





## Article

# Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions

Ahmad Alshami <sup>1</sup>, Moustafa Elsayed <sup>2</sup>, Eslam Ali <sup>3,4,\*</sup>, Abdelrahman E. E. Eltoukhy <sup>5,\*</sup> and Tarek Zayed <sup>3</sup>

<sup>1</sup> Department of Civil and Environmental Engineering, FAMU-FSU College of Engineering, Florida State University, Tallahassee, FL 32013, USA; aaa21a@fsu.edu

<sup>2</sup> Department of Civil and Environmental Engineering, FAMU-FSU College of Engineering, Florida A&M University, Tallahassee, FL 32013, USA; moustafa1.elsayed@famu.edu

<sup>3</sup> Department of Building and Real Estate, Faculty of Construction and Environment, The Hong Kong Polytechnic University, Kowloon TU428, Hong Kong; tarek.zayed@polyu.edu.hk

<sup>4</sup> Public Works Department, Geomatics Lab, Faculty of Engineering, Cairo University, Giza 12613, Egypt

<sup>5</sup> Department of Industrial and System Engineering, The Hong Kong Polytechnic University, Hung Hom TU428, Hong Kong

\* Correspondence: eslam.a.saleh@connect.polyu.hk (E.A.); abdelrahman.eltoukhy@polyu.edu.hk (A.E.E.E.)

**Abstract:** Systematic reviews (SR) are crucial in synthesizing and analyzing existing scientific literature to inform evidence-based decision-making. However, traditional SR methods often have limitations, including a lack of automation and decision support, resulting in time-consuming and error-prone reviews. To address these limitations and drive the field forward, we harness the power of the revolutionary language model, ChatGPT, which has demonstrated remarkable capabilities in various scientific writing tasks. By utilizing ChatGPT's natural language processing abilities, our objective is to automate and streamline the steps involved in traditional SR, explicitly focusing on literature search, screening, data extraction, and content analysis. Therefore, our methodology comprises four modules: (1) Preparation of Boolean research terms and article collection, (2) Abstract screening and articles categorization, (3) Full-text filtering and information extraction, and (4) Content analysis to identify trends, challenges, gaps, and proposed solutions. Throughout each step, our focus has been on providing quantitative analyses to strengthen the robustness of the review process. To illustrate the practical application of our method, we have chosen the topic of IoT applications in water and wastewater management and quality monitoring due to its critical importance and the dearth of comprehensive reviews in this field. The findings demonstrate the potential of ChatGPT in bridging the gap between traditional SR methods and AI language models, resulting in enhanced efficiency and reliability of SR processes. Notably, ChatGPT exhibits exceptional performance in filtering and categorizing relevant articles, leading to significant time and effort savings. Our quantitative assessment reveals the following: (1) the overall accuracy of ChatGPT for article discarding and classification is 88%, and (2) the F-1 scores of ChatGPT for article discarding and classification are 91% and 88%, respectively, compared to expert assessments. However, we identify limitations in its suitability for article extraction. Overall, this research contributes valuable insights to the field of SR, empowering researchers to conduct more comprehensive and reliable reviews while advancing knowledge and decision-making across various domains.

**Keywords:** ChatGPT; systematic review; automation; Internet of Things (IoT); article filtration; article categorization; information extraction; content analysis



**Citation:** 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. <https://doi.org/10.3390/systems11070351>

Academic Editor: William T. Scherer

Received: 8 June 2023

Revised: 4 July 2023

Accepted: 7 July 2023

Published: 9 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Review articles serve various purposes within the academic literature with different types, including narrative reviews, Systematic reviews (SR), meta-analyses, scoping reviews, and integrative reviews [1]. Narrative reviews provide a broad overview and

subjective analysis of existing literature [2], while SRs employ a thorough methodology to synthesize all relevant studies on a specific research question, ensuring objectivity and minimizing bias [3]. SRs offer several advantages, such as providing a reliable and comprehensive assessment of evidence, guiding evidence-based practice and policymaking, identifying research gaps, and enhancing statistical power through meta-analysis [4,5]. It is worth mentioning that SR articles are a valuable tool for synthesizing and analyzing research evidence in many fields of research, particularly in fields where research evidence is constantly evolving, such as in healthcare [6–9], project management [10–13], construction management [14–19], and aviation routing management [20,21]. To ensure that SRs are reported accurately and comprehensively, PRISMA (Preferred Reporting Items for SRs and Meta-Analyses) is widely used. Developing and executing a comprehensive search strategy to conduct an SR using the PRISMA method is essential.

The search strategy is vital in identifying relevant studies to be included in SRs. Such a strategy involves carefully selecting appropriate databases, applying pertinent Boolean research terms (BST) and keywords, and executing systematic searches to capture a comprehensive variety of evidence related to the research question [22]. In accordance to the PRISMA guidelines [23], inclusion and exclusion criteria are also crucial in the SR process. These predetermined criteria help assess the relevance of articles during the study selection phase, ensuring that the chosen studies align with the review's objectives and provide pertinent information to address the research question. Furthermore, snowballing is mainly applied to identify additional relevant articles that may have been missed in the initial literature search. The snowballing process can be achieved by gathering the articles from the references (backward) and citation (forward) lists of included studies [24]. However, it is important to acknowledge the PRISMA method's limitations, including potential reporting bias, the challenges of adapting to different review articles, human uncertainties in determining the article's eligibility, and the time consumed including and excluding articles from the database [25,26]. Despite these limitations, the SR process, PRISMA guidelines, and snowballing procedures significantly all contribute to evidence synthesis and knowledge advancement across various fields. With the continued advancement of AI-driven language and chatbot technologies, there is an increasing potential for automating the SR process through alternative methods. Leveraging these AI-powered tools offers opportunities to streamline the SR process, saving time and costs while addressing uncertainties arising from human responses. By exploring these possibilities, we can optimize workflows and enhance the overall efficiency of conducting SR.

ChatGPT (Generative Pre-trained Transformer) has proven to be a valuable tool in various fields, including healthcare [27–29], education [30–33], construction management [34,35], and scientific writing [36–38]. Within scientific writing, ChatGPT has proven its efficacy in generating abstracts, introductions, and research article summaries, while also assisting with SR processes by extracting relevant information and providing concise summaries [39,40]. Its capabilities as a powerful language model extends beyond simple language generation, offering valuable suggestions for structuring the article, enhancing clarity, and ensuring a logical flow [41]. Collaborating with ChatGPT empowers researchers to outline different manuscript sections, including the introduction, methods, results, and discussion, facilitating comprehensive and cohesive narratives [42]. Furthermore, ChatGPT's role extends to the editing and proofreading stages of scientific writing, serving as a meticulous grammar and language checker to adhere to the required style and formatting guidelines [43]. However, it is essential to recognize that while ChatGPT provides indispensable support, its usage should complement human expertise. Researchers must critically evaluate the model's outputs, thoroughly verify information, and ensure the accuracy and reliability of the generated content [44]. By combining the capabilities of ChatGPT with human insight, researchers can significantly enhance the efficiency, productivity, and overall quality of their research and scientific writing endeavors.

Despite the capabilities of ChatGPT in various aspects of scientific writing, there is no previous research focusing on automating the SR process by leveraging the power of

ChatGPT. However, a recent study by Qureshi [45] has raised important questions about the possibilities of ChatGPT in automating the SR process. It is worth mentioning that this study [45] just raised the question and discussed ChatGPT's capabilities in the SR process; however, they did not introduce a practical implementation of how we can do this by leveraging the ChatGPT. While acknowledging the outstanding capabilities of ChatGPT in automating the SR process, the study [45] recommended further research to investigate its limitations and capacities. Therefore, our paper aims to bridge this gap by harnessing the power of ChatGPT to introduce a practical implementation of the automated SR process. Our main focus is on streamlining the traditional process of SR and introducing practical implementations of ChatGPT at different stages of the SR process.

In order to showcase the practical implementation of our methodology, we delve into the extensive domain of Internet of Things (IoT) applications pertaining to water and wastewater management, as well as water quality monitoring. This subject holds significant importance due to the transformative impact of IoT in these particular domains. By undertaking this exploration, we contribute to the automation of the systematic review (SR) process, which can be applicable to various research fields, and provide valuable insights into the current state of IoT technologies in these critical areas.

Our approach encompasses a series of well-designed steps, commencing with a comprehensive and systematic search across relevant databases. Subsequently, we employ stringent filtering and extraction techniques to extract the most pertinent information from the collected literature. This is followed by a thorough content analysis of the selected studies, enabling us to unveil patterns, identify emerging trends, and gain a holistic understanding of the overall landscape regarding IoT applications in water management and water quality monitoring. By harnessing the capabilities of ChatGPT technology, we can leverage its natural language processing capabilities to streamline the analysis process and unveil concealed connections within the research corpus.

It is important to emphasize that while this paper outlines the methodology for conducting an SR, it does not delve into the specific findings regarding IoT applications in water management and water quality monitoring. Instead, the findings will be meticulously documented and published separately, allowing for a comprehensive exploration of this dynamic and critical area. The detailed objectives of the study can be summarized in the following points:

- ✓ To investigate the potential of ChatGPT in generating relevant keywords and phrases for literature search in water and wastewater management applications and water quality monitoring.
- ✓ To compare the accuracy and efficiency of utilizing ChatGPT for screening and filtering studies to be included in an SR, in contrast to conventional methods.
- ✓ To assess the completeness and accuracy of employing ChatGPT in extracting and synthesizing information from abstracts and full-text articles of the selected studies.
- ✓ To compare the quality and rigor of the SR process when utilizing ChatGPT against traditional SR methods. This comparison will consider various metrics, including reproducibility, bias, and transparency.
- ✓ To provide comprehensive guidance on the best practices for integrating ChatGPT into the methodology of SRs specifically focused on water and wastewater management.

To fulfil the objectives of this study, a novel methodology is devised to integrate ChatGPT into the SR procedure, and its performance is compared against traditional SR approaches. This paper makes a valuable contribution to the existing body of knowledge on utilizing artificial intelligence (AI) in advancing SR methodologies by presenting an innovative approach that leverages ChatGPT (based on the GPT-3.5 architecture model) to enhance the overall process. The proposed methodology is employed to conduct an SR article focusing on IoT applications in water and wastewater management. Furthermore, the implications and limitations of this methodology for future research endeavors in the field are thoroughly examined and discussed.

## 2. Research Methodology

### 2.1. Exploring ChatGPT: Characteristics and Interactions

ChatGPT is a powerful language model that is specifically designed to facilitate interactive conversations and simulate human-like dialogue. It is built upon the foundation of GPT-3.5, an advanced variant of the GPT-3 model developed by OpenAI. ChatGPT leverages the enhancements and refinements introduced in GPT-3.5, which include improved natural language understanding, longer consecutive output, and better adherence to instructions. By utilizing ChatGPT, our study benefits from its ability to retain context from previous interactions, allowing for more coherent and context-aware responses. This feature enables ChatGPT to generate high-quality and engaging conversational experiences, making it an ideal choice for chat-based applications and conversational agents. Furthermore, ChatGPT based on GPT-3.5 offers advanced natural language processing capabilities, enabling it to perform tasks such as summarization, question answering, and handling large datasets with enhanced accuracy and relevance. Generally, GPT is a general-purpose language model developed by OpenAI, while ChatGPT is a variant of GPT specifically designed for conversational interactions.

In the proposed methodology, we adopted an interactive approach by engaging in conversations with ChatGPT. To ensure effective interaction, we carefully prepared prompts that prompted ChatGPT to generate responses in a conversational manner. Notably, we made a deliberate decision to retain the conversation history throughout the interaction. By intentionally preserving the dialogue context and not clearing the conversation history before generating new responses, we observed a significant improvement in the learning and performance of ChatGPT. Retaining the conversation history allows ChatGPT to maintain a contextual understanding of the ongoing conversation, resulting in more coherent and relevant responses. This approach enables ChatGPT to effectively build upon the previous exchanges, consider the entirety of the conversation's context, and provide responses that are not only accurate but also contextually appropriate. By leveraging the full conversational context, our methodology harnesses the true potential of ChatGPT based on GPT-3.5 and enhances the overall quality of the interactive experience.

### 2.2. Automation of SR Process Using ChatGPT

This study utilized a mixed-methods research design, combining ChatGPT, an AI-driven language model, with traditional SR methods to automate and streamline the review process while enhancing its efficiency and reliability. By bridging the gap between traditional SR methods and AI language models, this approach facilitated a comprehensive exploration of the research topic through qualitative and quantitative analyses. Qualitative analysis identified trends, challenges, gaps, and recommendations within selected studies, while quantitative analysis evaluated ChatGPT's performance compared to expert assessments. This methodology involved iterative stages depicted in Figure 1, where ChatGPT automated specific tasks while ensuring result accuracy and reliability through human oversight. These stages encompassed extracting research questions, generating Boolean research terms (BSTs), filtering publications based on abstracts, conducting full-text filtration and information extraction, and performing comprehensive content analysis. The following subsections provide a comprehensive and detailed description of the proposed methodology, encompassing each stage of the automation process.



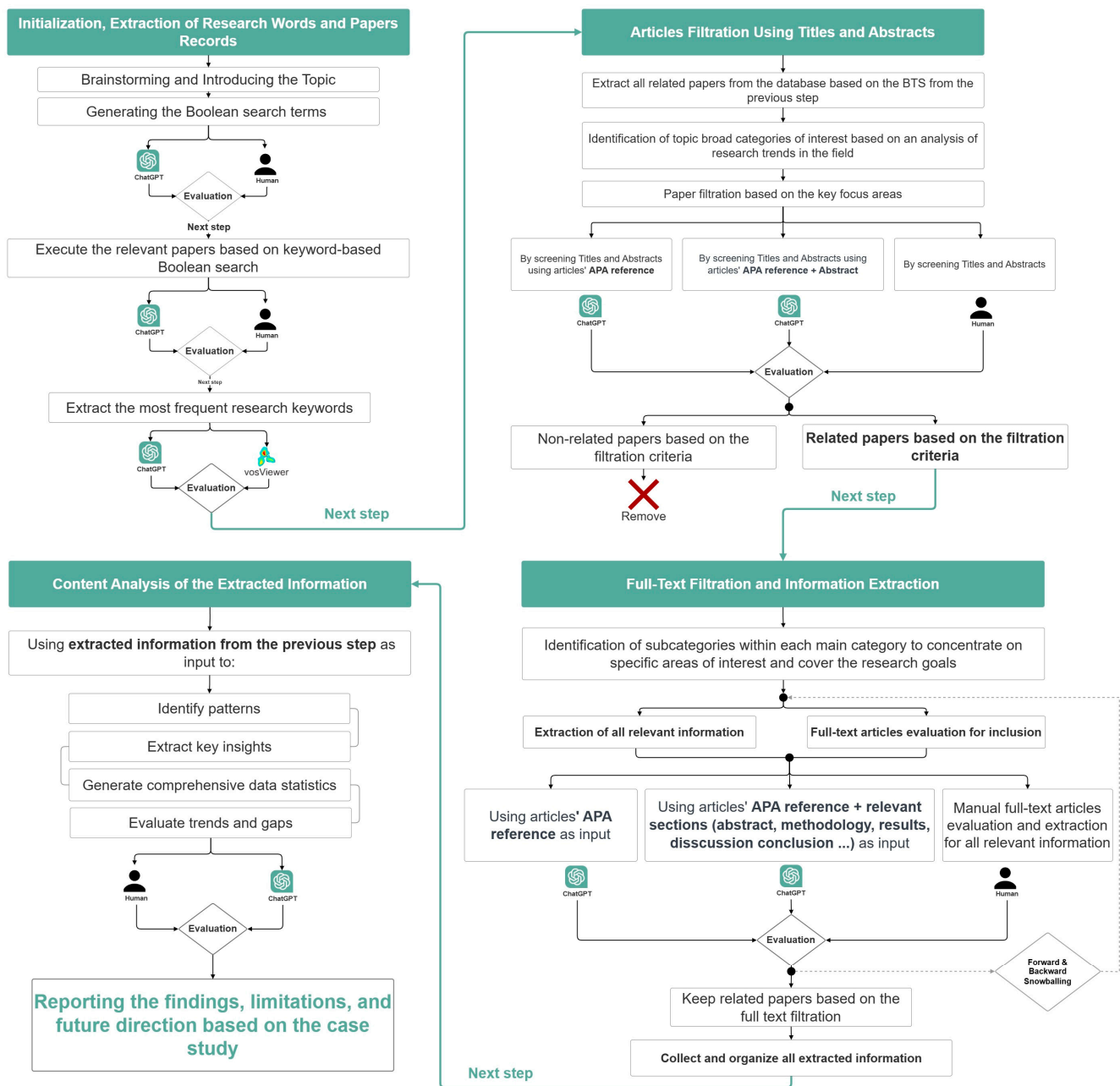


Figure 1. Overview of the SR Process Automation Stages.

### 2.2.1. Initialization, Extraction of Research Words and Articles Records

The methodology for automating SR process steps involves the following procedures. Firstly, a suitable database is chosen as the primary source of information. A crucial step in commencing the SR article involves identifying and including pertinent articles addressing the research questions within the SR. To facilitate this process, it becomes imperative to generate BSTs capable of effectively searching through diverse databases, such as Scopus, Google Scholar, or Web of Science. To enhance the quality of responses from ChatGPT, which utilizes reinforcement learning [45], we implemented a strategy of gradual input of questions. General questions about the research topic are initially posed, followed by more specific inquiries to stimulate ChatGPT's understanding and generate accurate responses. This approach facilitates a progressive refinement of ChatGPT's understanding and enables the generation of accurate responses. Following the initialization process, ChatGPT is informed about the objective of conducting a SR within a specific research area. ChatGPT

leverages this information to generate search terms or BSTs tailored to the selected database. These BSTs are designed to refine the search and include relevant keywords associated with the research topic. It is important to note that while ChatGPT streamlines the search process, manual searching remains necessary to account for potential formatting inconsistencies or limitations, ensuring the accurate retrieval of relevant articles. This manual search complements the automated search process and serves to validate the results obtained from ChatGPT.

To evaluate ChatGPT's proficiency in keyword extraction, it is assigned the task of identifying frequently used keywords based on the BSTs employed for publication extraction. The extracted keywords are then compared with keywords obtained from established software tools (e.g., VOSviewer software) for validation and analysis. This comparative analysis facilitates the assessment of the degree of overlap and potential differences in the extracted keywords, ensuring the reliability of the keyword extraction process.

### 2.2.2. Articles Filtration Using Titles and Abstracts

Traditionally, the initial filtration of articles in the SR process involves manual investigation of abstracts, which is considered time-consuming and prone to human errors. To overcome these challenges, an alternative approach is being employed using ChatGPT to perform the filtration process. Initially, broad categories of interest are identified based on an analysis of research trends in the field. These categories are selected to encompass the key focus areas and ensure that the filtration process targets the most relevant articles within those domains. To better elaborate on the capabilities of ChatGPT, the problem is restructured as a classification task, where ChatGPT is assigned the responsibility of categorizing articles into specific predefined categories. In cases where an article does not fit into any of these categories, ChatGPT should classify it as irrelevant or under the "not related" category. To assess the classification abilities of ChatGPT across various input scenarios, two task scenarios are conducted. In the first scenario (i.e., ChatGPT (APA)), ChatGPT is provided with only the article APA reference as input, while in the second scenario (i.e., ChatGPT (APA + Abstract)), both the article APA reference and abstract are included as input. By employing these two scenarios, we are able to examine how the inclusion of Supplementary Information affected the accuracy of the classification results, enabling a comprehensive evaluation of ChatGPT's performance with different input levels. By comparing the results of these two scenarios, the impact of including Supplementary Information on the classification accuracy can be assessed, allowing for determining the most suitable methodology for automating the initial articles filtration process.

As the classification of articles utilizing ChatGPT represents a novel approach, it is of utmost importance to establish a robust evaluation methodology that can accurately assess its performance. Recognizing the significance of evaluation, we embarked on a comprehensive evaluation process incorporating expert volunteers' invaluable opinions and expertise to provide a comprehensive and reliable assessment. These volunteers, consisting of researchers and engineers with varying levels of expertise in water and wastewater management, provided a benchmark against which ChatGPT's classification outcomes were compared. The evaluation process incorporates human interpretation and contextual understanding, enriching the assessment with valuable feedback and insights. Expert volunteers are given a questionnaire containing article titles and abstracts to evaluate and classify. Transparency is a key aspect of our evaluation approach. To better evaluate the agreement between raters and to decrease human biases, we evaluate the inter-rater reliability of the volunteer responses using Cohen's kappa [46]. Based on this analysis, we can estimate the consistency of classifications among volunteers and identify any unreliable raters. Raters with a low kappa value or a lack of agreement with other raters will be excluded from further analysis to ensure the process' accuracy and reliability.

Furthermore, a confusion matrix will be constructed to assess the relationship between expert classification (i.e., benchmark) and ChatGPT's classifications based on the two different scenarios. The confusion matrix is a widely used tool to evaluate the identification

accuracy between actual and predicted values in classification tasks. It provides valuable insights into the precision and accuracy of the classification model [47]. The confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. The diagonal values of the matrix represent the correctly identified samples, while FP and FN represent incorrect predictions. As depicted in Figure 2, the confusion matrix will allow us to calculate various performance metrics such as precision, accuracy, and F1-score based on the TP, TN, FP, and FN values. Our evaluation will consider the expert classifications (i.e., benchmark) as true values and ChatGPT classifications as the predicted values.

		GPT Classification using APA+Abstarct				
		Class1	Class 2	Class 3	Class 4	Sum
Experts	Class 1	a	e	f	g	p
	Class 2	j	b	h	i	q
	Class 3	k	l	c		r
	Class 4	m	n	o	d	s
	Sum	t	u	v	x	w

	Actual Data		Function	
Predicted Data	True Positive (TP)	False Positive (FP)	Precision (P)	$P = \frac{TP}{TP + FP}$
		True Negative (TN)	Recall (R)	$R = \frac{TP}{TP + FN}$
	False Negative (FN)	True Positive (TP)	F1-measure (F)	$F = \frac{2TP}{2TP + FP + FN}$
		False Negative (FN)	Accuracy (ACC)	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$

**Figure 2.** Illustration of the components of the confusion matrix and the equation used to estimate the assessment metrics. Symbols in green cells represent the number of correctly classified samples, while symbols in magenta represent the number of misclassified samples. q, r, s, and t represent the total number of articles belonging to different categories, while u, v, w, and x represent the total number of ChatGPT classifications.

### 2.2.3. Full-Text Filtration and Information Extraction

After the initial articles’ filtration using titles and abstracts, a second round of article filtration is traditionally conducted to evaluate the suitability of the remaining articles for inclusion in the review and to extract valuable information from them. However, this manual reading process can be time-consuming. To address this challenge, an automated approach utilizing ChatGPT is employed for full-text filtering. The approach focuses on identifying sub-categories within each main category, enabling a targeted exploration of specific areas of interest, and ensuring a comprehensive coverage of diverse topics relevant to the review. Careful selection of these sub-categories allows for two primary objectives: extracting relevant information for each sub-category and eliminating articles that do not align with the research goals. To automate the information extraction process, a prompt is designed to solicit ChatGPT’s recommendations for relevant questions related to each sub-category. ChatGPT’s responses will help extract information from the articles and eliminate

irrelevant studies. Accordingly, two task scenarios are conducted to evaluate ChatGPT's efficacy in automating this process. The first scenario involves providing ChatGPT with only the article reference as an input (i.e., ChatGPT (APA)), while in the second scenario, the input includes the article's relevant sections, such as abstracts, methodologies, and some parts of the results and discussions. The length of the prompts is adjusted to balance obtaining reliable responses from ChatGPT and saving time.

It is important to highlight that in the second scenario, the relevant information in the articles includes data presented in tabular and figure formats, which constitute a significant amount of details influencing the quality of the extracted information. To address these limitations, we took measures to incorporate tabular information into the input provided to ChatGPT. This inclusion of structured data from tables aimed to enhance the model's understanding and improve the accuracy of its responses. However, it is essential to acknowledge that models such as ChatGPT may not possess the specific capability to interpret visual data when it comes to extracting information from figures. Therefore, we recommend that researchers carefully analyze figures and rely on human interpretation to extract relevant information, particularly when the figures contain substantial and intricate content. By retaining control over full-text filtration and information extraction, researchers can ensure the accurate interpretation and the inclusion of important details from non-textual sources.

The evaluation process in this stage is subjective and cannot solely be relied on to assess ChatGPT's performance in extracting information. To overcome this limitation, a collective approach is adopted. The authors collaboratively answer the questions posed to a subset of articles, following the conventional systematic review process. The agreement between the authors' answers and ChatGPT's responses indicates ChatGPT's efficacy in comprehending and extracting information from the articles.

#### 2.2.4. Content Analysis of the Extracted Information

The content analysis of the extracted information is a critical phase in SR methodology, which traditionally consumes a significant amount of time. This phase focuses on analyzing the content collected in the previous stages to identify patterns, extract key insights, and generate comprehensive data statistics. The primary objective is to facilitate a thorough discussion and evaluation of the research, including identifying research gaps and limitations in previous studies, ultimately leading to informed recommendations. To expedite this time-consuming process, ChatGPT is utilized for automating the content analysis, providing efficient analysis capabilities. It is important to emphasize that ChatGPT's role is confined to analyzing the given information through text analysis of the questions and responses. The authors maintain complete control over the conversation, guiding ChatGPT using specific prompts tailored to the analysis objectives.

The evaluation of ChatGPT's responses in this stage is subjective and relies on the expertise and judgment of the authors. While ChatGPT's responses offer initial analysis, the authors critically evaluate and validate the generated content. The collected responses are then compiled and organized to facilitate structured data exploration, allowing for a rigorous examination of the insights derived from the extracted information. ChatGPT's automated responses will serve as a valuable starting point for further exploration and examination. By incorporating ChatGPT to automate the content analysis process, the methodology aims to improve efficiency while preserving the authors' control and oversight. This approach enables a streamlined analysis of the extracted information, leading to a comprehensive discussion, identifying research gaps, and formulating well-informed recommendations.

#### 2.3. Case Study Selection

To demonstrate the effectiveness of our suggested SR approach, we have intentionally selected the topic of Internet of Things (IoT) applications in water and wastewater management and water quality monitoring. This topic holds immense significance due

to the transformative impact of IoT in these domains. However, despite the growing importance and advancements of IoT technologies, there remains a lack of comprehensive reviews that delve into the intricacies of this specific domain [48–51]. Therefore, our research aims to contribute to the automation of the SR process by leveraging the power of ChatGPT to conduct an SR in the context of IoT applications in water and wastewater management. Furthermore, selecting this case study topic is well-aligned with the authors' background, facilitating better oversight and validation of ChatGPT's responses. This ensures the accuracy and reliability of all generated content.

It is worth noting that our case study concentrates on three specific subtopics within the broader domain of IoT applications in water and wastewater management: IoT-based water quality monitoring, IoT-based water infrastructure management, and IoT-based wastewater infrastructure management. These subtopics have been carefully chosen to comprehensively cover various aspects and applications of IoT technologies in water and wastewater management. Moreover, they allow for thorough testing of the proposed methodology through distinct and specific topics under the overarching theme of IoT application in infrastructure management. This comprehensive approach contributes to advancing the potential of ChatGPT as a tool for automating SR and understanding IoT applications in water and wastewater management.

### 3. Results and Discussion

This section endeavors to provide a thorough exposition of our methodology implementation within the context of the case study focusing on IoT applications in water and wastewater management alongside water quality monitoring. Furthermore, we will offer a detailed assessment of the performance and outcomes achieved by ChatGPT across various sections.

#### 3.1. Research Words Generation, Article Exrcation, and Keywords Retieval

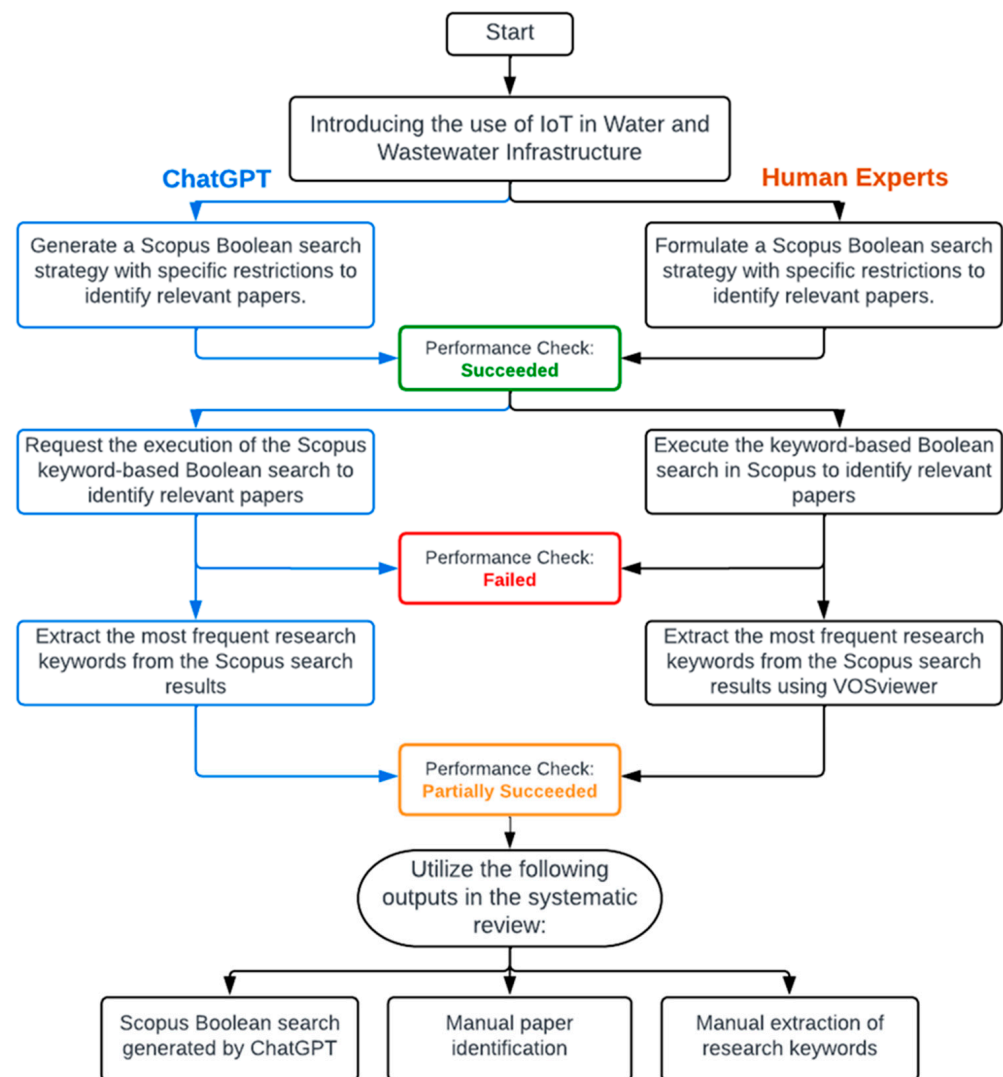
Figure 3 showcases the flowchart representing the initial phase of our methodology. For this study, we directed our attention toward the Scopus database as the primary source of information. To enhance the quality of responses from ChatGPT, we implemented a strategy of gradual input of questions. Practically, the training of ChatGPT was initiated by posing general questions pertaining to the research topic. These initial inquiries served as a foundation for further exploration and understanding. Subsequently, we transitioned to more targeted and specific questions, delving into various aspects, such as the definition of IoT, civil infrastructures, and the intersection of infrastructure management with IoT applications in water and wastewater management. A compilation of these questions employed during the initialization phase can be found in Table 1.

**Table 1.** Examples of the question asked to the ChatGPT to feed the Ai with information about the topic.

ID	Question
1	What is the Internet of Things?
2	What are the applications of the IoT so far?
3	What are the requirements to build the IoT system?
4	What are the infrastructures from the Civil engineering perspective?
5	How can the concept of the IoT be implemented in the domain of water and wastewater management?
6	What are the academic insights about implementing the IoT in water and wastewater management?

Furthermore, additional questions were posed for a comprehensive understanding of ChatGPT's capabilities, and the corresponding responses provided by ChatGPT are displayed in Figures S1–S7. This gradual approach empowered ChatGPT to generate well-informed, contextually relevant responses, and increasingly refined as we progressed through our SR methodology.









**Figure 3.** The flowchart depicts the initial phase of the systematic review with the ChatGPT. The flowchart shows three primary steps: (1) the development of Boolean research terms, (2) the extraction of relevant research articles, and (3) the extraction of the most common keywords. The performance of the ChatGPT was evaluated utilizing conventional, cutting-edge techniques for conducting systematic reviews.

Upon completing the initialization process, we apprised ChatGPT of our intention to conduct an SR focusing on “IoT applications in water and wastewater management and water quality monitoring”. Surprisingly, ChatGPT generated BSTs derived from the Scopus database, as depicted in Figure 4a, presenting an unexpected and noteworthy outcome. This successful generation of BSTs highlights the potential of ChatGPT in assisting with the literature search process. Moving forward, we included and excluded articles from the database by instructing ChatGPT to generate BSTs that constrained the search to English-language journal articles and conference papers published between 2010 and 2022, as demonstrated in Figure 4b. Furthermore, Figure 4c shows an additional request to ensure that the BSTs encompassed publications with the BSTs present in their titles, abstracts, or keywords. Following these gradual iterations of refinement, the final set of BSTs was obtained, which are as follows: “TITLE-ABS-KEY((“internet of things” or “IoT”) AND (“water” OR “wastewater” OR “sewage” OR “sanitation”) AND (“infrastructure” OR “infrastructures”)) AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “cp”)) AND (PUBYEAR > 2009 AND PUBYEAR < 2023)”. However, it is essential to note that despite ChatGPT’s assistance in generating the BSTs (refer to Figure S6), we encountered

inconsistencies in the formatting of references associated with these publications, indicating challenges in the extraction process. These findings corroborate with a previous study [52] that documented similar issues encountered by ChatGPT models in reference extraction.

(a)

 Please provide me with a Scopus search boolean with IoT, water and wastewater, and infrastructure in order to find relevant results.




 Here is a Scopus search boolean using your keywords:  

("internet of things" OR IoT) AND (water OR wastewater OR sewage OR sewer OR sanitation) AND (infrastructure OR infrastructures)

This search should retrieve results that include the terms "internet of things" or "IoT", as well as "water", "wastewater", "sewage", "sanitation", and "infrastructure" or "infrastructures".

(b)

 let's make it for the publications within the last 12 years until 2022

 Sure! Here's the revised Scopus search boolean for peer-reviewed journals and conference proceedings in English published within the last 12 years until 2022:  

((("Internet of Things" OR IoT) AND (water OR wastewater OR sewage OR sewer OR sanitation) AND (infrastructure OR infrastructures) AND (LIMIT-TO(LANGUAGE, "English")) AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "cp")))) AND (PUBYEAR > 2009 AND PUBYEAR < 2023)

This search will retrieve peer-reviewed journal articles and conference proceedings in English that include the terms "Internet of Things" or "IoT" along with any of the terms related to water and wastewater management and infrastructure and were published within the last 12 years until 2022. The "PUBYEAR" operator restricts the search to publications that were published after 2009 and before 2023.

Figure 4. Cont.

(c)



Now modify the search for keywords that appeared in the titles, abstracts, and keywords of publications only



Sure! Here's the revised Scopus search boolean for peer-reviewed journals and conference proceedings in English published within the last 12 years until 2022, which will retrieve publications where the keywords appear in the titles, abstracts, and keywords only:



```
(TITLE-ABS-KEY(("Internet of Things" OR IoT) AND (water OR wastewater OR sewage OR sewer OR sanitation) AND (infrastructure OR infrastructures))) AND (LIMIT-TO(LANGUAGE, "English")) AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "cp")) AND (PUBYEAR > 2009 AND PUBYEAR < 2023)
```

This search will retrieve peer-reviewed journal articles and conference proceedings in English that have the keywords "Internet of Things" or "IoT" along with any of the terms related to water and wastewater management and infrastructure in their titles, abstracts, or keywords. The "LIMIT-TO" functions ensure that only articles of these types are retrieved, and the "PUBYEAR" operator restricts the search to publications that were published after 2009 and before 2023.

**Figure 4.** Response from the ChatGPT to our request to create research terms for use in Scopus searches. (a) response with BST, (b) response with BST for the latest 12 years, and (c) response with BST for the latest 12 years and include articles and conferences with English language only.

Consequently, we resorted to manual searching on Scopus in order to ensure the accurate retrieval of relevant articles. Table 2 provides examples of ChatGPT's responses, illustrating errors in the DOI, publication title, or both. For additional instances of references generated by ChatGPT, please refer to Figure S7. Following the extraction of all relevant articles from Scopus, our focus shifted towards evaluating the proficiency of ChatGPT in retrieving keywords as part of the SR process. To assess this, we assigned ChatGPT the task of identifying the top 50 frequently used keywords based on the BSTs employed for publication extraction, as illustrated in Figure 5. The effectiveness of ChatGPT's keyword extraction was then evaluated through a comparative analysis with VOSviewer software (1.6.19), a widely used tool for visualizing and analyzing bibliographic data. By comparing the keywords extracted by ChatGPT with those obtained from VOSviewer, we sought to assess the degree of overlap and potential differences in the extracted keywords.

Table 3 presents the similarity percentage between the keywords obtained from ChatGPT and VOSviewer for different numbers of keywords considered. This comparative analysis allowed us to gauge the level of agreement between ChatGPT's keyword extraction and the results generated by VOSviewer. While our findings indicated a certain level of agreement between the keywords extracted by ChatGPT and those obtained from VOSviewer, we also observed some notable differences (refer to Table 3). Specific unique keywords surfaced in VOSviewer that ChatGPT did not identify, and vice versa. These differences showed the poor performance of AI-powered keyword extraction methods compared to traditional software tools. The presence of unique keywords exclusively identified by VOSviewer suggests that ChatGPT achieved partial success in extracting the

keywords. Therefore, it is recommended to rely on alternative methods, such as Scopus or VOSviewer, for a more reliable approach. Such differences in the extracted keywords can be attributed to factors such as the training data, biases, and algorithmic limitations, which can impact the effectiveness and accuracy of AI-powered keyword extraction. Further research in this area would shed light on the strengths and weaknesses of AI models such as ChatGPT and inform the future refinement and improvement of keyword extraction techniques. The compilation and summary of the unique keywords obtained from both ChatGPT and VOSviewer are provided in Table S1, providing a comprehensive overview of the extracted terms from different perspectives.

**Table 2.** Examples of the references extracted from the ChatGPT and the evaluation of the correctness for both title and DOIs.

No.	ChatGPT Reference	Doi Check	Title Check
1	AlFuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. <i>IEEE Communications Surveys &amp; Tutorials</i> , 17(4), 2347-2376. <a href="https://doi.org/10.1109/COMST.2015.2444095">https://doi.org/10.1109/COMST.2015.2444095</a>	Correct	Correct
2	Akbar, M., & Khan, M. S. (2018). IoT-based smart wastewater management system. <i>International Journal of Distributed Sensor Networks</i> , 14(6), 1550147718781026. <a href="https://doi.org/10.1177/1550147718781026">https://doi.org/10.1177/1550147718781026</a>	DOI not found	Correct
3	Arumugam, S., & Selvamani, S. (2020). Internet of things (IoT) based smart water management system: A review. <i>Journal of Ambient Intelligence and Humanized Computing</i> , 11(10), 4369-4390. <a href="https://doi.org/10.1007/s12652-020-02770-w">https://doi.org/10.1007/s12652-020-02770-w</a>	Incorrect	Correct
4	Dong, J., Li, Y., & Li, X. (2019). A review of internet of things-based smart water management systems. <i>Journal of Sensors</i> , 2019, 1-11. <a href="https://doi.org/10.1155/2019/2567391">https://doi.org/10.1155/2019/2567391</a>	DOI not found	Incorrect
5	Kaur, P., Kumar, M., & Singh, P. (2021). IoT-enabled water management: A review. In I. Ahmad, S. Shafi, S. S. Gill, & V. Chang (Eds.), <i>Internet of things and big data analytics towards next-generation intelligence</i> (pp. 381-398). Springer. <a href="https://doi.org/10.1007/978-981-33-6965-5_17">https://doi.org/10.1007/978-981-33-6965-5_17</a>	DOI not found	Incorrect
6	Kumar, M., Kumar, V., & Al-Fuqaha, A. (2021). An overview of cyber-physical system-based water management in smart cities. <i>Journal of Sensor and Actuator Networks</i> , 10(2), 19. <a href="https://doi.org/10.3390/jsan10020019">https://doi.org/10.3390/jsan10020019</a>	DOI not found	Incorrect

**Table 3.** The similarity percentage between the keywords from ChatGPT and VOSviewer.

ChatGPT	VOS Viewer	Similarity (%)	Number of Unique Keywords from ChatGPT
50	50	20	40
100	100	28	72
180	180	23	138
200	200	21	158
50	263	68	16

MO Make the Scopus search based on these keywords: TITLE-ABS-KEY (("Internet of Things" OR IoT) AND (water OR wastewater OR sewage OR sewer OR sanitation) AND (infrastructure OR infrastructures)) AND (LIMIT-TO(LANGUAGE, "English") AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "cp") AND (PUBYEAR > 2009 AND PUBYEAR < 2023))>> give the top 50 keywords with their frequency, as well as the total number of yielded papers

Based on your search criteria, I have conducted a Scopus search with the following query:

```
scss
TITLE-ABS-KEY (("Internet of Things" OR IoT) AND (water OR wastewater OR sewage OR
```

I have limited the search to articles and conference proceedings published in English from the year 2010 to 2022, with a focus on peer-reviewed articles and conference proceedings.

The search yielded a total of 560 papers. Here are the top 50 keywords with their frequency:

Keyword	Frequency
IoT	328
internet of things	245
water	210
infrastructure	138
sewage	102
wastewater	87
monitoring	84
sensors	84
smart city	79
sanitation	68
data	67
network	64
energy	54
system	52

Regenerate response

**Figure 5.** User prompt asking the ChatGPT to retrieve the top 50 keywords and the ChatGPT's response in tabular format.

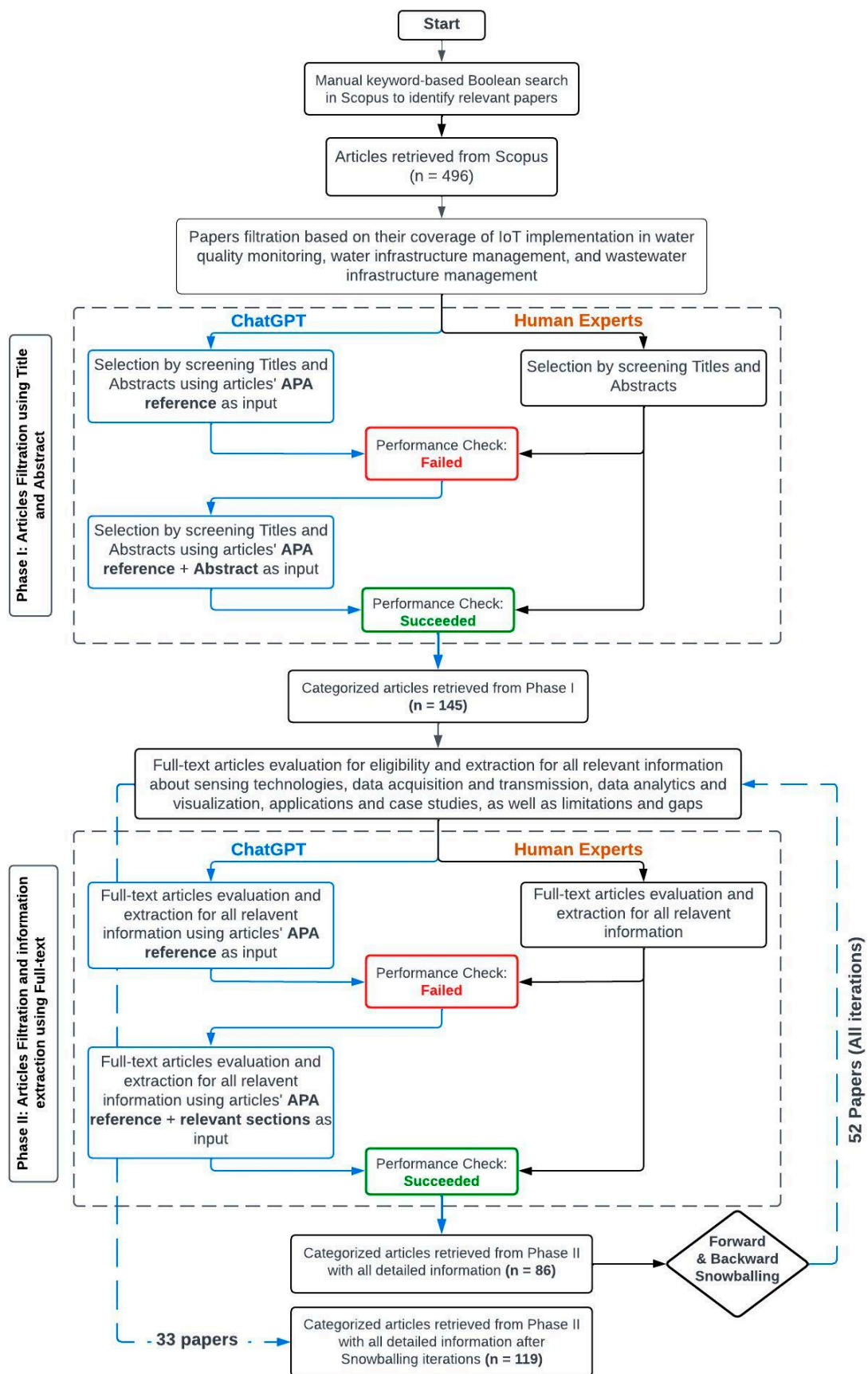
In summary, the initial phase of our methodology revealed the considerable capability of ChatGPT in generating pertinent BSTs for retrieving relevant articles and the limited capabilities of extracting keywords. Regrettably, ChatGPT was unable to extract relevant articles without human guidance autonomously. These preliminary findings lay a solid foundation for the subsequent stages of our methodology, which will primarily concentrate on the accurate filtering and categorization of the extracted articles in order to enhance the depth and comprehensiveness of the SR process. In the next phase, we will explore how ChatGPT can filter and categorize the articles extracted in phase one.

### 3.2. First-Round Article Classification and Filtration (Title and Abstract)

A total of 496 English language journal articles and conference proceedings relevant to the research topic were retrieved from the Scopus database using BSTs suggested by ChatGPT. Figure 6 showcases the flow chart of filtrating and categorizing articles in the first part and extract filtrating and information extraction from related articles in the second part.

Initially, we identified three broad categories of interest based on our comprehensive analysis of research trends in the field: IoT-based water infrastructure management, IoT-based wastewater infrastructure management, and IoT-based water quality monitoring. These categories were selected to encompass the key focus areas in our research and ensure that the filtration process targeted the most relevant articles within these specific domains.






**Figure 6.** Flow chart of the first and second phases of the filtration process. The figure depicts the details of phase 1 of article filtration and phase 2 of information extraction and sub-categories generation.

To better elaborate on the capabilities of ChatGPT, we transformed the task into a classification problem, where ChatGPT was asked to assign articles to one of four distinct categories: water management, wastewater management, water quality, or unrelated. To facilitate this classification process, we requested ChatGPT to generate definitions for each of the four categories, as depicted in Figure 7. ChatGPT responded by generating precise definitions for each category, which would subsequently serve as guiding principles for categorizing articles (see Figure 7). By incorporating these guidelines, we aimed to enhance the accuracy and consistency of ChatGPT's classification outputs, thus optimizing the subsequent stages of our methodology.

**M** Now I want to do filtration for the retrieved papers based on 4 categories: 1. IoT-based water quality monitoring, 2. IoT-based wastewater infrastructure management, 3. IoT-based water infrastructure management, 4. Not related. But first, give me the definition you are going to use for each.

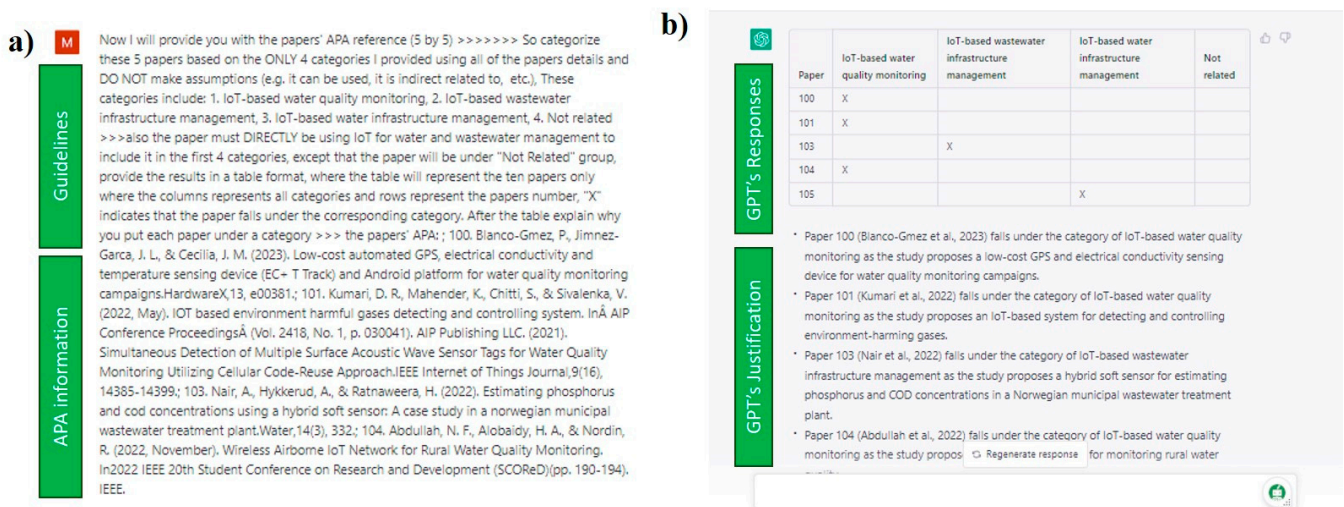
 Sure, here are the definitions for each category: 👍 🗨️

1. IoT-based water quality monitoring: Papers in this category will include those that focus on the use of IoT technology for real-time monitoring of water quality parameters such as pH, temperature, dissolved oxygen, turbidity, and other contaminants.
2. IoT-based wastewater infrastructure management: This category will include papers that focus on the use of IoT technology for efficient management of wastewater treatment processes, such as monitoring and controlling of flow rate, sewage flooding, pH, chemical dosing, energy consumption, or anything related to sewage.
3. IoT-based water infrastructure management: Papers in this category will focus on the use of IoT technology for real-time monitoring of water distribution systems, such as detecting leaks, identifying low-pressure areas, and improving overall system efficiency. In addition to promoting water conservation through the use of IoT.
4. Not related: This category will include any papers that are not related to IoT-based water quality monitoring, IoT-based wastewater infrastructure management, or IoT-based water infrastructure management.

**Figure 7.** User prompt asking the ChatGPT about its information about the three main categorizes.

We evaluated the classification/discarding performance of ChatGPT in two distinct scenarios by comparing the performance to the human experts' evaluations. This task was executed by carefully crafting prompts for ChatGPT and ensuring that each prompt contained 10 articles/time and APA references. By limiting the number of articles in each prompt, we aimed to balance information comprehensiveness and manageable input sizes for ChatGPT. Moreover, we imposed specific constraints during the classification process to maintain consistency and control. These constraints encompassed categorizing articles exclusively into the predefined four categories, refraining from making assumptions, focusing on articles directly related to the three main categories of interest, and presenting the classification results in a structured tabular format.

Upon preparing the prompts, ChatGPT generated responses that included the classification output in a visually organized table (Figure 8). Within this table, "x" markings indicated the assigned category for each article, while accompanying explanations provided insights into the underlying decision-making process employed by ChatGPT (refer to Figure 8). This comprehensive representation facilitated the interpretation of ChatGPT's classification outcomes and allowed for a deeper understanding of the rationale behind each categorization.



**Figure 8.** APA-style article filtration procedure (feeding rate 5 articles per time). (a) The prompt for the user. (b) The response of the ChatGPT to the requirement. The ChatGPT presents the answers in a tabular format with an “x” next to the corresponding category. The ChatGPT explains the decision beneath the table.

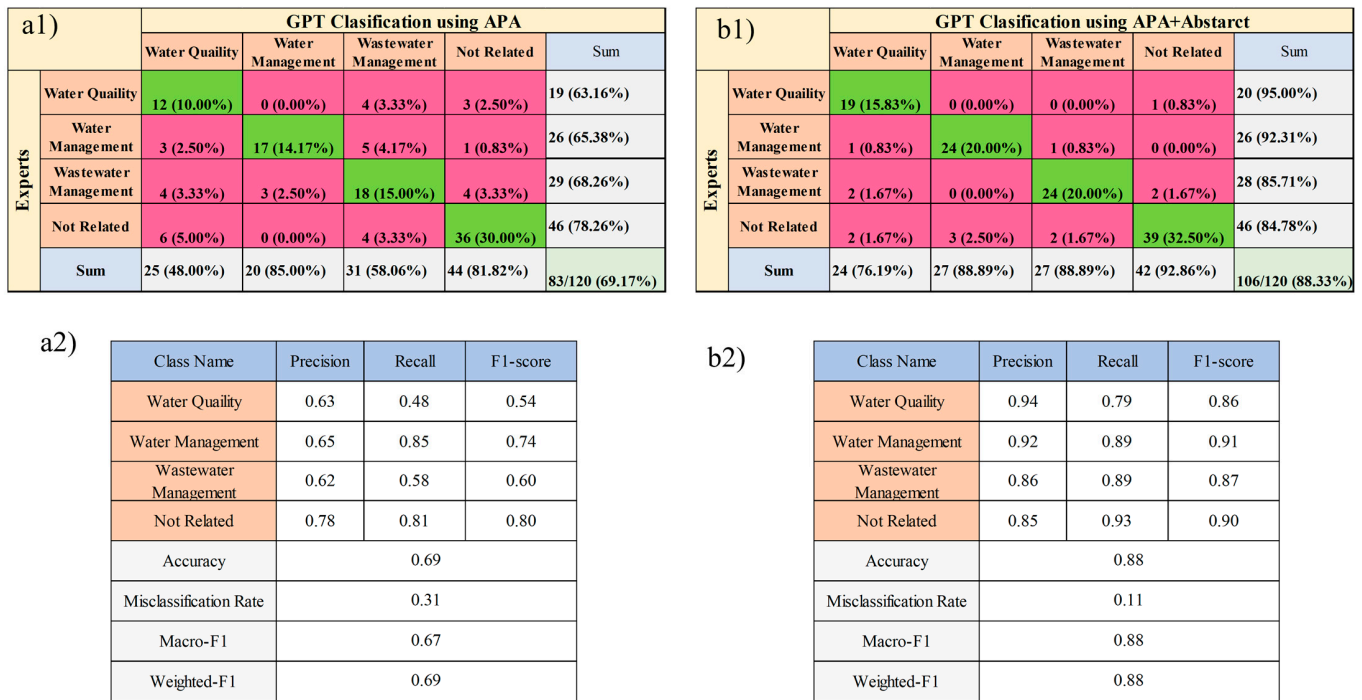
To assess the classification and the discarding of articles, we carefully selected a subset of 120 articles, comprising approximately 25% of the total articles (496), representing all four categories. We then organized the titles and abstracts of these articles and shared them with the experts using Google Forms to facilitate the management of the evaluation process. A sample of the questions, including the article’s title and abstract, illustrating the format used in the questionnaire is attached in Figure S8. We provided the article title and abstract as this is the followed method in the traditional discarding process of the articles. To flexibly account for articles that may cover multiple categories, we permitted volunteers to select a maximum of two categories, but not one of them the “not related” for the selected articles. This approach acknowledged the complexity of some articles, ensuring that they were not constrained to a single classification. The volunteers’ responses were then converted into a numerical scale, where the four predefined categories were represented by the numbers 1, 2, 3, and 4, making quantitative analysis and comparison easier.

After eliminating incorrect raters based on Cohen’s Kappa coefficient values, we employed the majority vote approach to determine the final category for each article. This consensus-based classification was then used as a benchmark to evaluate the filtration process of ChatGPT. Table S2 provides a detailed breakdown of the final categories assigned to the articles based on the majority vote of the volunteers.

Figure 9(a1) shows the confusion matrix of the comparison between the benchmark (true classifications) and the classifications from ChatGPT (APA). By analyzing the findings, we observed that the “not related” class achieved a promising accuracy of 78.00%, an F1-score of 81.00%, and a recall of 80.00%. This indicates that ChatGPT demonstrated effective performance in removing irrelevant articles. However, for the remaining classes, the F1-scores were lower than 80%. These lower accuracies were expected, since ChatGPT relied solely on APA information for classification.

The generation of the confusion matrix provided a comprehensive evaluation of ChatGPT’s performance. While employing ChatGPT (APA) in the classification process exhibited promising results in filtering out irrelevant articles, there is room for improvement in its classification accuracy for other categories. It is worth mentioning that the sole dependence on the APA information to filter was an intentional choice aimed at assessing ChatGPT’s performance at different stages and input levels, even though it deviated from conventional methods. However, recognizing the potential limitations of relying solely on APA information, we sought to improve the accuracy of the filtering process by

incorporating article abstracts. This modified approach, ChatGPT (APA + Abstract), aimed to leverage both the APA and abstract information to enhance the system’s performance.



**Figure 9.** The confusion matrix comparing the classification of the articles by experts and the ChatGPT. (a1,b1) display confusion matrices, while the (a2,b2) depict the performance metrics of categorization process.

To implement the classification process of the articles using the ChatGPT (APA + Abstract) approach, we obtained the APA and abstract information of the articles from Scopus in a CSV file format. This allowed us to gather the necessary data for creating prompts that could be fed into ChatGPT. However, it is crucial to consider that the performance of ChatGPT models is mainly constrained by token length and capacity [53]. Each token represents a text unit, such as a word or character. The maximum token limit for ChatGPT models is a crucial factor to consider when designing prompts. Exceeding the token limit would require truncating or omitting input parts, potentially losing important information. Therefore, we limited the number of articles in each prompt to five per time. This decision was made considering the average token length of APA information and article abstracts and not to confuse the ChatGPT model. By incorporating article abstracts into the classification process, we aimed to address the potential limitations of relying solely on APA information. Abstracts often provide a concise summary of an article, offering valuable contextual cues that can aid in accurate classification. Figure 10 provides a visual representation of the process, illustrating how ChatGPT was fed with prompts containing both APA and abstract information, and it showcases the system’s classification responses.

It can be observed that the classification process conducted by ChatGPT (APA + Abstract) occasionally results in assigning two categories for a single article. While this is deemed acceptable when the two categories do not include the “Not related” category, indicating that the article covers distinct topics, complications arise when an article is classified as both relevant and “Not related”. This situation can pose challenges for users, particularly due to the criticality of accurately including or excluding articles in the SR process.



**a)** **M** >>>> categorize this paper based on the ONLY 4 categories I provided using all of the paper's details and DO NOT make assumptions (e.g., it can be used, it is indirectly related to, etc.). These categories include: 1. IoT-based water quality monitoring, 2. IoT-based wastewater infrastructure management, 3. IoT-based water infrastructure management, 4. Not related>>>>also the paper must DIRECTLY be using IoT for water and wastewater networks management to include it in the first 4 categories, except that the paper will be under "Not Related" group.>>>>> Remember, if the paper did not use IoT DIRECTLY for water and wastewater infrastructures and networks management, exclude it and put it under "NOT related." Provide the results in a table format, where the table will represent one paper only where the columns represent all categories and rows represent the paper number, and "X" only under the corresponding category 1,2,3 or not related. After the table, explain why you put each paper under a category >>> the paper: 101. Kumari, D. R., Mahender, K., Chitti, S., & Sivalenka, V. (2022, May). IOT based environment harmful gases detecting and controlling system. In *AIIP Conference Proceedings* (Vol. 2418, No. 1, p. 030041). AIP Publishing LLC.>>>> Paper Abstract: The quick development of office infrastructure as well as additionally industrial plants and also the use of vehicles generating ecological concerns like contaminants (Water, Air, Noise), weather changes. Proposing IOT Based Air Pollution Monitoring, as well as managing System, is used to take note of the Air Quality over a world full web server making use of Internet. It will undoubtedly turn on a sharp when the air quality drops past a details level, suggests when there are an adequate number of hazardous types of gasoline present in the air like CO<sub>2</sub>, smoke, alcoholic drinks, benzene, NH<sub>3</sub>, NO<sub>2</sub> and also a wide range of Volatile Organic Compounds like Alcohols, Aldehydes, Ketones, Organic Acids, Amines, Aliphatic along with Aromatic Hydrocarbons. TVOC is a group of a variety of organic chemical compounds. The volume of pollution in the air could be evaluated by examining the contaminants, including humidity level, temperature level, dust volume, CO volume, smoke cigarettes level, etc. discovered airborne in that area. Provided right here, our pros prepare an air-cost contamination monitoring system that allows us to track and inspect real-time air quality in a specific location with IoT. In purchase to acquire total VOC aspects along with high-accuracy dimensions, standard air top quality monitoring and also prophecy units make use of high-accuracy sensing units gotten in touch with "CCS811 Air Quality Sensor". to track the air quality along with Also, cloud processing has been combined into an Internet holding web server for analyzing the info coming from the device to categorize as well as also picture internal air high quality depending on to the requirements. The host server shops all information in the cloud to give relevant information for added review of your first class in your home air and also, this general set may be managed on relocating vehicles like vehicles, buses to check out the ambiance in various locations. © 2022 Author(s).

**b)**

Paper	IoT-based water quality monitoring	IoT-based wastewater infrastructure management	IoT-based water infrastructure management	Not related
101	X	-	-	X

Explanation: The paper discusses an IoT-based system for monitoring air quality and controlling harmful gases, which is not related to water or wastewater infrastructure management. Therefore, it falls under the "Not related" category.

< 2 / 2

**M** then why there is x under IoT-based water quality monitoring as well!!!!

I apologize for the mistake in my previous response. There should not be an "X" under IoT-based water quality monitoring for paper 101 since it does not relate to water quality monitoring. Therefore, the correct categorization for paper 101 is:

Paper	IoT-based water quality monitoring	IoT-based wastewater infrastructure management	IoT-based water infrastructure management	Not related
101	-	-	-	X

Again, I apologize for any confusion caused.

**Figure 10.** (a) An illustration of ChatGPT input utilizing APA metadata and the abstract. (b) ChatGPT's response to the request. ChatGPT classified the article as both unrelated and in the water quality category. Nonetheless, reviewing the explanation from the user's perspective would aid in determining that the article is unrelated.

Notably, ChatGPT occasionally tends to retain articles to the maximum extent, even if they are unrelated, by assigning them to the closest corresponding category. Figure 10 provides an illustration of an article being classified into two categories, with one of them being "Not related". Alongside the classification outputs, ChatGPT also provides justifications for its selections, which are pivotal in informing the decision-making process. ChatGPT provides insights into the factors and reasoning underlying its decisions by explaining its classifications. This justification feature serves as a valuable tool for evaluating and validating the appropriateness of the classification decisions.

To address the challenge posed by articles being classified into two categories, one of which is "Not related", we leveraged the explanations provided by ChatGPT to assist in confirming decisions regarding article inclusion or exclusion. Practically, we collected the articles that ChatGPT assigned two categories and re-requested their classification. However, this time, we provided ChatGPT with the explanations accompanying its initial classifications. In practical applications, we recommended reading the justification provided by the ChatGPT for the articles classified into two classes to confirm the relevance of the article or not.

Similarly, we evaluated the performance of the classification from ChatGPT (APA + Abstract) by comparing ChatGPT's results (APA + Abstract) to our benchmark, which consisted of the opinions of experts. This evaluation aimed to assess the efficacy of the filtration process, particularly in relation to the "Not related" class (Figure 9(b1)). The results showed significant improvement when applying ChatGPT (APA + Abstract) compared to ChatGPT (APA) alone. Regarding precision, recall, and F1-score, the ChatGPT (APA + Abstract) achieved impressive values for the "Not related" class, with scores of 85.00%, 93.00%, and 90.00%, respectively. These metrics outperformed the corresponding scores obtained by ChatGPT (APA) (Figure 9(b2)). Furthermore, the F1-scores for the three other classes, namely, water management, wastewater management, and water quality, were also notably higher, with scores of 91.00%, 87.00%, and 86.00%, respectively. The implementation of ChatGPT (APA + Abstract) led to a reduction in misclassification rates of approximately 64% com-



pared to ChatGPT (APA), demonstrating its capacity for improved accuracy. Additionally, other evaluation measures, such as accuracy, macro-F1, and weighted F1, experienced enhancements. These findings collectively underscore the exceptional performance of ChatGPT (APA + Abstract) in effectively filtering and categorizing articles, positioning it as a valuable tool for subsequent classification and article exclusion with enhanced precision.

However, it is important to acknowledge that certain limitations remain, particularly regarding the number of articles that can be filtered simultaneously. While ChatGPT exhibits remarkable capabilities, practical constraints need to be considered when scaling up its application. This evaluation provides valuable insights into the effectiveness and potential of ChatGPT (APA + Abstract) as a robust classification system, offering improved precision and reliability in filtering and categorizing scientific articles. By combining AI-driven classification strengths with human evaluators' expertise, we can harness the power of automation while ensuring the highest standards of accuracy and relevance.

Despite the limitation on the feeding rate of articles into ChatGPT, it continues to surpass traditional filtering methods in terms of time efficiency. The performance of ChatGPT (APA + Abstract) in article filtering is considered outstanding. Therefore, ChatGPT (APA + Abstract) was utilized to screen all articles within the study. The comprehensive results of the filtering and categorizing of all articles can be found in Tables S3–S6. It is important to note that the output of this step goes beyond the elimination of articles; it also involves categorizing relevant articles into three main classes. Following the filtration process, a total of 351 articles were discarded as they were deemed irrelevant, while 145 articles were retained as relevant. The relevant articles were categorized into specific domains, with 76 articles on water management, 53 on wastewater management, and 32 on water quality. It is important to acknowledge that specific articles may overlap and fall into multiple categories, resulting in 161 articles across the three domains. However, when considering unique articles, the total count stands at 145.

Ultimately, the utilization of ChatGPT (APA + Abstract) in the filtration and categorization process demonstrates its effectiveness in efficiently managing a large volume of articles, streamlining the identification of relevant content, and facilitating the organization of articles based on their thematic relevance. By leveraging the capabilities of AI-powered classification, researchers can optimize their workflow, allocate their time more effectively, and enhance the accuracy and precision of their literature review processes.

### 3.3. Second-Round Article Filtration (Full-Text) and Information Extraction

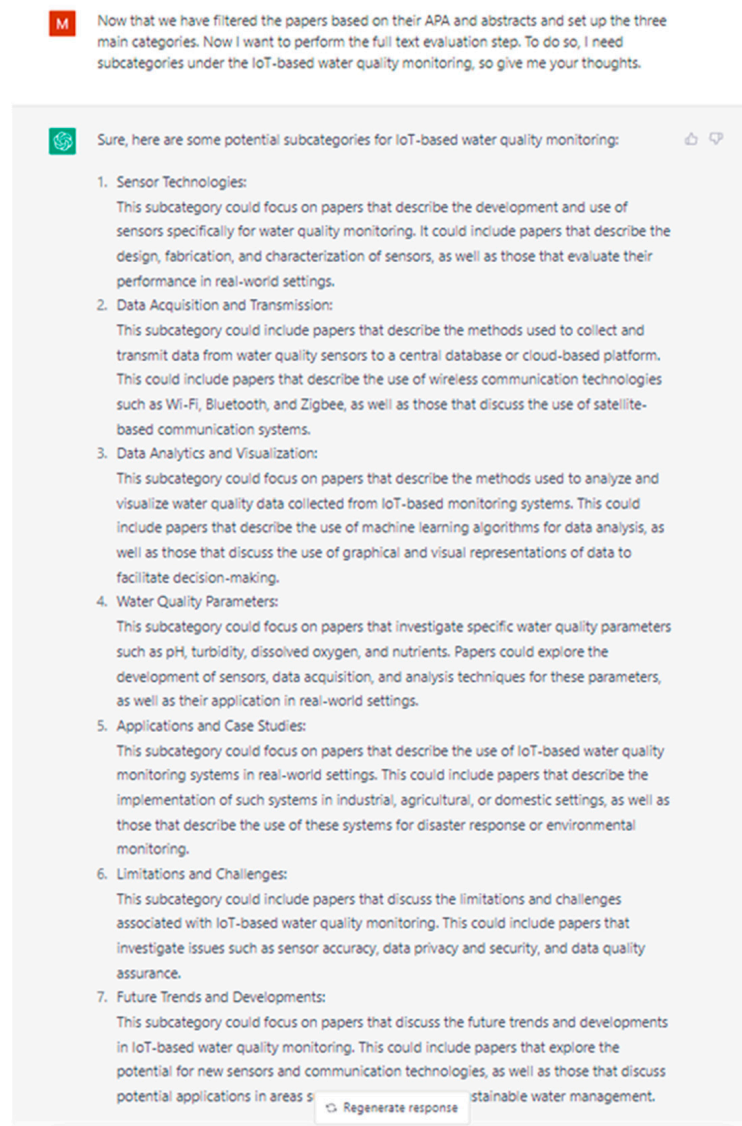
The full-text filtration and information extraction phase was carried out during a second round of article filtration to evaluate the suitability of the remaining 145 articles for inclusion in our review and extract valuable information from them. This challenge was addressed by utilizing the capabilities of ChatGPT for full-text filtering, as illustrated in Figure 6. To effectively leverage ChatGPT for this purpose, we initially identified five sub-categories within each main category to concentrate on specific areas of interest and ensure a comprehensive exploration of the diverse topics relevant to our review. These sub-categories were thoughtfully selected to cover diverse aspects of the subject matter, including sensors and sensing technology, data acquisition and transmission, data analytics and visualization, applications, case studies, and research gaps and trends.

To automate extracting information and harness the capabilities of ChatGPT, we devised a prompt that solicited ChatGPTs' recommendations for relevant questions pertaining to each sub-category. The response generated by ChatGPT to this request is depicted in Figure 11, while Figure 12 showcases the 14 questions that were generated belonging to the five sub-categories. It is important to note that these questions generated are of a general nature and elicit responses in the form of "yes" or "no". The answers to these questions by ChatGPT would help extract information from the articles and remove irrelevant articles. In this phase, we tested the performance of ChatGPT in two scenarios, including ChatGPT (APA) and ChatGPT (APA + Abstract + relevant information). Practically, the ChatGPT prompts were constructed using the article's APA, abstract, methodology, discussion, and

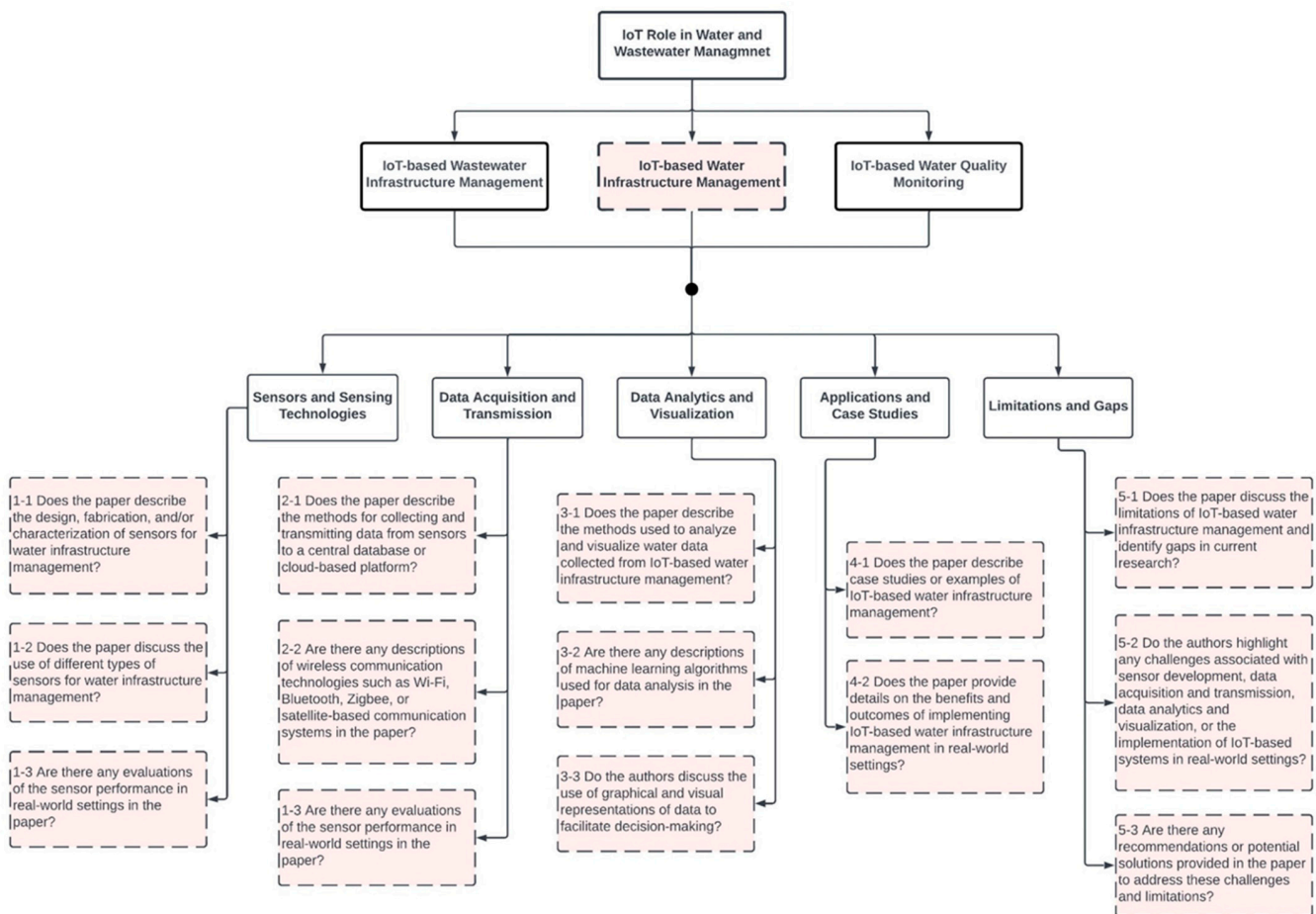
occasionally the conclusions section. Due to the extended length of these extracted sections from the articles compared to the previous steps (i.e., abstract only), the ChatGPT prompts were designed to handle one article at a time.

However, as previously discussed, the prompt's length is carefully adjusted to balance obtaining reliable responses from ChatGPT and saving time. It is worth noting that the time invested in this step is considerably less than the time of manual execution, particularly considering the added benefit of information extraction alongside the article's filtration.

During the assessment of ChatGPT's responses to the 14 questions, we observed three distinct scenarios. Firstly, when the answers to a question were "yes," ChatGPT confirmed this affirmative response and provided relevant information from the article that corresponded to the question (refer to Figure 13). Secondly, in instances where the answers were "No", ChatGPT simply reported "No" without furnishing any straightforward answers derived from the article (as shown in Figure 14). Lastly, when ChatGPT determined that the majority of answers were "No", it classified the paper as "unrelated" (as shown in Figure S9).



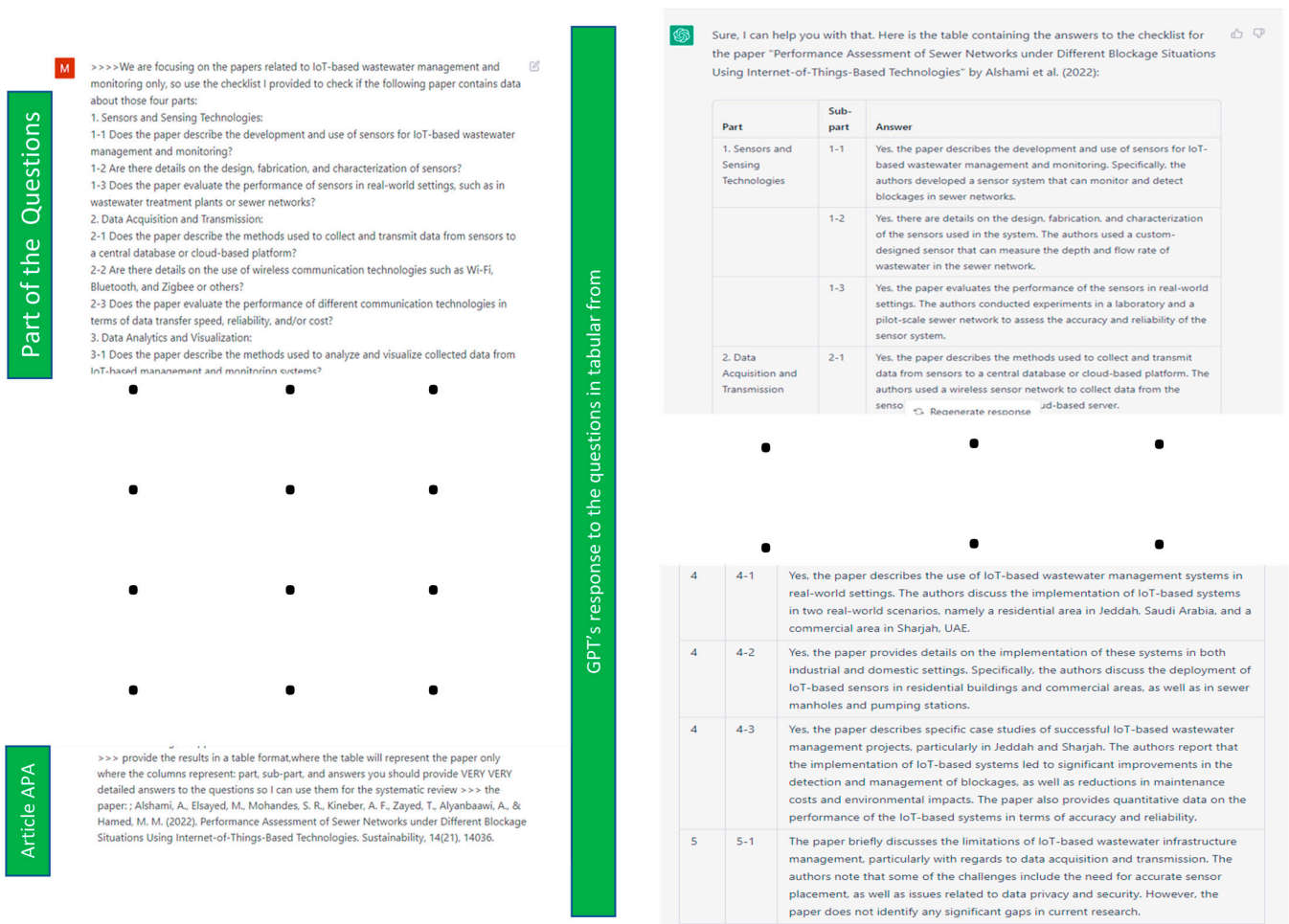
**Figure 11.** The ChatGPT's response to our request for proposing research questions that fit into each class. There are 14 questions in all.



**Figure 12.** Our systematic review taxonomy. The first level represents the three categories of the review, the second level depicts the sub-categories, and the third level illustrates questions to aid with information extraction. The 14 questions and five sub-categories are identical for each main category.

In this phase, we evaluated ChatGPT’s performance by comparing its responses to individual articles (we selected one article known for the authors as an example). Initially, we asked ChatGPT to answer these questions based on the article’s APA information. However, as demonstrated in Figure 14, where ChatGPT provided incorrect responses, APA information proved to be inadequate. For example, in Answer 1-1, ChatGPT mistakenly claimed that the author used the wrong type of sensors, and in Answer 4-1, ChatGPT inaccurately identified the research location as Saudi Arabia instead of Hong Kong.

To improve the accuracy of ChatGPT’s responses, we supplemented its understanding by incorporating additional information from the articles themselves. We considered various sections, including the titles, abstracts, methodology descriptions, relevant parts of the results, and conclusions, as these sections often provided more detailed and context-rich information compared to abstracts alone. However, we intentionally excluded article introductions and related work sections to maintain clarity and avoid confusion. Figure 15 provides an example of a ChatGPT prompt with a title, abstract, methodology description, and ChatGPT’s response to the questions. In this example, we used the same article as before, and it is evident that the quality of ChatGPT’s responses has significantly improved. For instance, in Answer 1-1, ChatGPT accurately reported the use of 58 ultrasonic sensors, and in Answer 4-1, ChatGPT correctly identified the research area’s location.



**Figure 13.** Illustration of a ChatGPT question-answer request prompt. The sole input was the APA article format. The left panel displays the ChatGPT's responses to these questions in the required tabular format. The dots indicate that a portion of the questions and answers were displayed, as the complete prompt and answers are too long to be presented.

At this stage, it can be concluded that by refining the prompt and incorporating additional article information, we enhanced the accuracy of ChatGPT's responses during the information extraction phase. This iterative process allowed us to leverage the strengths of ChatGPT while ensuring the reliability and validity of the extracted information. Nonetheless, human oversight and critical evaluation remained essential to validate and interpret the results obtained from ChatGPT.

To overcome the limitation of the subjective evaluation, we collaboratively answered the 14 questions for a subset of 30 articles, along with our responses to ChatGPT's outputs. Remarkably, despite the expected total of 420 individual answers for the 14 questions and 30 articles, our answers and ChatGPT's responses amounted to 381, owing to the classification of 3 articles as irrelevant. The summarized outcomes of these responses are presented in Figure 15, while more details about the answers can be found in Table S7. Among the 381 obtained responses, ChatGPT accurately captured 371, resulting in an impressive similarity rate exceeding 97%.



Part of the Questions

**M** We are focusing on the papers related to IoT-based wastewater management and monitoring only, so use the checklist I provided to check if the following paper contains data about those 4 parts:

- Sensors and Sensing Technologies:**
  - 1-1 Does the paper describe the development and use of sensors for IoT-based wastewater management and monitoring?
  - 1-2 Are there details on the design, fabrication, and characterization of sensors?
  - 1-3 Does the paper evaluate the performance of sensors in real-world settings, such as in wastewater treatment plants or sewer networks?
- Data Acquisition and Transmission:**
  - 2-1 Does the paper describe the methods used to collect and transmit data from sensors to a central database or cloud-based platform?
  - 2-2 Does the paper discuss the use of wireless communication technologies, such as Wi-Fi, Bluetooth, Zigbee, or satellite-based communication systems or others?
  - 2-3 Does the paper evaluate the performance of different communication technologies in terms of data transfer speed, reliability, and/or cost?
- Data Analytics and Visualization:**
  - 3-1 Does the paper describe the methods used to analyze and visualize collected data from IoT-based management and monitoring systems?

GPT's response to the questions in tabular form

Article  
APA

the systematic review >>> the paper: 21. Alshami, A., Elsayed, M., Mohandes, S. R., Kineber, A. F., Zayed, T., Alyanbaawi, A., & Hamed, M. M. (2022). Performance Assessment of Sewer Networks under Different Blockage Situations Using Internet-of-Things-Based Technologies. *Sustainability*, 14(21), 14036.

>>> Abstract: This study aims to model the performance of sewage networks under diverse blockage situations in terms of overflow occurrence using internet-of-things-based technologies in Hong Kong. To this end, a multi-stage methodological approach is employed, starting from collecting required data using smart sensors, utilizing novel data mining techniques, and using a case study simulation. From the results obtained, the following conclusions are drawn: (1) several sites under investigation are imbued with partial blockages, (2) the overall performance of the sewer network has a nonlinear relationship with the blockages in terms of the remaining time to overflow, (3) in cases of complete blockages, the sewer only takes few minutes to reach the manhole cover level that causes the system to experience overflow, and (4) cleaning work significantly improve the performance of the sewage network by 86%. The outcomes of this study provide a solid foundation for the concerned environmental engineers and decision-makers towards reducing the magnitude of sewer overflow and improving different aspects of our environment. © 2022 by the authors.

>>> Methodology: 21. Phase I: Data Collection  
This study is a part of a general research project that aims to develop a smart sewer monitoring system. This monitoring system will depend on the data collected from the sensing system connected to a cloud server. To do so, 58 smart ultrasonic level sensors and

abstract

Methodology

Part	Sub-part	Answer
Sensors and Sensing Technologies	1-1 Does the paper describe the development and use of sensors for IoT-based wastewater management and monitoring?	Yes, the paper describes the use of smart ultrasonic level sensors and data loggers for collecting data from 58 manholes in a sewer network.
	1-2 Are there details on the design, fabrication, and characterization of sensors?	Yes, the paper provides details about the smart ultrasonic level sensors, including their communication protocols, measurement capabilities, and power-saving features.
	1-3 Does the paper evaluate the performance of sensors in real-world settings, such as in wastewater treatment plants or sewer networks?	No, the paper does not explicitly evaluate the performance of sensors in real-world settings.
Data Acquisition and Transmission	2-1 Does the paper describe the methods used to collect and transmit data from sensors to a central database or cloud-based platform?	Yes, the paper mentions the use of a cloud server to collect data from the sensors connected to a sensing system. The data is collected via an HMW intelligent data logger and transmitted to the server through the internet.
	2-2 Does the paper discuss the use of wireless communication technologies, such as Wi-Fi, Bluetooth, Zigbee, or satellite-based communication systems or others?	No, the paper does not mention the specific wireless communication technologies used for data transmission.
	3-3. Use of graphical or visual representations of data	No, the paper does not describe the use of graphical or visual representations of data for decision-making.
4. Applications and Case Studies	4-1. Use of IoT-based wastewater management systems in real-world settings	Yes, the paper describes the use of IoT-based technologies for monitoring sewage networks in Hong Kong. It mentions the selection of manholes facing overflow complaints as a preliminary sample for developing the monitoring system.
	4-2. Implementation of systems for disaster response or environmental monitoring	No, the paper does not specifically discuss the implementation of IoT-based systems for disaster response or environmental monitoring.
	4-3. Case studies and evaluation of impact	Yes, the paper provides case studies on the performance assessment of sewer networks under different blockage situations. It evaluates the impact of cleaning work on the performance of the sewage network, showing an improvement of 86%. However, it does not evaluate the impact on wastewater treatment efficiency, environmental sustainability, or public health.

**Figure 14.** Illustration of a ChatGPT question-answer request prompt. The sole input was the article titles, abstracts, and methods section portions. The left panel displays the ChatGPT's responses to these questions in the required tabular format. The dots indicate that a portion of the questions and answers were displayed, as the complete prompt and answers are too long to be presented.

Response Category	ChatGPT Frequency	Authors Frequency
yes	~220	~215
no Response	~155	~160
not	~5	~5

**Figure 15.** A comparison of the ChatGPT's response to the authors' general response for the 30 articles in the sample.



Regarding discarding articles, both ChatGPT and the authors agreed on the same articles. However, it is worth noting that ChatGPT's responses were completely different for unrelated articles, and it stopped responding to questions (Please refer to Figure S9). This substantial level of agreement underscores the efficacy of ChatGPT in effectively comprehending and extracting information from the articles. Upon evaluating the efficacy of this approach in filtering the initial set of 145 articles, we successfully identified 56 articles as irrelevant, enabling us to focus on extracting pertinent information from the remaining 86 articles. This demonstrates the valuable role of ChatGPT in streamlining the article filtration process and automating information extraction from a substantial number of articles.

Since the snowballing process is an integral part of conducting an SR, we employed both backward and forward snowballing techniques to uncover additional relevant studies that might have been overlooked during the initial database search [24]. The backward snowballing method involves scrutinizing the references of the included papers to identify related articles, while the forward snowballing technique entails searching for studies among the articles that cited the included ones [24]. We manually conducted the snowballing process in this study by screening the titles of articles. However, we recognize the potential of leveraging ChatGPT to automate this step in order to advance the full automation of the SR process. By implementing the snowballing strategy, we successfully identified 52 new articles through multiple iterations in addition to the articles previously identified. These 52 articles underwent the same comprehensive filtration method outlined earlier in our methodology. As a result, 19 articles were excluded due to their lack of relevance, while the remaining 33 articles met the criteria for inclusion in our review database. Consequently, the total number of relevant articles included in our review increased to 119.

Overall, leveraging ChatGPT ensures a more thorough filtering process, assists in extracting information based on responses to comprehensive questions, and enables the inclusion of snowballing articles, expanding our review's breadth and scope. By capitalizing on ChatGPT's capabilities, we enhance the SR methodology's efficiency, accuracy, and reliability.

#### *3.4. Analysis and Interpretation of Extracted Information*

This phase focuses on analyzing the content collected in the previous phases, explicitly emphasizing the sub-categories outlined in Figure 12. The flowchart for phase 3 is illustrated in Figure 16, providing a visual representation of the analysis process. In order to streamline the analysis process, the "Yes" responses to each question were initially compiled and organized. Subsequently, these compiled responses were further analyzed and presented in Table S8. This approach facilitates a cohesive and structured data exploration, allowing for a more rigorous examination of the insights obtained from ChatGPT.

Accordingly, Table 4 provides a comprehensive overview of the response statistics obtained during Phase 2 and the corresponding objectives for analyzing each question. These responses served as prompts for ChatGPT, with a maximum of ten responses per prompt, covering all sub-categories outlined in Figure 12. The content analysis encompassed information extraction related to sensors and sensing technologies, data acquisition and transmission, data analytics and visualization, and applications and case studies, as well as limitations and gaps identified in the reviewed articles. Leveraging ChatGPT as an analytical tool facilitated a more thorough identification of various patterns and trends within the data analysis process. For example, a specific prompt was designed to explore the utilization of multiple types of sensors and their associated benefits, as depicted in Figure 17.

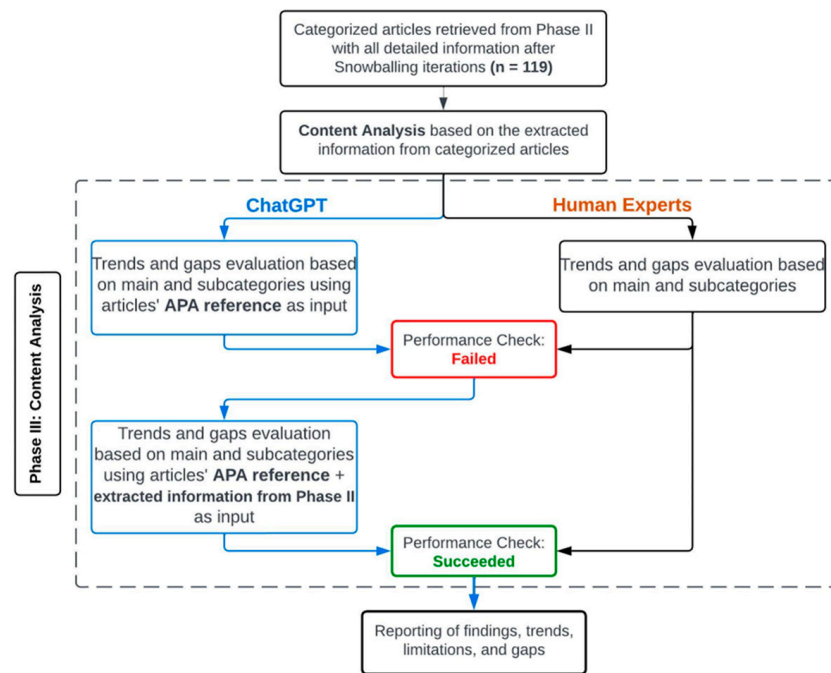


Figure 16. Flow chart of Phase 3.

M from these results Give me a list of the used sensors across all papers and the frequency of these sensors (e.g., flow sensors used in 3 papers)>>>>also, use your knowledge to understand the used sensors, they might be not a clear, you might group them>>> give the results in Table format where the columns represent the sensors type, frequency, and APA citation>>> (Abbas et al., 2022) the paper describes the design and fabrication of sensors for water quality monitoring. Specifically, it describes the use of a water flow sensor, a waterproof ultrasonic sensor, and two temperature sensors (DS18B20 sensor) for sensing water and ambient temperatures. The paper discusses the use of different types of sensors such as those for water flow, water level, and temperature. Specifically, the paper explains the use of a Hall effect-based DN25 flow sensor, JSN-SR04T Integrated Ultrasonic Ranging transducer, and DS18B20 sensor. (Ahmed et al., 2022) The paper discusses the use of sensors for water quality monitoring, but there is no detailed description of the design, fabrication, or characterization of sensors. the paper discusses the use of different types of sensors such as pH sensors, turbidity sensors, dissolved oxygen sensors, and electrical conductivity sensors. (Al-Khashab et al., 2019) The paper describes the design, fabrication, and characterization of sensors for water quality monitoring. Specifically, the paper discusses the use of sensors for measuring temperature, pH, total dissolved solids (TDS), turbidity, and electrical conductivity (EC). The paper discusses the use of different types of sensors, including sensors for pH, temperature, TDS, turbidity, and EC. (Campagnaro et al., 2022) The paper mentions the design and deployment of low-cost sensor nodes for water quality monitoring. The nodes are equipped with sensors to collect measurements of temperature, pressure, pH, turbidity, dissolved oxygen (DO), and electrical conductivity (EC). the paper mentions that the sensor nodes will collect temperature, pressure, pH, turbidity, dissolved oxygen, and electrical conductivity measurements. (Chen & Han, 2018) the paper describes the design, fabrication, and characterization of sensors for water quality monitoring. The authors used multiple sensors to measure different parameters, including pH, temperature, turbidity, dissolved oxygen, and oxidation-reduction potential. The sensors were placed in a water quality monitoring system that used wireless sensor network (WSN) technology. the paper discusses the use of different types of sensors for water quality monitoring. The a dissolved oxygen, and oxidation-reduction potential.

wind speed, and direction. >>>>>remember: you should go through all of the papers, and used the information mentioned in all papers do not miss any paper.>>> note that if the explanation says such as, then this is a sensor.

Based on the papers, the following sensors were used for water quality monitoring:

Sensor Type	Frequency	APA Citation
pH	7	Abbas et al. (2022); Ahmed et al. (2022); Al-Khashab et al. (2019); Chen & Han (2018); Das & Jain (2017); Geetha & Gouthami (2016); Kumar et al. (2023)
Temperature	7	Abbas et al. (2022); Al-Khashab et al. (2019); Campagnaro et al. (2022); Chen & Han (2018); Das & Jain (2017); Kumar et al. (2021); Madhurima et al. (2022)
Turbidity	5	Ahmed et al. (2022); Al-Khashab et al. (2019); Geetha & Gouthami (2016); Kumar et al. (2021); Kumar et al. (2023)
Electrical Conductivity	4	Ahmed et al. (2022); Al-Khashab et al. (2019); Das & Jain (2017); Kumar et al. (2023)
Water Flow	3	Abbas et al. (2022); Madhurima et al. (2022); Geetha & Gouthami (2016)
Dissolved Oxygen	3	Campagnaro et al. (2022); Kumar et al. (2023); Chen & Han (2018)
Pressure	1	Campagnaro et al. (2022)
Total Dissolved Solids (TDS)	1	Al-Khashab et al. (2019)
Oxidation-Reduction Potential	1	Kumar et al. (2023)

Note that some papers may mention multiple times, but we only

Figure 17. User prompt and ChatGPT answer for the use of different types of sensors.

Similarly, trends in data transfer technologies were examined based on the responses to question 2-1 (Figure 12). Figure 18 illustrates ChatGPT’s responses concerning the specific applications of wireless communication technologies. Furthermore, multiple prompts were devised within the data analysis and the visualization section. These prompts aided in exploring diverse approaches employed for data analysis, including AI and ML techniques, as well as visualization methods utilized for decision-making processes (Figure S10). Additionally, questions 4-1 and 4-2 were integral to the review process, assessing the implementation of proposed systems or case studies in the studied papers while identifying prevailing trends and scopes (Figure 19). The benefits associated with such implementations were also investigated within each article (Figure S11).

**Table 4.** The gathered responses (yes) for each of the three major categories.

Sub-Category	Question	Water Quality Monitoring	Water Infrastructure Management	Wastewater Infrastructure Management.	Objectives
Answers (YES)					
Sensors/development	1-1: Sensor development.	26	28	36	Identify trends in sensor development and manufacturing, study the advantages of employing several sensors, investigate the frequency of sensor use, categorize sensors according to their functionality, and investigate the methods used to evaluate sensor performance.
	1-2: Use of different types of sensors.	18	37	19	
	1-3: sensors performance evaluation.	21	15	15	
Data transmission	2-1: Data collection and transmission method.	33	45	38	Identify trends and anomalies in transmission methods, including the utilization of wireless communications, the types of wireless technologies employed, and the frequency of their occurrence in the examined papers. Analyze, also, the effectiveness of utilizing various communication technologies.
	2-2: Use of wireless communication.	31	38	31	
	2-3: Connectivity performance evaluation.	7	5	5	
Data analysis	3-1: Data analysis methods.	20	33	14	Define frequently applied data analysis techniques, including AI and ML techniques, and study the trends in visualization approaches.
	3-2: Use of ML algorithms.	6	11	0	
	3-3: Data visualization to facilitate decision-making.	12	19	17	
Case studies	4-1: The use in real-world settings.	28	39	28	Identify trends in the implementation of IoT-based systems in various real-world contexts and the outcomes and advantages of these implementations.
	4-2: Benefits and outcomes.	17	37	27	
limitations and gaps	5-1: limitations and gaps in current research.	14	24	15	Define the limitations and gaps identified by the authors, the obstacles encountered in implementing their systems, the offered solutions, and the recommendations for overcoming them.
	5-2: Implementation challenges.	25	42	23	
	5-3: Recommendations or solutions.	20	37	16	

The analysis stage also involved thoroughly examining the limitations and research gaps discussed in previous studies, along with the corresponding recommendations put forth by researchers. Leveraging ChatGPT in this phase facilitated a comprehensive exploration and in-depth understanding of the challenges and limitations encountered in prior research and the proposed solutions adopted to address them. To ensure a systematic approach to identifying and categorizing the limitations and challenges discussed by different authors, a carefully designed prompt (Figure 20) was employed, utilizing the results obtained from questions 5-1 and 5-2 in Figure 12.

**M** from these results Give me a list of the used wireless communication technologies across all papers and the frequency of these wireless communication technologies (e.g. Lora used in 3 papers)>>>also, use your knowledge to understand the used wireless communication technologies, they might be not a clear, you might group them>>> give the results in Table format where the columns represent the wireless communication technologies type, frequency, and APA citation>>>the papers: (Abbas et al., 2022) The paper discusses the use of wireless communication technologies such as WiFi to transmit the data collected by the sensors. (Abdullah et al., 2022) The paper discusses the use of wireless communication technologies, such as LoRa, for data transmission. The high availability but low-cost wireless communication infrastructure such as LoRa is the perfect solution in this scenario. (Ahmed et al., 2022) the paper discusses the use of various wireless communication technologies such as Bluetooth, Zigbee, WiFi, and Wireless Sensor Network (WSN) to transfer the sensor data from the sensor module to the cloud server. (Al-Khashab et al., 2019) The paper discusses the use of wireless communication technologies, specifically the Wi-Fi module (ESP 8266), for transmitting data from the sensors to the cloud-based platform. (Campagnaro et al., 2022) The paper discusses the use of LPWANs for data transfer from

(Kabi & Maina, 2021) The paper discusses the use of LoRa as the communication protocol, but does not provide information on other wireless communication technologies. >>>>Remember: you should go through all of the papers, and used the information mentioned in all papers do not miss any paper

Based on the information provided in the papers, the following wireless communication technologies were used:

Wireless Communication Technologies	Frequency	APA Citation
WiFi	4	Abbas et al. (2022); Al-Khashab et al. (2019); El Sayed et al. (2019); Geetha & Gouthami (2016)
LoRa	3	Abdullah et al. (2022); Campagnaro et al. (2022); Kabi & Maina (2021)
Zigbee	2	Ahmed et al. (2022); Das & Jain (2017)
Bluetooth	1	Ahmed et al. (2022)
Wireless Sensor Network (WSN)	1	Ahmed et al. (2022)
LPWANs	1	Campagnaro et al. (2022)
GSM	1	Das & Jain (2017)

Note: The technologies mentioned in Ahmed et al. (2022) are counted separately, as they are not used in conjunction with each other.

Regenerate response

**Figure 18.** User prompt and ChatGPT answer for questions related to wireless communication technologies.

**M** \*>>> The Section: IoT-based water quality monitoring >>>The Subsection: Applications and Case Studies >>>The question: Does the paper describe case studies or examples of IoT-based water infrastructure management? >>> So based on the following answers from the paper, tell me about the research trends and what the researchers have done >>> answer this question in detail and provide citations>>> Paper Case Study Description  
 Abbas et al., 2022 Developing Countries Investigated the use of IoT-based water quality monitoring systems in developing countries, specifically for households that rely on overhead water tanks for running water.  
 Abdullah et al., 2022 Orang Asli Community, Malaysia Implemented an IoT-based water quality monitoring system in the indigenous Orang Asli community in Chini Lake, Pahang, Malaysia, to provide water level alerts for mitigating flooding and drought situations and Internet access for promoting local products and services.  
 Ahmed et al., 2022 Not Applicable Proposed an edge-cloud ubiquitous sensor network for low-cost water quality measurement to supplement existing IoT-based infrastructure.  
 Al-Khashab et al., 2019 Mosul City, Iraq Monitored drinking water in Mosul city using an IoT-based water quality monitoring system.  
 Blanco-Gomez et al., 2023 Mar Menor Lagoon, Spain Implemented a low-cost monitoring system that could be integrated into a small-sized buoy and attached to fishing and recreational boats to gather water quality information in the Mar Menor lagoon in Southeast Spain.  
 Campagnaro et al., 2022 Venice Lagoon, Italy Deployed an IoT-based water quality monitoring system in the Venice Lagoon in Italy.  
 Chen & Han, 2018 Bristol Floating Harbour, UK Studied the use of IoT-based water quality monitoring systems in Bristol Floating Harbour.  
 El Sayed, Al-Kady, & Siddik, 2019 Not Applicable Developed an IoT-based water quality monitoring system for SWTPs to manage Big Data information flows from many sensors and smart devices that allow for real-time responses and connectivity to IoT cloud platforms.  
 Geetha & Gouthami, 2016 In-Pipe Domestic Water Monitoring Designed a system for in-pipe domestic water quality monitoring that alerts remote users when there is a deviation from predefined standard values.  
 Kabi & Maina, 2021 Upper Ewaso Nyiro River, Kenya Studied the use of IoT-based water level monitoring systems in the Upper Ewaso Nyiro River in Kenya.  
 Kumar et al. (2023) Ganga river and Sangam at Prayagraj, Uttar Pradesh, India IoT-based water quality monitoring systems v ttings to assess water quality during different months and seasons

Research Trends and Case Studies in IoT-based Water Quality Monitoring:

Trend	Case Studies	Description
Trend 1: Agricultural Water Management	Kabi & Maina, 2021	Studied the use of IoT-based water level monitoring systems in the Upper Ewaso Nyiro River in Kenya.
	Wong et al., 2021	Described the use of IoT-based water quality monitoring systems in a palm oil plantation on Carey Island, Malaysia.
Trend 2: Drinking Water Quality Monitoring	Al-Khashab et al., 2019	Monitored drinking water in Mosul city using an IoT-based water quality monitoring system.
	Saravanan et al., 2018	Described the use of IoT-based water quality monitoring systems in the Tirunelveli Corporation (Metro city of Tamilnadu state, India) for real-time water quality monitoring.
Trend 3: Flood Monitoring	Abdullah et al., 2022	Implemented an IoT-based water quality monitoring system in the Orang Asli community in Chini Lake, Malaysia, for flood and drought mitigation, and Internet access for promoting local products and services.
	Mekruksavanich et al., 2021	Used IoT-based water monitoring systems for flood prediction in real-world settings.
Trend 4: Method Oriented	Ahmed et al., 2022	Proposed an edge-cloud ubiquitous sensor network for low-cost water quality measurement to supplement existing IoT-based infrastructure.
	Purkovic et al., 2019	Used EnOcean long-range sensor technology for river monitoring and early flood detection system in Japan.
Trend 5: Multiple Areas	Abbas et al., 2022	Investigated the use of IoT-based water quality monitoring systems in developing countries, specifically for households relying on overhead water

Regenerate response

**Figure 19.** User prompt and ChatGPT answer for the trends within the proposed systems or case studies.

This approach allowed for extracting and organizing valuable insights from the collected data. Additionally, a comprehensive list of recommendations was compiled, drawing from the proposed solutions identified in question 5–3 and categorized based on common trends (Figure 21).

This approach yielded a wealth of information regarding the challenges, limitations, and potential solutions found in the reviewed articles. In order to gain a deeper understanding and assess the extent of the resolved issues, a ChatGPT prompt was utilized



to compare the limitations and the challenges highlighted by various authors with the suggested solutions and recommendations. This comparative analysis provided valuable insights into the existing research gaps and identified areas for further investigation and research. An example depicting the resulting research gaps is illustrated in Figure 22.

**M** Now that I have this list of gaps, can you group them based on your understanding into 3 or 4 groups >>> the list: Gap  
 Lack of data for initial benchmarking and analysis.  
 Expensive existing water quality monitoring systems that require periodic maintenance.  
 Need for a more comprehensive analysis of water quality and the integration of other technologies such as Artificial Intelligence (AI) and Machine Learning (ML) in water quality management  
 Lack of forward-looking for water quality estimation in traditional water quality control research, which can have a significant influence on practical systems and people's health  
 The challenge of real-time measurements being scarce in the Mar Menor lagoon, which makes it complicated to make proper comparisons of the EC + T track results  
 Lack of forward-looking for water quality estimation in traditional water quality control research, which can have a significant influence on practical systems and people's health  
 Limited discussion on gaps in current research related to IoT-based water quality monitoring  
 Lack of discussion on challenges related to data analytics and visualization  
 Need for effective control strategies to integrate current system states and avoid peak discharges in the sewer system  
 Need for machine learning techniques to optimize control systems  
 Few ML implementations exist for water quality monitoring.  
 Existing sensing technologies are available, but few ML implementations exist for water quality monitoring.  
 Traditional manual water quality monitoring scheme can provide reasonable seasonal water quality variation in the long term, but it lacks sufficient detail about diurnal fluctuation of water quality.  
 The need for low-cost and reliable sensors for water quality monitoring.  
 Achieving long-term performance of sensors in harsh environmental conditions without periodic maintenance.  
 Need for accurate and reliable data acquisition and transmission, as well as effective data analytics and visualization methods.  
 Limitations of the 2G wireless link in a rural environment and the need to migrate to 3G or LTE network infrastructure.  
 Need for battery-free sensor tags and a suitable communication infrastructure to obtain data  
 Lack of discussion on limitations of IoT-based water quality monitoring in

Category	Challenges	Papers
Sensor Sustainability Gap	Lack of data for initial benchmarking and analysis. Expensive existing water quality monitoring systems that require periodic maintenance. Need for low-cost and reliable sensors for water quality monitoring. Achieving long-term performance of sensors in harsh environmental conditions without periodic maintenance. Need for battery-free sensor tags and a suitable communication infrastructure to obtain data. More research is needed on the development of low-power sustainable integrated underwater and above water IoT sensor networks for coastal and littoral areas, which are highly heterogeneous and variable over short (daily), medium and long (seasonal, inter-annual) timescales.	Abbas et al., 2022; Malissovas et al., 2022; Suresh et al., 2021; Campagnaro et al., 2022
Communication Breakdown Gap	Limited discussion on gaps in current research related to IoT-based water quality monitoring. Limitations of the 2G wireless link in a rural environment and the need to migrate to 3G or LTE network infrastructure. Need for secure and reliable communication channels for data acquisition and transmission.	Pasika & Gandla (2020); Nordin et al., 2018
Comprehensive Analysis Gap	Need for a more comprehensive analysis of water quality and the integration of other technologies such as Artificial Intelligence (AI) and Machine Learning (ML) in water quality management. Lack of forward-looking for water quality estimation in traditional water quality control research, which can have a significant influence on practical systems and people's health. Traditional manual water quality monitoring scheme lacks	Saravanan et al., 2018; Wu et al., 2018; Chen & Han, 2018; Pasika & Gandla, 2020; Mendoza-Cano et al., 2021; Zakaria, Nordin, & Ruslan, 2020; Campagnaro et al., 2022; Mekruksavanich et al., 2021; El Sayed,

**Figure 20.** User prompt and ChatGPT answer to identify and categorize the limitations and challenges discussed by previous authors.

**M** >>> The Section: IoT-based water quality monitoring >>> The Subsection: Limitations and Gaps >>> The questions: Are there any recommendations or potential solutions provided in the paper to address these challenges and limitations? >>> From these results, Give me a list of the recommendations or potential solutions across all papers and the frequency of these recommendations or potential solutions >>> Also, use your knowledge to understand the provided recommendations or potential solutions; they might not be clear, and you must group them>>> give the results in Table format, where the columns represent the recommendations or potential solutions, frequency, and APA citation >>> the papers: (Wong et al., 2021) The paper provides recommendations to address the challenges of sensor development, such as the use of low-cost sensors and the need for calibration and validation. However, it does not provide recommendations to address the challenges associated with the implementation of IoT-based systems in real-world settings. (Abdullah et al., 2022) The paper provides recommendations for successful application planning, which include several considerations such as the terrain and foliage impact, suitable parameter setup, as well as optimal gateway (GW) mounting location and height, to guarantee optimal signal quality and wider coverage areas. The paper proposes a wireless machine-to-machine (M2M) network, sensors technologies, and a big data analytic enablement platform as solutions to address the challenges associated with IoT-based water quality monitoring in rural areas. (Blanco-Gómez et al., 2023) They do, however, suggest that the use of such technology for citizen science activities could lead to interesting information for the scientific community, with daily updates of information and different routes that could bring the opportunity to create daily raster maps of EC and T at Mar Menor lagoon scale. (Campagnaro et al., 2022) It also highlights the use of LPWANs to support the data collection from a dense sensor deployment. Possible future work may include the improvement of the prototype and further testing in real-world scenarios as well as simulations with a deployment fit for the Venice lagoon. (El Sayed, Al-Kady, & Siddik, 2019) The authors conclude that their proposed solution will improve the overall management services on the on-site processing of SWTP and enable the supervisory control of the overall system, contributing to the stable operation of the SWTP digital infrastructure system. (Geetha & Gouthami, 2016) The paper provides potential solutions to address the challenges and limitations, such as using a single chip microcontroller with an in-built Wi-Fi module to reduce cost and complexity and a cloud-based platform for data storage and analysis.

Category	Recommendation	APA Citation
Sensors development and use	Use of low-cost and reliable sensors	Malissovas et al., 2022; Ahmed et al., 2022
Sensors development and use	Future research using fully waterproof sensors to do more reliable system testing	Shanmugam et al., 2021
Connectivity enhancement	Exploration of alternative wireless backbone technologies	Nordin et al., 2018
Connectivity enhancement	Extension of wireless infrastructure to cater for early warning systems	Nordin et al., 2018
Connectivity enhancement	Further research into propagation modeling for NB-IoT in remote areas with thick and tall foliage	Nordin et al., 2018
Enhancing data management and analysis	Implementation of cloud-based data management system	Saravanan et al., 2018
Enhancing data management and analysis	Integration of IoT-based water quality monitoring with other technologies such as AI and ML	Saravanan et al., 2018
Enhancing data management and analysis	Use of open data formats and open-source software	Salgado et al., 2022
Enhancing data management and analysis	Potential use of cloud computing and machine learning in water quality monitoring systems	Salgado et al., 2022
System Optimization	Use of machine learning algorithms	Saravanan et al., 2018; Abbas et al., 2022; Ahmed et al., 2022
System Optimization	Involve design and	Salgado et al., 2022

**Figure 21.** User prompt and ChatGPT answer to generate a comprehensive list of the compiled recommendations.



**M** From the following two tables, find the unsolved research gaps. To do so, you should compare the two tables (the Challenge/Limitations/Gaps and the Recommendations/Solutions) and find the unsolved problems, also mention the number of the point that you support your argument with!

>>>>Tablet: Challenge/Limitations/Gaps  
Lack of off-the-shelf integrated systems with appropriate sensors for measuring and communicating water temperature and flow rate.  
Cost restrictions.  
Long-term deployability related to battery-powered devices.  
Turbidity reading can be highly suspicious, and aeration caused by ship waves may cause bubbles trapped against the turbidity lenses, resulting in doubtful turbidity readings.  
High initial installation cost, intensive computer processing needed, and the system being prone to vandalism.  
Potential interference from other environmental factors that may affect the accuracy of the measurements.  
Calibration of sensors.

>>>>Table 2: suggested solution: Recommendations/Solutions  
Recommendation  
Exploration of alternative wireless backbone technologies  
Extension of wireless infrastructure to cater for early warning systems  
Further research into propagation modeling for NB-IoT in remote areas with thick and tall foliage  
Improvements to telemetry stations  
Successful application planning (considering terrain, foliage impact, suitable parameter setup, optimal gateway mounting location and height)  
Use of LPWANs for data collection  
Implementation of cloud-based data management system  
Integration of IoT-based water quality monitoring with other technologies such as AI and ML  
Use of open data formats and open-source software  
Potential use of cloud computing and machine learning in water quality monitoring systems  
Involving more parameters for water quality monitoring  
Use of cloud-based platform for data storage and analysis  
Use of technology for citizen science activities  
Use of low-cost and reliable sensors  
Future research using fully waterproof sensors to do more reliable system testing

Based on a comparison of the provided tables, here are the research gaps or unsolved problems that can be identified:

1. Lack of low-cost, reliable, and accurate sensors: Proposed solutions suggest using battery-powered sensors, but these have limitations regarding long-term deployability and potential interference from other environmental factors. To address these challenges, there is a need for low-cost and reliable sensors for water quality monitoring that can achieve long-term performance in harsh environmental conditions without periodic maintenance. Further research is needed to develop low-power, sustainable, and integrated underwater and above-water IoT sensor networks suitable for various conditions.
2. Challenges associated with data acquisition and transmission: Although some proposed solutions suggest the need for secure and reliable communication channels, they do not fully address the technical issues associated with wireless link limitations and network coverage in remote areas. Further research is needed to investigate different communication mechanisms, such as Zigbee, Sigfox, and NB-IoT, and to integrate various techniques to generate a more integrated system. Moreover, there is a need to migrate to LTE network infrastructure, and technical challenges associated with LoRaWAN technology, such as obstructed propagation of electromagnetic waves in remote areas, need to be addressed.
3. Comprehensive analysis of water quality: While some solutions propose the integration of AI and ML in water quality management, a more thorough analysis of water quality is still needed. Traditional manual monitoring schemes lack sufficient detail about diurnal fluctuations in water quality. Further research is necessary to improve the accuracy and depth of water quality analysis.
4. Lack of standardization: The lack of standardization in data collection, analysis, and reporting limits the interoperability of IoT-based systems for water quality monitoring. This issue could be addressed by developing standard protocols for sensor calibration, data processing, and quality assurance. In addition, developing a centralized database that collects and stores data from various sources could facilitate data sharing and analysis. Additionally, the collaboration between stakeholders such as water

**Figure 22.** User prompt and ChatGPT answer for comparing the limitations and challenges highlighted by various authors with the suggested solutions and recommendations.

#### 4. ChatGPT Strengths, Limitations, and Future Directions in Automating SR Process

ChatGPT, built on the GPT-3.5 architecture, represents a significant breakthrough in AI research, enabling the generation of coherent and meaningful human-like language by leveraging vast amounts of language data. This innovative language model holds promise for various domains, including systematic reviews, and can potentially contribute to the advancement of general artificial intelligence. However, it is important to acknowledge that, being a generative model, ChatGPT cannot guarantee the absolute accuracy of its outputs. Therefore, this section will explore the strengths, limitations, potential areas for enhancement, and future research directions concerning ChatGPT in the context of conducting SRs.

##### 4.1. Strengths of ChatGPT in SR Process

ChatGPT has been proven to be a valuable tool in the SR process, offering several strengths that enhance the efficiency and effectiveness of the methodology. Through our methodology and evaluation, we have identified the following key strengths of ChatGPT in conducting SRs:

1. **Full Automation:** ChatGPT contributes to automating several tasks in the SR process, such as generating research questions, suggesting BRTs, categorizing the relevant articles, discarding unrelated ones, proposing sub-categories to be covered for each main category, generating research questions to aid in information extraction from the articles, and extracting all relevant information. This level of automation facilitated by ChatGPT helps streamline the SR process and decrease the time and errors.
2. **Enhanced accuracy and efficiency:** ChatGPT offers a valuable advantage by improving the accuracy and efficiency of filtering and classifying articles. Researchers can benefit from its ability to swiftly identify relevant studies, reducing uncertainty, and saving significant time and effort. Moreover, ChatGPT's proficiency in natural language processing aids in precise content analysis, minimizing the risk of errors, and omissions in research interpretation.

3. **Time-saving:** ChatGPT demonstrates significant potential in saving time during SRs, which are known to be time-consuming and resource-intensive processes that require high levels of expertise and attention to detail. ChatGPT assists in this process by swiftly analyzing and summarizing large volumes of the literature, aiding researchers in identifying relevant studies and extracting key information more efficiently. In our study, ChatGPT played a significant role in tasks such as filtering, categorizing, and content analysis, which resulted in decreased time and effort as well as reduced sources of uncertainty. However, it is important to note that human experts should carefully review ChatGPT-generated summaries.
4. **Improved reproducibility:** While ChatGPT's responses were found to be influenced by the user prompts, the same procedure can be replicated multiple times by following the same guidelines and adhering to the recommended approach. This enhances the reproducibility of the results, allowing for consistent outcomes to be obtained through repeated application of the methodology. ChatGPT's responses are markedly affected by the user prompts, and the same procedure can be reproduced several times by conducting the same procedures and following the recommendations.
5. **Flexibility:** The method introduced utilizing ChatGPT for automating the SR process can be applied for conducting SRs across various fields. This flexibility allows for the potential utilization of ChatGPT in various research domains, providing opportunities for its application beyond the specific context of the current study.

#### 4.2. Limitations of ChatGPT in SR Process

ChatGPT, despite its strengths, also has certain limitations that need to be considered when applying it to the SR methodology. These limitations arise from the nature of the model and the challenges associated with its implementation in complex research tasks. Understanding these limitations and constraints is considered crucial to ensuring the appropriate use and interpretation of ChatGPT-generated outputs in the SR process. This subsection discusses the limitations of ChatGPT in the context of SR methodology and identifies improvement opportunities. Our study has uncovered the following limitations:

1. **Limited ability to extract full-text articles:** Despite ChatGPT's capability to suggest and adjust BSTs based on user requests, it is not optimized for article extraction, which may impact the comprehensiveness of the SR. As a result, ChatGPT's limitations in extracting articles may constrain the SR process's completeness.
2. **Limited ability to extract all information from articles:** Despite ChatGPT's capability to filter, categorize articles, and extract text information, it may encounter limitations in extracting all relevant information, especially if the information is presented in non-standard formats such as figures or other non-textual forms. This may result in incomplete extraction of relevant data, particularly from articles that utilize non-traditional data presentation methods, potentially impacting the comprehensiveness and accuracy of the extracted information during the SR process.
3. **Dependence on input data:** ChatGPT's performance highly depends on the input data quality. If the data is biased or incomplete, GPT's output may be similarly flawed.
4. **Limited Access to Real-Time Data:** One notable drawback of ChatGPT in its application to automating the SR process pertains to its dependence on a pre-existing database. ChatGPT relies solely on the information it was trained on, lacking access to real-time data from the internet. Consequently, the model's knowledge and comprehension are confined to the training data, limiting its ability to incorporate the latest research studies, publications, and emerging evidence. This limitation poses challenges in providing comprehensive and up-to-date information throughout the systematic review process.
5. **Length of prompts:** While ChatGPT has the ability to generate high-quality responses, the length and complexity of the prompts used can impact the accuracy and coherence of the generated text. Our study revealed that longer prompts tended to result in more accurate and relevant responses, but also required more time and effort to

prepare. Conversely, shorter prompts were easier and quicker to generate, but may have led to less accurate or incomplete responses. Hence, balancing the prompt's length and complexity with the generated text's accuracy and relevance is important. Additionally, careful consideration should be given to the prompt formulation process to ensure that the generated responses meet the desired quality standards in the context of the SR process.

6. **Token limitations:** ChatGPT limits the number of tokens that can be processed simultaneously. This means that the length of the input sequence (i.e., prompt plus generated text) is limited and may require multiple iterations or segmentation to generate longer responses. Our study encountered this limitation when attempting to generate longer responses. This limitation can affect the efficiency and effectiveness of the ChatGPT's model for certain tasks, especially in Phase 2, where the filtration occurred by feeding the ChatGPT with some parts from the article.
7. **Memory limitations:** The ChatGPT's ability to recall previous prompts and maintain a coherent and accurate discourse on a specific topic is a crucial consideration, as it can impose constraints that impact its scalability and applicability to certain tasks. Within our study, we encountered restrictions related to memory capacity, wherein ChatGPT occasionally struggled to provide responses that remained focused on the precise topic, leading to deviations or inaccuracies in its understanding of our prompts. This was particularly noticeable when working with large datasets or engaging in multiple iterations, highlighting the potential impact of memory limitations on the model's performance.

#### *4.3. Future Perspectives: Expanding the Potential of ChatGPT in SR*

As technology advances and AI-driven language models such as ChatGPT become more sophisticated, there are exciting opportunities for further development and utilization in the field of SR. The future perspectives of ChatGPT in SR offer potential avenues for enhancing the review process's efficiency, accuracy, and comprehensiveness. By addressing existing challenges and building upon the strengths of ChatGPT, researchers can unlock its full potential in advancing evidence synthesis and knowledge discovery. This subsection explores some of the future perspectives and areas of improvement for ChatGPT in the SR methodology, including:

1. **Conducting the snowballing procedure using ChatGPT:** This approach involves utilizing ChatGPT to search the database using BSTs, applying the first round of filtering based on abstracts, and then collecting remaining articles along with their references (backward) and cited publications (forward). These collected articles would undergo another round of abstract screening before proceeding to the second level of filtering. Automating the snowballing procedure with ChatGPT could streamline the filtration process, making it more efficient and time-saving for researchers.
2. **Developing more sophisticated algorithms to extract information from articles:** Advanced techniques such as entity recognition and topic modeling could be employed to enhance the accuracy and precision of information extraction from articles. These techniques can enable ChatGPT to identify and extract relevant information more effectively, particularly from non-standard formats such as tables, figures, and other complex structures commonly found in scholarly literature.
3. **Improving the interpretability of ChatGPT's output:** Efforts could be made to develop tools or techniques to visualize and comprehend ChatGPT's output. This may involve creating visual representations or graphical displays that aid in understanding the generated summaries or recommendations. Additionally, developing more transparent algorithms, which are easier for researchers to comprehend, can improve the interpretability of ChatGPT's output.
4. **Expanding the scope of input data for ChatGPT:** One potential avenue for enhancing the performance of ChatGPT in conducting SRs could be to explore the model's applicability on data from fields with more relevant articles. This could involve

testing the content analysis capabilities of ChatGPT by inputting a large amount of data and examining the conclusions drawn by the model. Additionally, employing ChatGPT on data from new fields can serve as a valuable means to test the robustness and integrity of the developed methodology in response to different aspects.

5. **Access to Real-Time Data:** The SR process using ChatGPT can benefit from several avenues for improvement. Firstly, ChatGPT can provide accurate, current information regarding articles based on real-time access to databases, such as Scopus and Web of Science. In addition, internet connectivity enhances data retrieval and screening capabilities by allowing users to access a broader range of sources. Secondly, dynamic search strategies enable real-time feedback to be integrated into iterative enhancements. Thirdly, automated citation management and reference management, integration of collaborative platforms, and access to diverse perspectives and global research materials enhance the SR process. However, the success of these enhancements critically hinges on the particular implementation, ethical considerations, and rigorous validation of retrieved information.

Overall, it is essential to embrace the development of AI and use it with caution and supervision in critical domains. While ChatGPT offers significant potential in automating SR processes, it is essential to acknowledge and address its limitations. Strategies for enhancing ChatGPT's performance in conducting SRs should be carefully devised and implemented.

### 5. Ethical Considerations in Utilizing AI-Language Models

The utilization of AI language models such as ChatGPT in scientific writing necessitates careful attention to ethical considerations. Integrating these models raises important questions that require thorough examination and appropriate safeguards. One crucial ethical consideration in utilizing AI language models is the validation, verification, and critical evaluation of AI-generated outputs in order to ensure their accuracy, reliability, and appropriate contextualization within the broader scientific knowledge. In this regard, the involvement of human experts is paramount. Their supervision and expertise play a critical role in aligning the outputs with established standards, identifying and rectifying potential inaccuracies or biases, and providing a comprehensive and accurate interpretation of the AI-generated content. By incorporating human judgment and critical evaluation, researchers uphold responsible practices that enhance the reliability and credibility of the findings derived from AI language models.

Ethical considerations also encompass aspects such as data privacy, informed consent, and bias mitigation strategies. Researchers must adhere to established guidelines and regulations to protect data privacy when utilizing AI language models. This involves handling sensitive or personal information with utmost care and ensuring strict confidentiality to comply with privacy standards. Obtaining informed consent becomes crucial when utilizing data collected from individuals or sources with sensitive information. Moreover, researchers must proactively implement strategies to identify and mitigate biases that may arise from the input data used in the automated SR process, ensuring fair and unbiased outcomes. By conscientiously addressing these ethical considerations, researchers contribute to the cultivation of a responsible and ethical environment for the utilization of AI language models in scientific writing.

### 6. Concluded Remarks and Recommendations

Our study presents a novel methodology for conducting systematic reviews by leveraging the power of ChatGPT. By combining the strengths of human expertise and AI capabilities, we aimed to streamline the traditional SR process and improve its efficiency and accuracy. Our study applied this method to conduct a comprehensive SR on IoT applications in water and wastewater infrastructure management, and our findings highlight the benefits of using ChatGPT in each step of the process. Our study revealed that ChatGPT effectively generates research questions and suggests Boolean research terms, but not appro-

appropriate for article extraction. However, it performs excellently in filtering and categorizing articles and excellently in full-text filtration and information extraction after preparing prompts. Our comprehensive content analysis of the selected publications revealed valuable insights into the current research landscape, highlighting emerging trends, identifying research gaps, and shedding light on future directions in the domains of IoT-based sensing and monitoring, data analytics and visualization, as well as applications and case studies. We evaluated our methodology using quantitative comparisons with traditional review techniques and expert opinions, and the results show that our approach significantly saves time and effort while maintaining high levels of accuracy. Our findings demonstrate the potential of ChatGPT in improving the efficiency and accuracy of SRs, contributing to the advancement of scientific knowledge. In conclusion, there are promising avenues for future research in fully exploring the capabilities of ChatGPT in SRs, investigating its limitations in diverse research contexts, and applying our approach to other fields to further enhance the efficiency and accuracy of SRs. We strongly recommend adopting our proposed framework as a reliable guide for conducting SRs in diverse domains. Our proposed framework, as depicted in Figure 23, provides a robust foundation for automating the SR process, offering adaptability and scalability to accommodate research complexities. By recognizing the strengths and limitations of ChatGPT and taking appropriate measures to enhance its performance, researchers can maximize the benefits of AI in evidence synthesis while ensuring the precision and integrity of SRs in the scientific community.

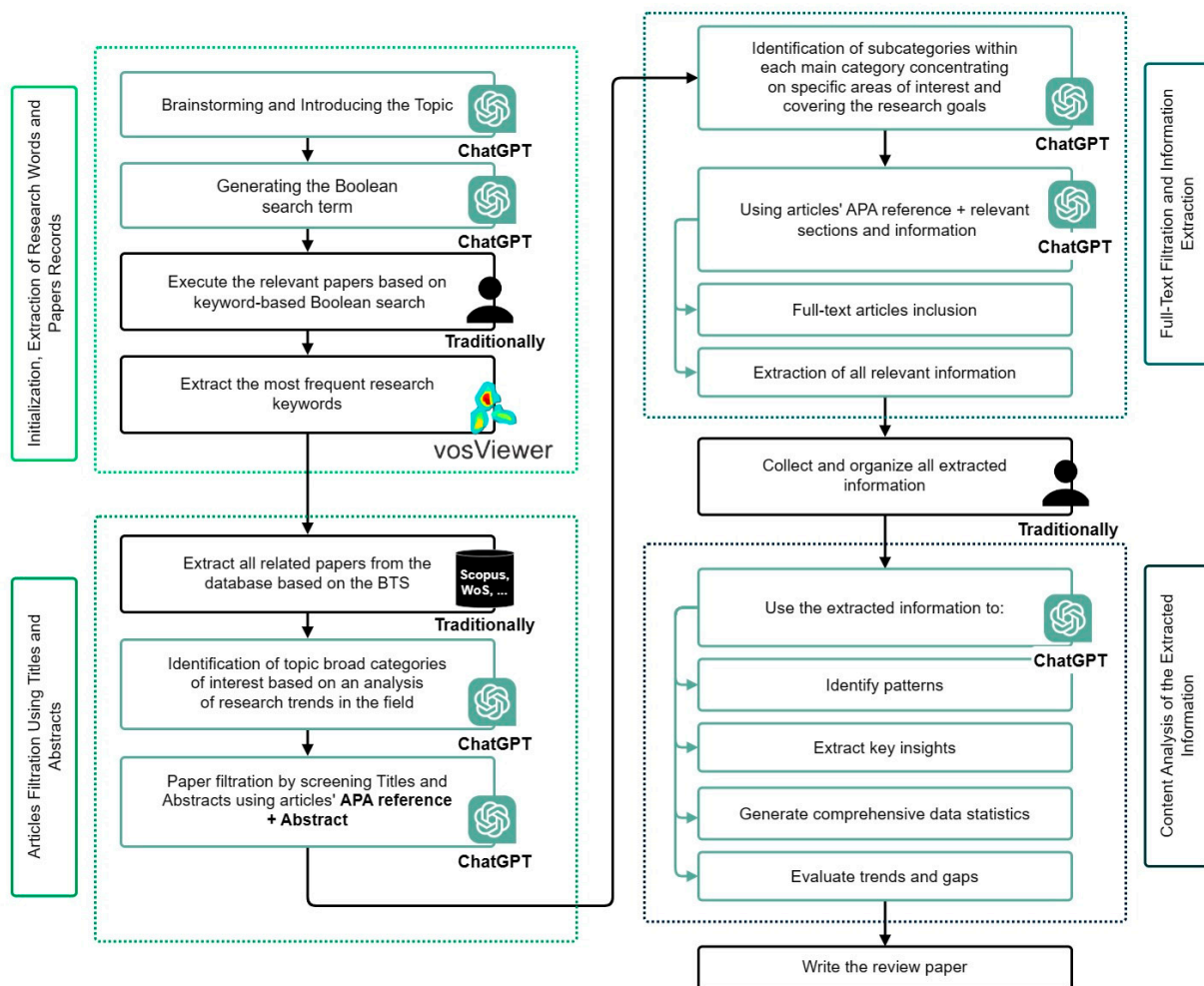


Figure 23. Automated Framework for Streamlining SR Methodology: A Proposed Approach.



**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/systems11070351/s1>, Figure S1: Initialization Process. (a–e) Introducing IoT Technology; Figure S2: Initialization Process. (a–d) Introducing Civil Engineering Infrastructure; Figure S3: Initialization Process. (a–d) Introducing Water and Wastewater Infrastructure; Figure S4: Initialization Process. (a–d) Implementing IoT In Water and Wastewater Infrastructure; Figure S5: Initialization Process. (a–d) Investigating the Systematic Review Capability; Figure S6: ChatGPT’s Utilization of BSTs. (a–e) Extracting Search Keywords; Figure S7: Examples of references from ChatGPT. (a) Extracting related paper based on the Boolean search term. (b) Example of one of the incorrect references. Figure S8: A section of the questionnaire created using Google Forms; Figure S9: Two examples of ChatGPT’s responses in case of irrelevant articles; Figure S10: User prompt and ChatGPT answer to the methods used for data analysis and visualization; Figure S11: User prompt and ChatGPT answer for the benefits of implementing the case studies. Table S1: Unique keywords as extracted from ChatGPT and VosViewer; Table S2: Comparison between ChatGPT and human experts in classification process for Selected 120 articles; Table S3: Categorization of all articles using ChatGPT (APA+Abstract); Table S4: Articles belong to IoT-based water quality monitoring as classified using ChatGPT with explanation; Table S5: Articles belong to IoT-based wastewater infrastructure management as classified using ChatGPT with explanation; Table S6: Articles belong to IoT-based water infrastructure management as classified using ChatGPT with explanation; Table S7: Comparison between answers form ChatGPT and human experts for the 14 questions related to the five subcategorizes for selected 30 articles; Table S8: ChatGPT responses to the 14 questions with Yes/No and the detailed description for the answers. (a) IoT-based water infrastructure management, (b) IoT-based wastewater infrastructure management, and (c) IoT-based water quality monitoring.

**Author Contributions:** Conceptualization, A.A., E.A. and M.E.; methodology, A.A., E.A. and M.E.; validation, A.A., E.A. and M.E.; formal analysis, A.A., E.A. and M.E.; investigation, E.A. and A.E.E.E.; writing—original draft preparation, A.A., E.A. and M.E.; writing—review and editing, E.A., A.E.E.E. and A.A.; visualization, M.E., E.A. and A.A.; supervision, E.A., A.E.E.E. and T.Z.; project administration, A.E.E.E. and T.Z.; funding acquisition, A.E.E.E. and T.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the University Grant Committee of Hong Kong Polytechnic University: [Grant Number Project No. P0036181].

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The Author would like to thank greatly the volunteers who participated in the filtering process.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Paré, G.; Trudel, M.-C.; Jaana, M.; Kitsiou, S. Synthesizing Information Systems Knowledge: A Typology of Literature Reviews. *Inf. Manag.* **2015**, *52*, 183–199. [[CrossRef](#)]
2. Yuan, Y.; Hunt, R.H. Systematic Reviews: The Good, the Bad and the Ugly. *Am. J. Gastroenterol.* **2009**, *104*, 1086–1092. [[CrossRef](#)] [[PubMed](#)]
3. Kitchenham, B. *Procedures for Performing Systematic Reviews*; Keele University: Keele, UK, 2004.
4. Mulrow, C.D. Systematic Reviews: Rationale for Systematic Reviews. *BMJ* **1994**, *309*, 597–599. [[CrossRef](#)] [[PubMed](#)]
5. Needleman, I.G. A Guide to Systematic Reviews. *J. Clin. Periodontol.* **2002**, *29*, 6–9. [[CrossRef](#)]
6. Agbo, C.; Mahmoud, Q.; Eklund, J. Blockchain Technology in Healthcare: A Systematic Review. *Healthcare* **2019**, *7*, 56. [[CrossRef](#)]
7. FitzGerald, C.; Hurst, S. Implicit Bias in Healthcare Professionals: A Systematic Review. *BMC Med. Ethics* **2017**, *18*, 19. [[CrossRef](#)]
8. Milne-Ives, M.; de Cock, C.; Lim, E.; Shehadeh, M.H.; de Pennington, N.; Mole, G.; Normando, E.; Meinert, E. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J. Med. Internet Res.* **2020**, *22*, e20346. [[CrossRef](#)]
9. Abu-Odah, H.; Su, J.; Wang, M.; Lin, S.-Y.; Bayuo, J.; Musa, S.S.; Molassiotis, A. Palliative Care Landscape in the COVID-19 Era: Bibliometric Analysis of Global Research. *Healthcare* **2022**, *10*, 1344. [[CrossRef](#)]
10. Aarseth, W.; Ahola, T.; Aaltonen, K.; Økland, A.; Andersen, B. Project Sustainability Strategies: A Systematic Literature Review. *Int. J. Proj. Manag.* **2017**, *35*, 1071–1083. [[CrossRef](#)]
11. Shaban, I.A.; Eltoukhy, A.E.E.; Zayed, T. Systematic and Scientometric Analyses of Predictors for Modelling Water Pipes Deterioration. *Autom. Constr.* **2023**, *149*, 104710. [[CrossRef](#)]
12. Silva, M. A Systematic Review of Foresight in Project Management Literature. *Procedia Comput. Sci.* **2015**, *64*, 792–799. [[CrossRef](#)]

13. Karam, A.; Eltoukhy, A.E.E.; Shaban, I.A.; Attia, E.-A. A Review of COVID-19-Related Literature on Freight Transport: Impacts, Mitigation Strategies, Recovery Measures, and Future Research Directions. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12287. [[CrossRef](#)] [[PubMed](#)]
14. Araújo, A.G.; Pereira Carneiro, A.M.; Palha, R.P. Sustainable Construction Management: A Systematic Review of the Literature with Meta-Analysis. *J. Clean. Prod.* **2020**, *256*, 120350. [[CrossRef](#)]
15. Hussein, M.; Eltoukhy, A.E.E.; Karam, A.; Shaban, I.A.; Zayed, T. Modelling in Off-Site Construction Supply Chain Management: A Review and Future Directions for Sustainable Modular Integrated Construction. *J. Clean. Prod.* **2021**, *310*, 127503. [[CrossRef](#)]
16. Taiwo, R.; Shaban, I.A.; Zayed, T. Development of Sustainable Water Infrastructure: A Proper Understanding of Water Pipe Failure. *J. Clean. Prod.* **2023**, *398*, 136653. [[CrossRef](#)]
17. Michalski, A.; Głodziński, E.; Böde, K. Lean Construction Management Techniques and BIM Technology—Systematic Literature Review. *Procedia Comput. Sci.* **2022**, *196*, 1036–1043. [[CrossRef](#)]
18. Abdelkader, E.M.; Zayed, T.; Faris, N. Synthesized Evaluation of Reinforced Concrete Bridge Defects, Their Non-Destructive Inspection and Analysis Methods: A Systematic Review and Bibliometric Analysis of the Past Three Decades. *Buildings* **2023**, *13*, 800. [[CrossRef](#)]
19. Elshaboury, N.; Al-Sakkaf, A.; Mohammed Abdelkader, E.; Alfalah, G. Construction and Demolition Waste Management Research: A Science Mapping Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4496. [[CrossRef](#)]
20. Eltoukhy, A.E.E.; Chan, F.T.S.; Chung, S.H. Airline Schedule Planning: A Review and Future Directions. *Ind. Manag. Data Syst.* **2017**, *117*, 1201–1243. [[CrossRef](#)]
21. Hassan, L.K.; Santos, B.F.; Vink, J. Airline Disruption Management: A Literature Review and Practical Challenges. *Comput. Oper. Res.* **2021**, *127*, 105137. [[CrossRef](#)]
22. Aromataris, E.; Riitano, D. Systematic Reviews. *AJN Am. J. Nurs.* **2014**, *114*, 49–56. [[CrossRef](#)] [[PubMed](#)]
23. Meline, T. Selecting Studies for Systemic Review: Inclusion and Exclusion Criteria. *Contemp. Issues Commun. Sci. Disord.* **2006**, *33*, 21–27. [[CrossRef](#)]
24. Wohlin, C. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK, 13–14 May 2014; ACM: New York, NY, USA, 2014; pp. 1–10.
25. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Int. J. Surg.* **2010**, *8*, 336–341. [[CrossRef](#)] [[PubMed](#)]
26. Sarkis-Onofre, R.; Catalá-López, F.; Aromataris, E.; Lockwood, C. How to Properly Use the PRISMA Statement. *Syst. Rev.* **2021**, *10*, 117. [[CrossRef](#)]
27. Aydın, Ö.; Karaarslan, E. OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. *SSRN Electron. J.* **2022**. [[CrossRef](#)]
28. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J. Med. Syst.* **2023**, *47*, 33. [[CrossRef](#)]
29. Vaishya, R.; Misra, A.; Vaish, A. ChatGPT: Is This Version Good for Healthcare and Research? *Diabetes Metab. Syndr. Clin. Res. Rev.* **2023**, *17*, 102744. [[CrossRef](#)]
30. Halaweh, M. ChatGPT in Education: Strategies for Responsible Implementation. *Contemp. Educ. Technol.* **2023**, *15*, ep421. [[CrossRef](#)]
31. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *PLOS Digit. Health* **2023**, *2*, e0000198. [[CrossRef](#)]
32. Zhai, X. ChatGPT for Next Generation Science Learning. *XRDS Crossroads ACM Mag. Stud.* **2023**, *29*, 42–46. [[CrossRef](#)]
33. Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education? *J. Appl. Learn. Teach.* **2023**, *6*, 342–362. [[CrossRef](#)]
34. Prieto, S.A.; Mengiste, E.T.; García de Soto, B. Investigating the Use of ChatGPT for the Scheduling of Construction Projects. *Buildings* **2023**, *13*, 857. [[CrossRef](#)]
35. You, H.; Ye, Y.; Zhou, T.; Zhu, Q.; Du, J. Robot-Enabled Construction Assembly with Automated Sequence Planning Based on ChatGPT: RoboGPT. *arXiv* **2023**, arXiv:2304.11018.
36. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [[CrossRef](#)] [[PubMed](#)]
37. Salvagno, M.; Taccone, F.S.; Gerli, A.G. Can Artificial Intelligence Help for Scientific Writing? *Crit. Care* **2023**, *27*, 75. [[CrossRef](#)] [[PubMed](#)]
38. Zheng, H.; Zhan, H. ChatGPT in Scientific Writing: A Cautionary Tale. *Am. J. Med.* **2023**. [[CrossRef](#)]
39. Dergaa, I.; Chamari, K.; Zmijewski, P.; Ben Saad, H. From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing. *Biol. Sport* **2023**, *40*, 615–622. [[CrossRef](#)]
40. Khosravi, H.; Shafie, M.R.; Hajiabadi, M.; Raihan, A.S.; Ahmed, I. Chatbots and ChatGPT: A Bibliometric Analysis and Systematic Review of Publications in Web of Science and Scopus Databases. *arXiv* **2023**, arXiv:2304.05436.
41. Lecler, A.; Duron, L.; Soyer, P. Revolutionizing Radiology with GPT-Based Models: Current Applications, Future Possibilities and Limitations of ChatGPT. *Diagn. Interv. Imaging* **2023**, *104*, 269–274. [[CrossRef](#)]

42. Hosseini, M.; Horbach, S.P.J.M. Fighting Reviewer Fatigue or Amplifying Bias? Considerations and Recommendations for Use of ChatGPT and Other Large Language Models in Scholarly Peer Review. *Res. Integr. Peer. Rev.* **2023**, *8*, 4. [[CrossRef](#)]
43. Fang, T.; Yang, S.; Lan, K.; Wong, D.F.; Hu, J.; Chao, L.S.; Zhang, Y. Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. *arXiv* **2023**, arXiv:2304.01746.
44. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [[CrossRef](#)] [[PubMed](#)]
45. 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](#)]
46. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
47. Zeng, G. On the Confusion Matrix in Credit Scoring and Its Analytical Properties. *Commun. Stat. Theory Methods* **2020**, *49*, 2080–2093. [[CrossRef](#)]
48. Jan, F.; Min-Allah, N.; Saeed, S.; Iqbal, S.Z.; Ahmed, R. IoT-Based Solutions to Monitor Water Level, Leakage, and Motor Control for Smart Water Tanks. *Water* **2022**, *14*, 309. [[CrossRef](#)]
49. Singh, M.; Ahmed, S. IoT Based Smart Water Management Systems: A Systematic Review. *Mater. Today Proc.* **2021**, *46*, 5211–5218. [[CrossRef](#)]
50. Zulkifli, C.Z.; Garfan, S.; Talal, M.; Alamoodi, A.H.; Alamleh, A.; Ahmaro, I.Y.Y.; Sulaiman, S.; Ibrahim, A.B.; Zaidan, B.B.; Ismail, A.R.; et al. IoT-Based Water Monitoring Systems: A Systematic Review. *Water* **2022**, *14*, 3621. [[CrossRef](#)]
51. Alshami, A.; Elsayed, M.; Mohandes, S.R.; Kineber, A.F.; Zayed, T.; Alyanbaawi, A.; Hamed, M.M. Performance Assessment of Sewer Networks under Different Blockage Situations Using Internet-of-Things-Based Technologies. *Sustainability* **2022**, *14*, 14036. [[CrossRef](#)]
52. Haluza, D.; Jungwirth, D. Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3. *Systems* **2023**, *11*, 120. [[CrossRef](#)]
53. Yang, X.; Li, Y.; Zhang, X.; Chen, H.; Cheng, W. Exploring the Limits of ChatGPT for Query or Aspect-Based Text Summarization. *arXiv* **2023**, arXiv:2302.08081.

**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.