experiments will gauge their potential in automating aspects of data cleaning, aiming to optimize the efficiency of the SR process.

6. Conclusions

This article demonstrates a hybrid method to integrate LLMs with human input into SR processes. The results demonstrate the potential to reduce workload, increase accuracy, and decrease the usual temporal lag of producing an SR. This study may contribute to a transformative step in the evolution of evidence-based research methodologies. The strategic fusion of the sophisticated capabilities of LLMs with the critical and interpretive acumen of human experts has the potential to significantly bolster both the efficiency and accuracy of knowledge synthesis. This innovative hybrid approach not only addresses the reliability concerns associated with fully automated AI screening but also potentially navigates the constraints conventionally imposed by the scope and temporal lag inherent in the SR process.

This identifier and verifier approach is a significant development for human–AI teaming in the context of large fixed-format document processing applications using LLM. Because of the technical requirements of maintaining high-end infrastructure for LLM service, it is foreseeable that such a service will be provided by the cloud. This would allow users from various fields to use the services. As such, this prompting technique, which is suitable for non-programmers, can be widely used across several LLM-based applications.

Author Contributions: Conceptualization, A.Y.; methodology, A.Y., A.M. and M.S.; software, A.Y.; validation, A.M., M.S. and S.J.P.; formal analysis, A.Y. and A.M.; investigation, A.Y. and A.M.; resources, A.Y., M.S. and S.J.P.; data curation, A.Y.; writing—original draft preparation, A.Y. and A.M.; writing—review and editing, M.S. and S.J.P.; visualization, A.Y. and A.M.; supervision, A.M., M.S. and S.J.P. All authors have read and agreed to the published version of this manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this article. But we call for readers to follow https://github.com/AnjiaYe/pdfMinerPlus, accessed on 4 May 2024, for further updates and if readers want to collaborate and participate in further testing.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; Group, P. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Int. J. Surg.* **2010**, *8*, 336–341. [CrossRef] [PubMed]
2. Chalmers, I.; Haynes, B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ* **1994**, *309*, 862–865. [CrossRef] [PubMed]
3. Higgins, J.P.T.; Green, S. *Cochrane Handbook for Systematic Reviews of Interventions*; Wiley: Hoboken, NJ, USA, 2008.
4. Robinson, K.A.; Whitlock, E.P.; Oneil, M.E.; Anderson, J.K.; Hartling, L.; Dryden, D.M.; Butler, M.; Newberry, S.J.; McPheeters, M.; Berkman, N.D.; et al. Integration of existing systematic reviews into new reviews: Identification of guidance needs. *Syst. Rev.* **2014**, *3*, 60. [CrossRef] [PubMed]
5. Ahn, E.; Kang, H. Introduction to systematic review and meta-analysis. *Korean J. Anesthesiol.* **2018**, *71*, 103–112. [CrossRef] [PubMed]

6. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gotzsche, P.C.; Ioannidis, J.P.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ* **2009**, *339*, b2700. [CrossRef] [PubMed]
7. Borah, R.; Brown, A.W.; Capers, P.L.; Kaiser, K.A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **2017**, *7*, e012545. [CrossRef] [PubMed]
8. Michelson, M.; Reuter, K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp. Clin. Trials. Commun.* **2019**, *16*, 100443. [CrossRef] [PubMed]
9. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can large language models replace humans in the systematic review process? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *arXiv* **2023**, arXiv:2310.17526. [CrossRef]
10. Syriani, E.; David, I.; Kumar, G. Assessing the ability of ChatGPT to screen articles for systematic reviews. *arXiv* **2023**, arXiv:2307.06464. [CrossRef]
11. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
12. Goodyear-Smith, F.A.; van Driel, M.L.; Arroll, B.; Del Mar, C. Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: A case study. *BMC Med. Res. Methodol.* **2012**, *12*, 76. [CrossRef] [PubMed]
13. Tsafnat, G.; Glasziou, P.; Choong, M.K.; Dunn, A.; Galgani, F.; Coiera, E. Systematic review automation technologies. *Syst. Rev.* **2014**, *3*, 74. [CrossRef] [PubMed]
14. Aromataris, E.; Fernandez, R.; Godfrey, C.M.; Holly, C.; Khalil, H.; Tungpunkom, P. Summarizing systematic reviews: Methodological development, conduct and reporting of an umbrella review approach. *Int. J. Evid. Based Healthc.* **2015**, *13*, 132–140. [CrossRef] [PubMed]
15. Meline, T. Selecting studies for systemic review: Inclusion and exclusion criteria. *Contemp. Issues Commun. Sci. Disord.* **2006**, *33*, 21–27. [CrossRef]
16. Bannach-Brown, A.; Przybyła, P.; Thomas, J.; Rice, A.S.C.; Ananiadou, S.; Liao, J.; Macleod, M.R. Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst. Rev.* **2019**, *8*, 23. [CrossRef] [PubMed]
17. Yu, Z.; Menzies, T. FAST2: An intelligent assistant for finding relevant papers. *Expert Syst. Appl.* **2019**, *120*, 57–71. [CrossRef]
18. van de Schoot, R.; de Bruin, J.; Schram, R.; Zahedi, P.; de Boer, J.; Weijdema, F.; Kramer, B.; Huijts, M.; Hoogerwerf, M.; Ferdinands, G.; et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **2021**, *3*, 125–133. [CrossRef]
19. Marshall, I.J.; Wallace, B.C. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst. Rev.* **2019**, *8*, 163. [CrossRef] [PubMed]
20. Alshami, A.; Elsayed, M.; Ali, E.; Eltoukhy, A.E.E.; Zayed, T. Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems* **2023**, *11*, 351. [CrossRef]
21. Qureshi, R.; Shaughnessy, D.; Gill, K.A.R.; Robinson, K.A.; Li, T.; Agai, E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst. Rev.* **2023**, *12*, 72. [CrossRef] [PubMed]
22. Guo, E.; Gupta, M.; Deng, J.; Park, Y.J.; Paget, M.; Naugler, C. Automated paper screening for clinical reviews using large language models: Data analysis study. *J. Med. Internet Res.* **2024**, *26*, e48996. [CrossRef] [PubMed]
23. van Dijk, S.H.B.; Brusse-Keizer, M.G.J.; Bucsán, C.C.; van der Palen, J.; Doggen, C.J.M.; Lenferink, A. Artificial intelligence in systematic reviews: Promising when appropriately used. *BMJ Open* **2023**, *13*, e072254. [CrossRef] [PubMed]
24. de la Torre-López, J.; Ramírez, A.; Romero, J.R. Artificial intelligence to automate the systematic review of scientific literature. *Computing* **2023**, *105*, 2171–2194. [CrossRef]
25. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. *arXiv* **2021**, arXiv:2109.01652.
26. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv*, 2023; arXiv.2302.13971. [CrossRef]
27. Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv* **2022**, arXiv:2202.12837.
28. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data cleaning. In Proceedings of the 2016 International Conference on Management of Data, New York, NY, USA, 26 June–1 July 2016; pp. 2201–2206.
29. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
30. Lusa, S.; Punakallio, A.; Manttari, S.; Korkiakangas, E.; Oksa, J.; Oksanen, T.; Laitinen, J. Interventions to promote work ability by increasing sedentary workers' physical activity at workplaces—A scoping review. *Appl. Ergon.* **2020**, *82*, 102962. [CrossRef] [PubMed]
31. Wei, J.; Wei, J.; Tay, Y.; Tran, D.; Webson, A.; Lu, Y.; Chen, X.; Liu, H.; Huang, D.; Zhou, D. Larger language models do in-context learning differently. *arXiv* **2023**, arXiv:2303.03846.

32. Gemini, T.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805. [CrossRef]
33. Horsley, T.; Dingwall, O.; Sampson, M. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database Syst. Rev.* **2011**. [CrossRef] [PubMed]
34. AMSTAR Checklist. Available online: https://amstar.ca/Amstar_Checklist.php (accessed on 19 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.