SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2026
ORGANIZATION AND MANAGEMENT SERIES NO. 246

EVALUATING SELECTED AI TOOLS FOR LITERATURE REVIEWS: A CASE STUDY IN LOCAL ADMINISTRATION CYBERSECURITY

Maciej SZMIT[1*], Dominika LISIAK-FELICKA[2]

[1] Department of Computer Science, Faculty of Management, University of Lodz; maciej.szmit@uni.lodz.pl, ORCID: 0000-0002-6115-9213
[2] Department of Economic and Medical Informatics, Faculty of Economics and Sociology, University of Lodz; dominika.lisiak@uni.lodz.pl, ORCID: 0000-0001-8451-4268
* Correspondence author

Purpose: The objectives of this article are to identify scientific articles on the cybersecurity of local administration and to assess how useful the aid of two Artificial Intelligence tools: SciSpace.com (Advanced Plan) and Consensus.app (Pro tier) can be in this type of research.
Design/methodology/approach: The method of conducting the screening and verifying its results based on manual screening partially assisted by LLMs is described in detail in the article. The PRISMA 2020 protocol was used for a part of the preliminary screening.
Findings: The study indicates that both tools have limited usefulness for screening based on abstract content and better for full text document screening. When using the Scopus database, the study showed minimal utility of the PRISMA protocol implementation within the SciSpace tool, particularly when weighed against its usage cost.
Research limitations/implications: Due to both: the nature of LLM tools, whose immanent feature is indeterminism, often leading to the lack of reproducibility of obtained results, and the business model, where the tools are continually developed and their behavior changes over time, the obtained results are not general or universal.
Practical implications: Local administration cybersecurity is a very specific field of knowledge and practice. This type of literature review can be helpful to those working in said field in different countries. and the identification of the tested AI tools' limitations – particularly the relatively low quality of screening based solely on article abstracts – may allow for more effective use of similar tools.
Originality/value: The results provide knowledge about publications in the field of cybersecurity in local government administration and include calculations of the basic statistics (accuracy, sensitivity, and specificity) of the discussed tools in the examined case. To our knowledge, this is the most current and comprehensive literature review on this topic to date, while the calculations constitute a contributory study.
Keywords: information security management systems, local administration cybersecurity.
Category of the paper: research paper.

1. Introduction

The rapid expansion of scientific literature has made traditional systematic literature reviews (SLRs) increasingly time-consuming. In response, artificial intelligence (AI) has emerged as a modern tool, offering automation across multiple stages of the SLR process. As in a number of other applications, AI can be a promising tool for this process but - also similarly to other AI applications - serious doubts arise regarding the admissibility of its use, from the threat of losing creative originality of the researcher's work (see e.g. Lund et al., 2023, Hosseini et al., 2023; Ateriya et al., 2025; Cheng et al., 2025), to the issue of the quality of the obtained results, including their reliability, especially in the context of the tendency of some AI tools - mainly Large Language Models (LLM) - to "hallucinate", which can manifest itself e.g. in the production of non-existent bibliographic entries (see e.g. Walters et al., 2023: Chelli et al., 2024; Tosi, 2025).

Most existing studies examine AI-assisted screening in broad or well-established research areas, often in medical or interdisciplinary contexts. Less attention has been given to narrowly defined and semantically heterogeneous domains, where relevance cannot be determined solely by keyword matching but contextual and institutional interpretation is also required.

Cybersecurity in local administration is one such field. Both "local administration" and "cybersecurity" are conceptually broad and interpreted differently across countries and disciplines. This diversity complicates query construction and screening criteria and raises questions about how reliably AI tools can identify relevant publications.

Against this background, the study addresses the research problem: to what extent can selected AI tools help with literature screening in such a field? The research question is: how accurate and useful are SciSpace (Advanced Plan) and Consensus (Pro tier) in identifying publications devoted to cybersecurity in local administration?

This article has two objectives: first, to identify all relevant publications across major academic databases; and second, to evaluate the classification performance and utility of the AI tools in question in abstract-based and full-text screening procedures.

2. Literature review

AI can help with SLRs in several ways: early use covered automating repetitive tasks such as literature search and screening using artificial neural networks (ANN) and text mining (de la Torre-López et al., 2023). More recent developments leverage advanced machine learning (ML), natural language processing (NLP), and LLMs to support not only screening but also data extraction, synthesis, reporting and even automated literature reviews, enhancing

scalability and uncovering interdisciplinary connections (Malik, Terzidis, 2025; Bolanos et al., 2024; Li et al., 2025; Susnjak et al., 2025; Ofori-Boateng et al., 2024). Given the known weaknesses of AI, several frameworks have been proposed that use human supervision over artificial intelligence, emphasizing transparency and reliability (Lee et al., 2025; Le Dinh et al., 2025). These frameworks typically use a cyclical approach, including human testing and validation of AI tools to reduce bias and ensure methodological soundness (called "human-in-the-loop"). A human-assisted approach remains crucial, enabling real-time adjustments to inclusion criteria and continuous quality assurance (van Dijk et al., 2023).

There are a lot of articles devoted to issues related to evaluating the usefulness of this type of tools (Bernard et al., 2025; Pinzolits, 2023; Fuller-Tyszkiewicz et al., 2025; Vallamchetla et al., 2025; Abogunrin et al., 2025; Tosi, 2025; Mogoale et al., 2025), especially reduction of workload, precision, accuracy, as well as developing methodologies and tools intended to assist researchers in conducting SLRs (van de Schoot et al., 2021). Arhin in the article (Arhin et al., 2025) declares that a thorough search performed across databases including Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar using keywords such as "AI and academic writing", "AI in literature reviews", and "machine-assisted literature review" gave 1153 peer-reviewed studies produced from 2020 to 2024 (after eliminating non-empirical publications and those irrelevant to AI in academic literature review procedures). At the time of writing this (December 2025) there are 2340 findings published in 2025 indexed in Web of Science alone, one could therefore venture that the use of AI tools in this context is currently at a stage named Peak of Inflated Expectations in Gartner's Hype Cycle model (see e.g. Czerwonka, Podgórski, 2025) perhaps one should even call it overhyped. However taking into account that there are many AI solutions on the market and new ones and new versions and new extensions of previously available ones appearing and moreover, different sciences have different methodological conditions, and the terminology used within them may be amenable to analysis in different degrees (taking into account the methods used by particular tools) it seems useful to conduct similar research for different issues.

In existing literature, evaluations of AI-assisted screening have typically been conducted in broad or well-established research areas, often in medical or interdisciplinary contexts (van de Schoot et al., 2021; Ofori-Boateng et al., 2024). In such domains, inclusion criteria are usually well structured and terminology is relatively standardized, which may facilitate automated classification. However, several authors underline that the effective use of AI in systematic reviews requires careful validation, human supervision, and methodological transparency (van Dijk et al., 2023; Lee et al., 2025).

Less attention has been devoted to narrowly defined and semantically heterogeneous research areas, where relevance is not based on keyword matching. In such areas, classification may depend on institutional or contextual interpretation or regulatory recommendations concerning particular terms, increasing the risk of classification errors. Moreover, generative AI systems may exhibit a tendency to interpret prompts too literally and sensitivity to contextual

framing, which further endangers the stability and reproducibility of screening outcomes (Tosi, 2025).

Cybersecurity in local administration constitutes such a domain. It combines elements of public administration and GRC (Governance, Risk Management and Compliance), especially concerning information security management. Overlapping term ranges – for example between "local government", "municipality", and "regional authority", and between "cybersecurity", "information security", and "data protection" – increases the methodological complexity of query construction and screening procedures. This condition makes for a suitable test case for examining how AI tools perform when semantic nuance plays a significant role.

Despite the growing number of publications on AI-supported systematic reviews, there isn't much empirical research assessing tool performance in narrowly defined domains and comparing abstract-based and full-text screening within such contexts. This study aims to contribute to this discussion by providing a domain-specific evaluation of selected AI tools and by quantifying their classification characteristics.

3. Research methods

3.1. Databases selection and query construction

One of the most important issues concerning local administration cybersecurity is that neither "local administration" nor "cybersecurity" have an unambiguous meaning nor syntax. Instead of "cybersecurity", it is possible to use such semi-equivalent terms such as "information security", "data security", "cybersafety" etc. Additionally different (conjunctive and separated) spelling the "cyber-" prefix are used in the literature. One can find a few synonymous terms for "local administration" (e.g. "local government", "municipal administration", "local authority" etc.) also.

Taking the above into account, the queries we constructed for each database were quite complex, as they involved searching for conjunctions of two long phrases (in at least one of the most important fields i.e. title, abstract, keywords): one related to local administration and one related to cybersecurity. Formally, we were looking for conjunctions of complex alternatives. A sample query directed at the IEEEXplore database is shown in the fugure below.

```
((((("Document Title":"local government") OR ("Document
Title":municipality)OR ("Document Title":"local administration")OR
("Document Title":"municipal administration")OR ("Document Title":"city
government")OR ("Document Title":"regional government")OR ("Document
Title":"local authority"))AND (("Document Title":"cybersecurity")OR
("Document Title":"cyber security")OR ("Document Title":"cybersafety")OR
("Document Title":"cyber safety")OR ("Document Title":"cyberthreat")OR
("Document Title":"cyber threat")OR ("Document Title":"cyber risk")OR
("Document Title":"information security")OR ("Document Title":"data
security")OR ("Document Title":"security of data")OR ("Document Title":"IT
security")OR ("Document Title":"cyber protection")OR ("Document
Title":"security awareness")OR ("Document Title":"network security")OR
("Document Title":"computer security")OR ("Document Title":"cyber
resilience")OR ("Document Title":"ICT security")OR ("Document Title":"cyber
defence")OR ("Document Title":"cyberdefence")OR ("Document Title":"GDPR")OR
("Document Title":"data protection"))) OR (((„Abstract”:"local government")
OR („Abstract”:municipality)OR („Abstract”:"local administration")OR
(„Abstract”:"municipal administration")OR („Abstract”:"city government")OR
(„Abstract”:"regional government")OR („Abstract”:"local
authority"))AND((„Abstract”:"cybersecurity")OR („Abstract”:"cyber
security")OR („Abstract”:"cybersafety")OR („Abstract”:"cyber safety")OR
(„Abstract”:"cyberthreat")OR („Abstract”:"cyber threat")OR
(„Abstract”:"cyber risk")OR („Abstract”:"information security")OR
(„Abstract”:"data security")OR („Abstract”:"security of data")OR
(„Abstract”:"IT security")OR („Abstract”:"cyber protection")OR
(„Abstract”:"security awareness")OR („Abstract”:"network security")OR
(„Abstract”:"computer security")OR („Abstract”:"cyber resilience")OR
(„Abstract”:"ICT security")OR („Abstract”:"cyber defence")OR
(„Abstract”:"cyberdefence")OR („Abstract”:"GDPR")OR („Abstract”:"data
protection")))OR(((„Author Keywords”:"local government") OR („Author
Keywords”:municipality)OR („Author Keywords”:"local administration")OR
(„Author Keywords”:"municipal administration")OR („Author Keywords”:"city
government")OR („Author Keywords”:"regional government")OR („Author
Keywords”:"local authority"))AND ((„Author Keywords”:"cybersecurity")OR
(„Author Keywords”:"cyber security")OR („Author Keywords”:"cybersafety")OR
(„Author Keywords”:"cyber safety")OR („Author Keywords”:"cyberthreat")OR
(„Author Keywords”:"cyber threat")OR („Author Keywords”:"cyber risk")OR
(„Author Keywords”:"information security")OR („Author Keywords”:"data
security")OR („Author Keywords”:"security of data")OR („Author
Keywords”:"IT security")OR („Author Keywords”:"cyber protection")OR
(„Author Keywords”:"security awareness")OR („Author Keywords”:"network
security")OR („Author Keywords”:"computer security")OR („Author
Keywords”:"cyber resilience")OR („Author Keywords”:"ICT security")OR
(„Author Keywords”:"cyber defence")OR („Author Keywords”:"cyberdefence")OR
(„Author Keywords”:"GDPR")OR („Author Keywords”:"data protection"))))
```

Figure 1. IEEEXplore query.

Source: Own survey.

3.2. Screening methodology

The SciSpace agent called "Systematic Literature Review" offers systematic literature review using PRISMA 2020 Methodology (Page et al., 2021). We used it to perform screening of Scopus database search results. Taking into account that the cost of the "advanced" subscription (and additional credits because starting number of SciSpace credits is not sufficient to conduct more than one study using advanced agents) is relatively high and automating such screening in the case of high-quality literature databases would only slightly reduce the workload of the process compared to manual review (only 1 record excluded and 8 flagged for manual review, which constitutes just over one percent of the initial number of records), we decided not to use the PRISMA protocol for the rest of the examined databases and instead construct AI prompts in a way that allows eliminate abstracts not related to the relevant subject.

For the screening procedure, we decided to use the algorithm shown in Figure 2.

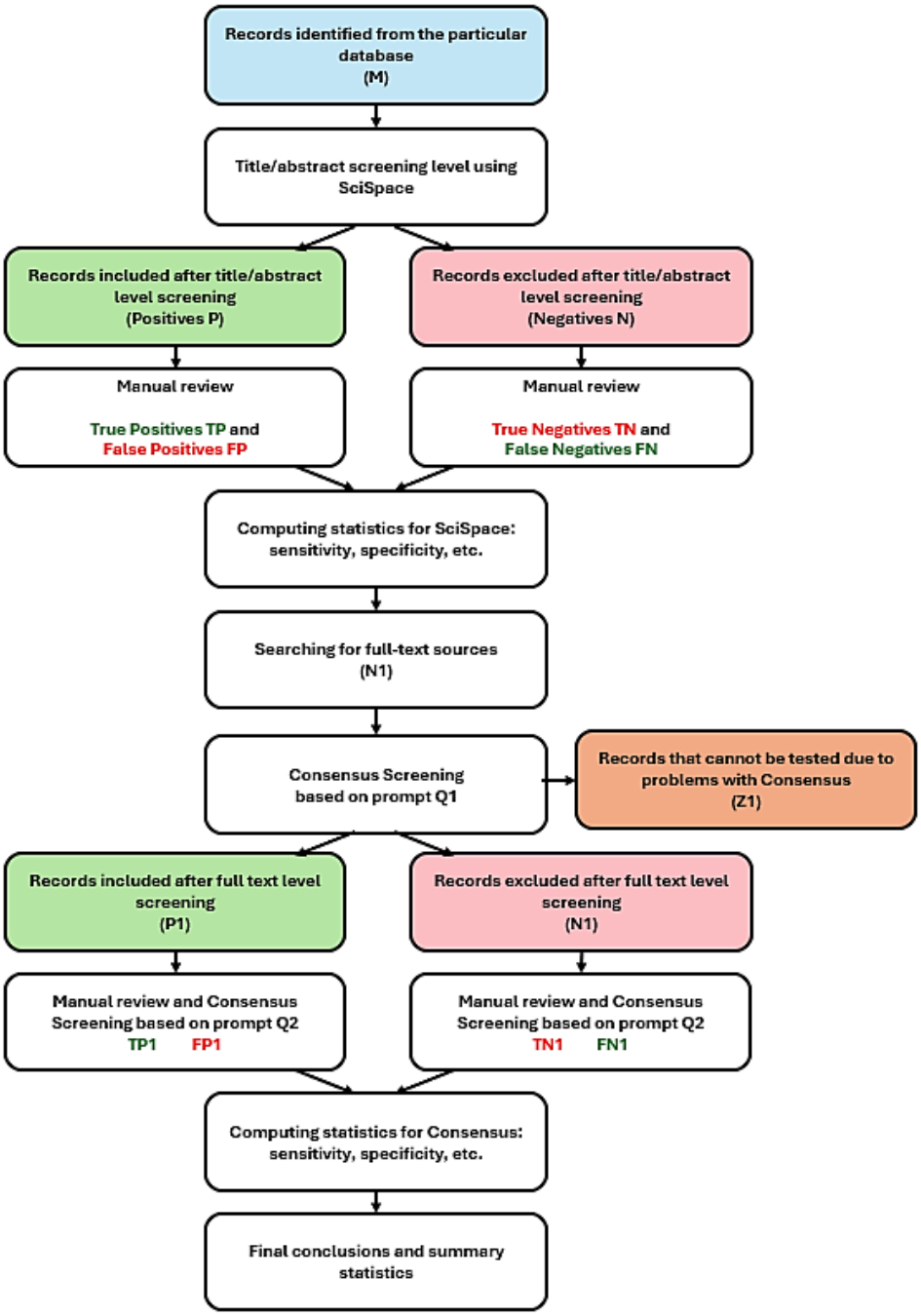

Figure 2. Procedure algorithm.

Source: Own survey using SciSpace.com and Consensus.

The initial screening (using the SciSpace tool) only accounted for abstracts, titles, and keywords, whereas the screenings utilizing the Consensus tool accounted for the full texts of articles (and therefore only for those articles that are available in full-text form). "Records cannot be tested due to problems with Consensus (Z1)" indicates a situation where the process of importing files containing the publication into the tool has failed. We asked the tool's authors why certain texts could not be imported into Consensus. In response, they explained that Consensus relied on standard journal metadata – such as journal name, publication year, DOI, and volume/issue – and that book chapters, despite being valid scholarly

works, were more difficult to index and upload automatically. Consequently, the system was unable to process such texts.

For all three screenings, we constructed prompts according to the schemas shown in the figures below.

```
You got file "xxx.CSV" with bibliographic data. I need two CVS: 1. with list
of all of text are not devoted to local government (including municipal, city
etc.) cybersecurity or information security 2. with list of all other texts
```

Figure 3. SciSpace prompt schema.
Source: Own survey.

```
Is this paper devoted to cybersecurity in local administration (Yes/No)?
```

Figure 4. Consensus prompt Q1 schema.
Source: Own survey.

```
Summarize the article in Polish. What are the main conclusions? Does the
article concern information security in local or municipal government
administration? Does it contain any guidelines regarding the organization or
management of information security?
```

Figure 5. Consensus prompt Q2 schema translation.
Source: Own survey.

We decided to pose two questions to the Consensus tool: one aimed at generating a binary (yes/no) answer, and the other at providing a more in-depth discussion and explanation of the decision made. We anticipated that the requirement to give a concise answer could lead to situations where the tool's made decision might be considered incorrect from a human perspective. Therefore, the second question (along with human control) served to verify its accuracy.

4. Results

4.1. Initial queries results

The number of records identified in each database, as well as the number of records remaining after manual and consensus screening (Q2), are shown in table 1.

Table 1.
Initial queries results

Database	Number of records	Number of full text sources	Final records after Q2 screening
Scopus	692	310	137
Web of Science	215	144	61
IEEEXplore	56	18	13
ACM Digital Library	14	14	7

Source: Own survey.

Some publications are indexed simultaneously in several databases. Figure 6 and table 2 illustrate this situation by showing the number of items in each set and their intersections. The number of unique findings after Q2 screening is 169, while the number of full text sources available is 218.

Table 2.
Cardinality of the intersection of full-text sets

Database	Number of records
Scopus AND Web of Science	36
Scopus AND WoS AND IEEExplore	1
Scopus AND WoS AND ACM Digital Library	4
Web of Science AND ACM Digital Library	1

Source: Own survey.

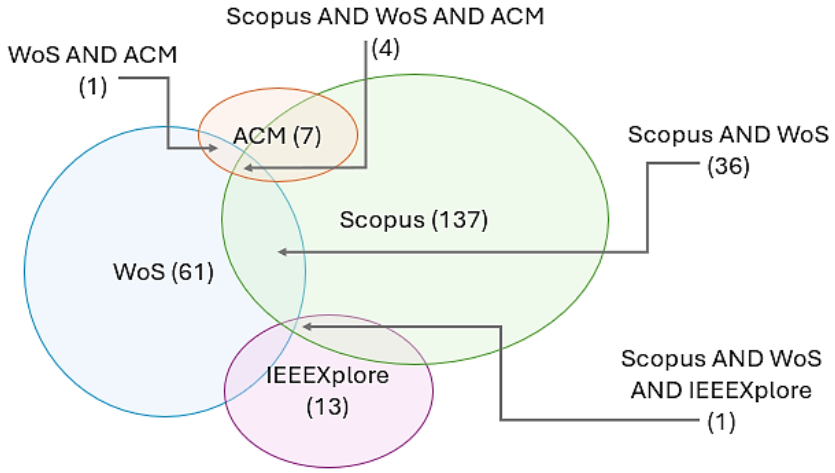

Figure 6. Cardinalities of full-text sets and their intersections.
Source: Own survey.

4.2. SciSpace aided use of PRISMA 2020 on Scopus' search results

Results of having the SciSpace agent "Systematic Literature Review" use the PRISMA 2020 protocol on Scopus Database search results are shown in Figure 7.

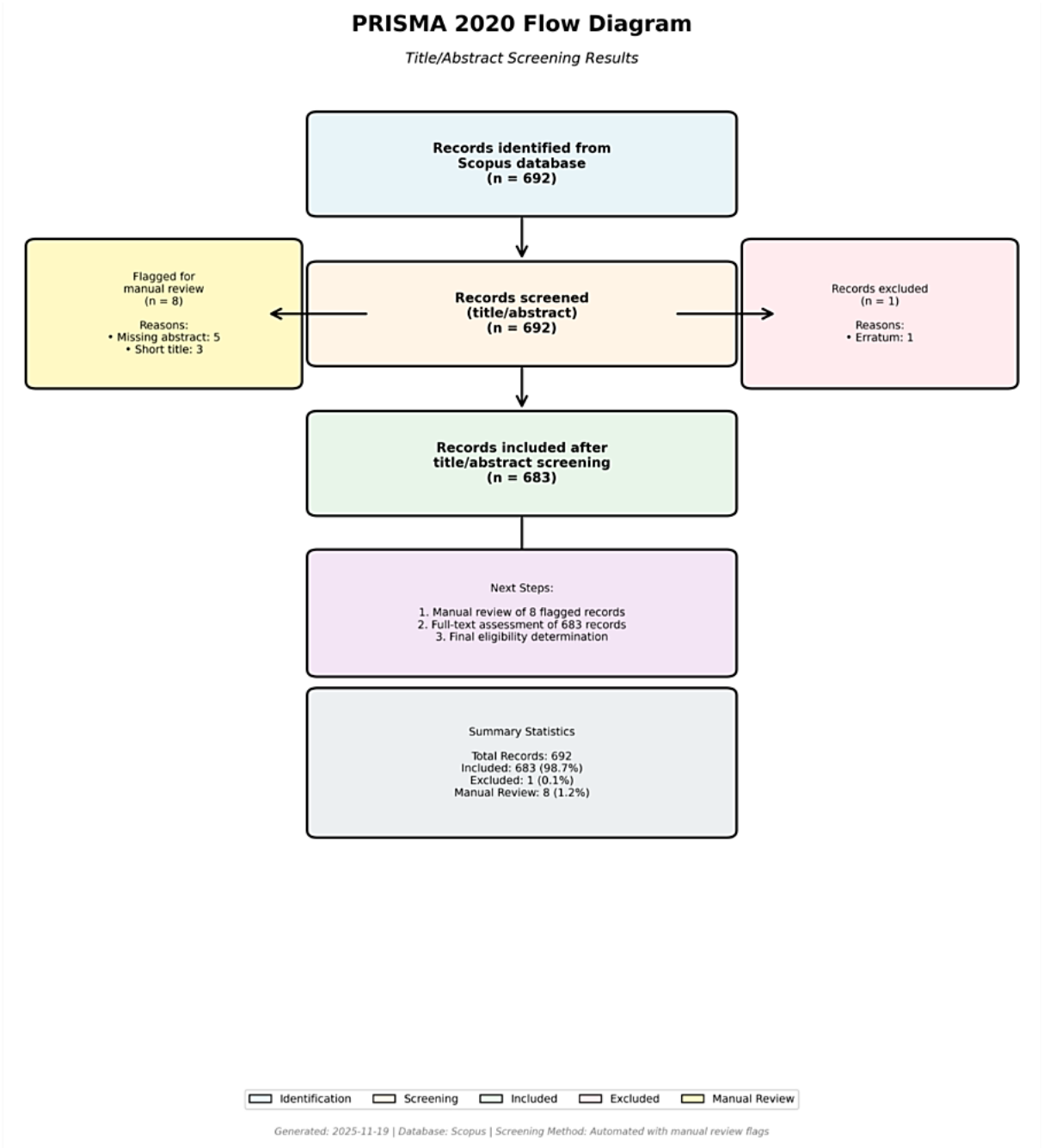

Figure 7. PRISMA report created by SciSpace "Systematic Literature Review" agent on Scopus search results.

Source: Own survey using SciSpace.com.

As mentioned above, due to an insufficient cost-benefit ratio, the use of this protocol for subsequent databases was abandoned in further research.

The study doesn't include a formal time or cost-per-record analysis for the PRISMA screening conducted using SciSpace. Based on practical experience, manual screening of several hundred records (e.g., 692 in Scopus) typically requires several working days, whereas the automated PRISMA procedure was completed minutes after data upload but the cost is tens of dollars.

However, only a very low number of records were automatically excluded or flagged, and most still required manual verification. Considering the subscription and additional credit costs, the cost–benefit ratio the decision to discontinue PRISMA screening for the remaining databases was made.

4.3. Screening statistics

For this article's purpose, among all possible classifier characteristics, the most significant are Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate) and Precision – formulas (1), (2), (3) and (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where:
TP is the number of True Positives,
TN is the number of True Negatives,
FP is the number of False Positives,
FN is the number of False Negatives.

Statistics calculated for each individual database and both tools are shown in Table 4. All classification metrics presented in Table 4 were calculated directly from the TP, FP, TN and FN values reported in Table 3.

Table 3.
Confusion matrix values for each database and tool

Database	SciSpace		Consensus	
Scopus	TP	388	TP1	137
	FP	199	FP1	8
	TN	99	TN1	145
	FN	6	FN1	5
WoS	TP	128	TP1	61
	FP	81	FP1	5
	TN	5	TN1	66
	FN	1	FN1	5
IEEE	TP	12	TP1	13
	FP	5	FP1	0
	TN	38	TN1	5
	FN	1	FN1	0

Cont. table 3.

ACM	TP	7	TP1	7
	FP	7	FP1	0
	TN	0	TN1	7
	FN	0	FN1	0

Source: Own survey.

Table 4.
AI tools statistics

	Accuracy	**Sensitivity**	**Specificity**	**Precision**
SciSpace abstracts screening	Scopus			
	0.704	0.985	0.332	0.661
	Web of Science			
	0.619	0.992	0.058	0.612
	IEEE			
	0.893	0.923	0.887	0.706
	ACM Digital Library			
	0.500	1.000	0.000	0.500
Consensus full-text screening (Q2)	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
	Scopus			
	0.956	0.965	0.948	0.945
	Web of Science			
	0.927	0.924	0.930	0.924
	IEEE			
	1.000	1.000	1.000	1.000
	ACM Digital Library			
	1.000	1.000	1.000	1.000

Source: Own survey.

The study doesn't include a controlled experimental comparison of the time required formanual full-text screening versus AI-assisted screening using Consensus.

Based on practical experience, careful manual evaluation of a single full-text scientific article may require from several minutes to almost an hour, depending on length and complexity. Screening over 200 full-text documents would therefore realistically require several days of focused work.

In contrast, once the documents were successfully imported into the Consensus system, initial classification and summarization were generated within minutes. The human researcher's role was limited primarily to verification of the AI-generated classification. While this observation strongly suggests a substantial productivity gain, it should be emphasized that this conclusion is qualitative and experience-based rather than derived from a formally measured time-efficiency experiment. Future research should include controlled timing and costingto quantify the productivity improvement.

5. Discussions

As with other AI tools, it seems crucial to formulate precise prompts. In several cases, asking whether an article is “devoted” to a given topic yielded different results than asking whether it is “concerning” that topic. Additionally, the examined tools appeared to display a tendency toward literal interpretation of prompts, which resulted in arbitrary classification that rejected texts not devoted entirely and exclusively to the given subject (local administration cybersecurity), even when that subject appeared within them as one of the topics (including cases where it was of relatively substantial importance and scope). However, even when trying to force a clear answer, the question (with “Yes/No" clause) sometimes Consensus gave an ambiguous answer (such as "Yes, the paper is devoted to cybersecurity, but not specifically to local public administration", which is not particularly helpful for someone trying to use an AI tool to extremely simplify text screening). It should be remembered that LLMs such as generative pre-trained transformers (GPT) do not understand the processed text but rely on the knowledge acquired during the training process, so in the case of uncommon and complex language structures (both: the analysed texts as well as the questions asked) they may provide inaccurate or vague answers.

Likewise, when it comes to the analysis of complex and multi-faceted texts (as scientific articles sometimes are), it would be worth considering abandoning the enforcement of unequivocal (“Yes/No”) responses and allowing, perhaps, intermediate answers (“partially”). Admittedly, as is always the case with classifiers, a reduction in type I errors will cause an increase in type II errors (and vice versa); nevertheless, in the context of screening scientific articles, the risk of rejecting relevant texts is a greater concern than a slightly increased workload at the final stage of reviewing articles that have successfully passed the screening process (at least when one is not dealing with a large number of texts).

It’s also important to remember that even the very concepts of cybersecurity, local government data security etc. can be understood differently. The scope of local government activities and the technologies they employ may differ in different countries, so issues that don’t strictly fall under the umbrella of local government cybersecurity in one country (such as medical data privacy, information security in cyber-physical systems etc.) may, in some cases, be a very important aspect of such administration’s operations in another country.

These observations are consistent with existing research indicating that AI-assisted screening may achieve high sensitivity during abstract-only screening, while still requiring substantial human validation (van de Schoot et al., 2021; van Dijk et al., 2023). AI-based tools can accelerate the screening of large volumes of scientific literature; however, they should be supervised by a human to ensure methodological reliability (de la Torre-López et al., 2023; Arhin et al., 2025; Ofori-Boateng et al., 2024).

In the present study, abstract-based screening demonstrated high sensitivity but relatively low specificity in several databases, which reduced the risk of excluding relevant publications but did not significantly decrease the workload associated with false positives. Similar observations have been reported in studies analysing the use of AI research assistants in systematic review frameworks, where AI tools helped identify relevant publications but still required careful human verification (Bernard et al., 2025).

In contrast, the very high values of both: sensitivity and specificity observed in full-text screening using Consensus suggest that access to richer contextual information may partially explain the better classification performance.

These findings support the view that AI tools should be treated as methodological support instruments rather than replacements for expert judgement, and that their effectiveness depends on the type of screening and the characteristics of the research domain.

6. Limitations and Future Research

First, the research did not include a controlled experimental measurement of time efficiency or a formal cost-per-record analysis. Although practical experience suggests substantial differences between manual and AI-assisted screening, the conclusions regarding productivity gains remain qualitative rather than based on systematic timing.

The evaluation was also limited to two specific AI tools: SciSpace (Advanced Plan) and Consensus (Pro tier) and to their functionality at the time of the study. Given the frequent updates and rapid development of AI models, tool performance may change over time. Therefore, the results should not be interpreted as universal or permanent signifiers of these tools' quality.

In addition, the analysis was conducted within a single, narrowly defined research domain which limits generality of results. Tool performance in other domains – particularly those with more standardized terminology – may differ significantly.

Furthermore, the screening procedure relied on specific prompt formulations, alternative formulations might have produced different classification outcomes.

Future research should therefore include formal timing and costing comparisons of manual and AI-assisted screening. It would also be valuable to replicate the study across different research domains and to compare a broader range of AI tools. It also may be interesting use multi-class classification approaches (e.g., allowing "partially relevant" categories).

7. Conclusions

The statistics for abstract screening clearly indicate that the adopted procedure is characterized by high sensitivity but relatively low specificity, and thus leaves significant workload in verifying texts classified as relevant. For unknown reasons, the results are better in the case of IEEEXplore. It may be due to the different characteristics of the database itself or changes that might have occurred in the tools between uses. In comparison, the AI-assisted screening of full texts yields very good results in terms of both sensitivity and specificity. In practice, this means that for large text databases, AI can be a useful assistant in the screening process, provided it works on the full texts of articles, not just their abstracts, titles, and keywords.

Acknowledgements

The subscription costs of both AI tools were covered by the authors' private funds. Authors are not affiliated with any AI software developers.

References

1. Abogunrin, S., Muir, J.M., Zerbini, C., Sarri, G. (2025). How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. *Frontiers in Pharmacology*, *16*. https://doi.org/10.3389/fphar.2025.1454245
2. Ateriya, N., Sonwani, N.S., Thakur, K.S., Kumar, A., Verma, S.K. (2025). Exploring the ethical landscape of AI in academic writing. *Egyptian Journal of Forensic Sciences*, *15*(1), 36. https://doi.org/10.1186/s41935-025-00453-1
3. Bernard, N., Sagawa Jr, Y., Bier, N., Lihoreau, T., Pazart, L., Tannou, T. (2025). Using artificial intelligence for systematic review: the example of elicit. *BMC Medical Research Methodology*, *25*(1), 75. https://doi.org/10.1186/s12874-025-02528-y
4. Bolanos, F., Salatino, A., Osborne, F., Motta, E. (2024). Artificial Intelligence for Literature Reviews: Opportunities and Challenges. *ArXiv*, abs/2402.08565. https://doi.org/10.48550/arxiv.2402.08565
5. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference

Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, *26*, e53164. https://doi.org/10.2196/53164

6. Cheng, A., Calhoun, A., Reedy, G. (2025). Artificial intelligence-assisted academic writing: recommendations for ethical use. *Advances in Simulation*, *10*(1), 22. https://doi.org/10.1186/s41077-025-00350-6
7. Czerwonka, P., Podgórski, G. (2025). *Technologie transformacji cyfrowej przedsiębiorstw produkcyjnych*. Wydawnictwo Uniwersytetu Łódzkiego. https://doi.org/10.18778/8331-912-4
8. de la Torre-López, J., Ramírez, A., Romero, J.R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, *105*(10), 2171-2194. https://doi.org/10.1007/s00607-023-01181-x
9. Edmond Osei Arhin, Festus Cobena Ainoo, Kwame Amponsah (2025). Integration of Artificial Intelligence (AI) Into Academic Literature Reviews: An overview. *World Journal of Advanced Science and Technology*, *7*(1), 006-023. https://doi.org/10.53346/wjast.2025.7.1.0010
10. Fuller-Tyszkiewicz, M., Jones, A., Vasa, R., Macdonald, J.A., Deane, C., Samuel, D., Evans-Whipp, T., Olsson, C.A. (2025). Artificial Intelligence Software to Accelerate Screening for Living Systematic Reviews. *Clinical Child and Family Psychology Review*. https://doi.org/10.1007/s10567-025-00519-5
11. Hosseini, M., Resnik, D.B., Holmes, K. (2023). The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics*, *19*(4), 449-465. https://doi.org/10.1177/17470161231180449
12. Le Dinh, T., Le, T.D., Uwizeyemungu, S., Pelletier, C. (2025). Human-Centered Artificial Intelligence in Higher Education: A Framework for Systematic Literature Reviews. *Information*, *16*(3), 240. https://doi.org/10.3390/info16030240
13. Lee, K., Paek, H., Ofoegbu, N., Rube, S., Higashi, M.K., Dawoud, D., Xu, H., Shi, L., Wang, X. (2025). A4SLR: An Agentic Artificial Intelligence-Assisted Systematic Literature Review Framework to Augment Evidence Synthesis for Health Economics and Outcomes Research and Health Technology Assessment. *Value in Health*, *28*(11), 1655-1664. https://doi.org/10.1016/j.jval.2025.08.002
14. Li, Y., Datta, S., Rastegar-Mojarad, M., Lee, K., Paek, H., Glasgow, J., Liston, C., He, L., Wang, X., Xu, Y. (2025). Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *Journal of the American Medical Informatics Association*, *32*(4), 616-625. https://doi.org/10.1093/jamia/ocaf030
15. Lund, B.D., Wang, T., Mannuru, N.R., Nie, B., Shimray, S., Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence- written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570-581. https://doi.org/10.1002/asi.24750

16. Malik, F.S., Terzidis, O. (2025). A hybrid framework for creating artificial intelligence-augmented systematic literature reviews. *Management Review Quarterly*. https://doi.org/10.1007/s11301-025-00522-8
17. Mogoale, P.D., Pretorius, A.B., Mogase, R.C., Segooa, M.A. (2025). Evaluating the Efficacy of AI Tools in Systematic Literature Reviews: A Comprehensive Analysis. *Journal of Information Systems and Informatics*, *7*(1), 870-888. https://doi.org/10.51519/journalisi.v7i1.1035
18. Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., Moreno-Garcia, C.F. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial Intelligence Review*, *57*(8), 200. https://doi.org/10.1007/s10462-024-10844-w
19. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuiness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *n71*. https://doi.org/10.1136/bmj.n71
20. Pinzolits, F.J.R. (2023). AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities*, *4*(1), 37-50. https://doi.org/10.53880/2744-2373.2023.4.37
21. Susnjak, T., Hwang, P., Reyes, N., Barczak, A.L.C., McIntosh, T., Ranathunga, S. (2025). Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning. *ACM Transactions on Knowledge Discovery from Data*, *19*(3), 1-39. https://doi.org/10.1145/3715964
22. Tosi, D. (2025). Comparing Generative AI Literature Reviews Versus Human-Led Systematic Literature Reviews: A Case Study on Big Data Research. *IEEE Access*, *13*, 56210-56219. https://doi.org/10.1109/ACCESS.2025.3554504
23. Vallamchetla, S.K., Abdelkader, O., Elnaggar, A., Ramadan, D., Islam Shourav, M.M., Riaz, I.B., Lin, M.P. (2025). Do it faster with PICOS: Generative AI-Assisted systematic review screening. *Journal of Biomedical Informatics*, *168*, 104860. https://doi.org/10.1016/j.jbi.2025.104860
24. van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., Oberski, D.L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125-133. https://doi.org/10.1038/s42256-020-00287-7
25. van Dijk, S.H.B., Brusse-Keizer, M.G.J., Bucsán, C.C., van der Palen, J., Doggen, C.J.M., Lenferink, A. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open*, *13*(7), e072254. https://doi.org/10.1136/bmjopen-2023-072254