From Abstracts to Full Texts: The Impact of Context Positioning in LLM-Based Screening Automation

Elias Sandner
CERN
Geneva, Switzerland
e-mail: elias.sandner@cern.ch

Marko Zeba
University of Technology
Graz, Austria
e-mail: m.zeba@student.tugraz.at

Igor Jakovljevic
CERN
Geneva, Switzerland
e-mail: igor.jakovljevic@cern.ch

Alice Simniceanu
WHO
Geneva, Switzerland
e-mail: simniceanua@who.int

Luca Fontana
WHO
Geneva, Switzerland
e-mail: fontanal@who.int

Andre Henriques
CERN
Geneva, Switzerland
e-mail: andre.henriques@cern.ch

Andreas Wagner
CERN
Geneva, Switzerland
e-mail: andreas.wagner@cern.ch

Christian Gütl
University of Technology
Graz, Austria
e-mail: c.guetl@tugraz.at

***Abstract*—Screening for relevant research is among the most time-intensive phases of a Systematic Review (SR), significantly impacting its timeliness and resource requirements. Automation through Large Language Model (LLM) promises substantial efficiency gains, potentially reducing the human screening workload and mitigating the risk of reviews becoming outdated prior to publication. Existing research has primarily explored LLM applications in Title & Abstract (TiAb) screening, achieving promising sensitivity but limited investigation into Full-Text (FT) screening. This study extends the 5-tier prompting approach, originally developed for TiAb screening, to FT screening. An experimental evaluation was conducted using the LLaMA 3.1 8B model on five real-world SR datasets. Two FT prompting strategies were tested: one that directly adapted the 5-tier TiAb approach to FT screening, and another addressing the known 'lost-in-the-middle' phenomenon by positioning eligibility criteria before and after the full text. Findings indicate that providing FT context improves workload reduction considerably, nearly doubling it in some cases, though sensitivity slightly decreased compared to TiAb screening. Notably, positioning eligibility criteria both before and after FT significantly improved performance, highlighting the importance of the prompt structure. These results demonstrate that careful prompt engineering enhances LLM effectiveness in FT screening, balancing the critical trade-off between sensitivity and workload reduction. Overall, this research underscores the potential of LLM-based FT screening, providing valuable insights into prompt optimization for systematic review automation.**

Keywords-systematic review; screening automation; full-text screening; LLM.

I. INTRODUCTION

The screening process, where researchers evaluate the relevance of papers to a predefined research question based on eligibility criteria, is one of the most time-consuming aspects of a Systematic Review (SR) [1]. The automation of this process is essential for reducing human workload, thereby enabling timely, high-quality evidence-based research, particularly in time-sensitive situations or projects with limited resources. Furthermore, automation addresses the challenge that some SRs become outdated already by the time they are published [2].

Several studies have evaluated Large Language Models (LLMs) for screening automation in Title & Abstract (TiAb) and, more recently, Full-Text (FT) screening. However, only a few of the LLM-powered TiAb screening approaches are extensible for FT screening.

Since token limits no longer restrict LLM-based screening automation to the TiAb phase, the traditional separation between TiAb and FT screening can be reconsidered in the context of automated approaches. The 5-tier prompting approach, which acts as a prefiltration mechanism by removing records where the LLM is highly confident that the eligibility criteria are not met, has demonstrated promising results for TiAb screening [3]. Given its inherent scalability, this study focused on extending and evaluating the 5-tier prompting method for FT screening. The following research question is addressed through an experimental evaluation in which two FT prompting strategies are benchmarked against the original TiAb prompt, using LLAMA 3.1 on five real-world datasets:

Does providing the FT as additional context during screening yield higher sensitivity and greater workload reduction compared to LLM-powered TiAb screening?

The remainder of the paper is organized as follows: In Section II, the steps needed to conduct a SR are described, followed by a summary of related work in LLM-powered screening in SRs. Furthermore, Section III describes how the experiments have been conducted and which SRs were used for evaluation. The results are presented in Section IV, while Section V answers the research question by interpreting them. Finally, Section VI concludes this study and points out potential

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

future work based on the findings of this study.

II. Background and Related Work

SRs follow a structured approach by i) retrieving potentially relevant primary research, ii) evaluating the eligibility of those candidate studies, and iii) synthesizing the relevant findings [4].

In the first phase of a SR researchers define a research question. Based on this, the corresponding eligibility criteria, which are divided into inclusion and exclusion criteria, are defined. An insensitive search string is used to retrieve potentially relevant papers from multiple academic libraries, followed by deduplication. The deduplicated records are then passed to the second phase to evaluate the relevance of these candidate studies [5]. This is typically done in a double-blinded mode, meaning two reviewers screen all records to minimize human errors and bias. In case of conflicts, a third reviewer's opinion is used to resolve it [6]. Initially, researchers evaluate the relevance of each paper based on its title and abstract, comparing it to the already defined eligibility criteria. Records that meet all inclusion criteria and do not violate any of the exclusion criteria in this initial screening stage are subsequently subject to FT screening based on the same eligibility criteria. After the second screening stage, the appropriate data gets extracted from the remaining papers and included in a descriptive analysis and a flow diagram to ensure transparency and reproducibility [5].

The mean duration of an SR from the PROSPERO registry [7] is approximately 67 weeks [8], with TiAb and FT screening being the most time-critical phases. [9] analyzed 319 SR requests from the SR request data from Weill Cornell Medicine's service. Out of the 319 SR requests, 30% were abandoned during TiAb and 24% during FT screening, underscoring the criticality of these two screening stages.

Due to the remarkable performance improvements of LLMs across various downstream tasks over the last few years, several studies have experimented with automating screening in SRs using such models. By introducing Instruction Structure Optimized (ISO) prompting and their ISO-ScreenPrompt [10], researchers achieved results over 90% in terms of accuracy, sensitivity, and specificity on the training and validation datasets for FT screening. [11] demonstrated how Retrieval Augmented Generation (RAG) with GPT-4 [12] can effectively be used for FT screening. In their setting, the FT of each paper served as the document set from which the LLM retrieved information. The evaluation of their approach on one completed SR resulted in a specificity of 99.6%.

Other studies with either insufficient [13] or only in some cases on par with human [14] results underline the difficulty of automating FT screening with LLMs.

Given the limited amount of related work in FT screening, this study further focused on related work covering TiAb screening approaches. Studies were considered relevant if they were extensible and achieved a high sensitivity. Several studies on LLM-based TiAb screening have been found [15]–[17], but rarely any reach a sensitivity of greater than 99%. To be used in real-world practice as a replacement for human screeners, any automation approach must meet this sensitivity level, as required by Cochrane [18]. The 5-tier approach [3] is one study that introduced a scalable prompting strategy and reached the Cochrane sensitivity requirement. This was achieved by classifying papers into five categories, ranging from 1 (highly relevant) to 5 (not relevant). Papers that the LLM assigns to category 5 are excluded automatically, while those in categories 1 to 4 remain subject to human screening. This approach, which excludes only studies where the LLM is highly confident of ineligibility, maximizes sensitivity. However, the effectiveness of the 5-tier approach on open-source LLMs and its application to FT screening have not yet been investigated.

III. Methodology

Compared to the 5-Tier-Prompting case study [3], LLaMA 3.1 8B has served as LLM for evaluation instead of GPT-4. Figure 1 depicts the required steps to conduct experiments on TiAb and FT screening, which are subsequently described in detail. The code used to conduct the experiments can be found in the supplementary material provided through Zenodo [19].

As the SYNERGY dataset [20] was used for evaluation, the first stage of the pipeline included FT retrieval via the BioC API for PMC Open Access [21] followed by parsing the XML response. Due to restricted access to retrieve the FT of various papers, only the 5 largest SRs after retrieval were considered for evaluation as they most closely mirrored real-world conditions. The selected SRs had a substantial number of studies to screen with a relatively small proportion of studies ultimately included. In this paper, the term 'dataset' was used for each of the 5 selected SRs. Table I gives a brief overview of each dataset regarding the topic of the SR, number of total records, and number of records with decision 'include' as ground truth after retrieval.

TABLE I. DATASETS AFTER FT RETRIEVAL

Dataset	Topic(s)	Records	Included
Bos_2018	Medicine	1163	5
Brouwer_2019	Psychology, Medicine	6482	11
Leenaars_2020	Medicine	791	75
van_Dis_2020	Psychology, Medicine	1753	15
Walker_2018	Biology, Medicine	3234	88

In prompt construction, three different prompts have been used. The 5-Tier-Prompting approach [3] for TiAb screening served as the baseline, while two FT screening approaches have been introduced:

- **FT1**: The 5-Tier prompt has been adjusted accordingly by exchanging the terms 'title and abstract' to 'fulltext' (see Table II) and the FT has been passed instead of TiAb.
- **FT2**: In addition to the changes in FT1, the eligibility criteria have been positioned before and after the FT. This approach was inspired by the study of [10] to address the 'lost-in-the-middle' phenomenon [22]–[24].

Table III summarizes the prompt structure used in the three approaches. For the baseline (TiAb), the same approach as in [3] was used. The structure for FT1 was similar to that in [3], where the only adjustments were the new system prompt and

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

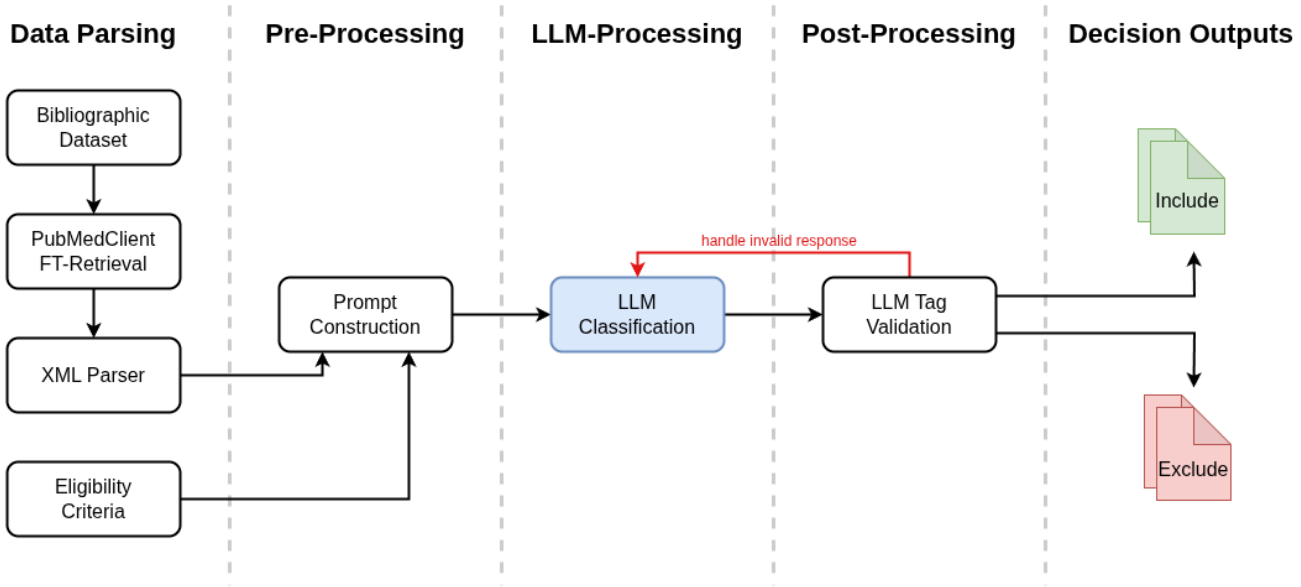


Figure 1. Conceptual Architecture.

TABLE II. SYSTEM PROMPT FOR FT SCREENING

Structure	Prompt
Role Play Instruction	You are a researcher rigorously screening full-texts of scientific papers for inclusion or exclusion in a review paper. Based on the provided inclusion and exclusion criteria listed below, you are asked to assign the paper to one of the following groups:
5-Tier-Group Definition	"1 Highly Relevant": Based on the given fulltext, the paper meets all inclusion criteria and no exclusion criteria. Therefore, the paper will be included. "2 Probably Relevant": The information provided in the full-text indicates that the paper is likely relevant. "3 Undecidable": The given full-text does not contain enough information to evaluate whether the inclusion and exclusion criteria are met. "4 Probably Irrelevant": Based on the given full-text, it is likely that at least one inclusion criterion is not met or that at least one of the exclusion criteria is met. "5 Not Relevant": Based on the full-text, it is clear that the paper does not meet the criteria. Therefore, the paper will be excluded.
Response Instruction	Based on the probability of a paper meeting all inclusion criteria and no exclusion criteria, assign it to one of the five categories. Only type the number of the group as "1", "2", "3", "4" or "5" in your response. Do not type anything else.

TABLE III. PROMPT STRUCTURE FOR SCREENING

TiAb	FT1	FT2
System Prompt [3]	System Prompt Table II	System Prompt Table II
Title and Abstract	Full-Text	Eligibility Criteria
Eligibility Critera	Eligibility Criteria	Full-Text
		Eligibility Criteria

the provision of FT instead of TiAb. FT2 is an extension of FT1, which aimed to address the 'lost-in-the-middle' phenomenon.

The constructed prompt was then passed to a locally hosted LLaMA 3.1 8B model for screening. By including a validation function to check whether the response of the LLM was only a number between 1 and 5 (as requested in the prompt), invalid responses were eliminated. If, after five retries, the response for a paper was still invalid, the paper got manually assigned '1' as tag value. By assigning '1', the paper was subject to human screening in the setting, as the LLM was not able to provide a valid screening decision.

This modular architecture allows the approach to be adapted to alternative FT retrieval mechanisms and other LLMs. Additionally, the prompt construction is independent of the underlying eligibility criteria and candidate studies, enabling applicability to large-scale systematic reviews without restriction to specific topics or domains.

Similar to [3], evaluation is based on two metrics: sensitivity and workload reduction. Sensitivity as defined in (1) is the fraction of ground-truth includes classified by the LLM as include. It measures the risk of missing relevant studies. Workload reduction (2) is based on the assumption that screening automation is integrated into the systematic review workflow as a filtration step, whereby records classified as exclude are removed prior to the human screening phase. Consequently, it represents the proportion of papers excluded by the model.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (1)$$

$$\text{Workload reduction} = \frac{\text{True Negative + False Negative}}{\text{n}} \quad (2)$$

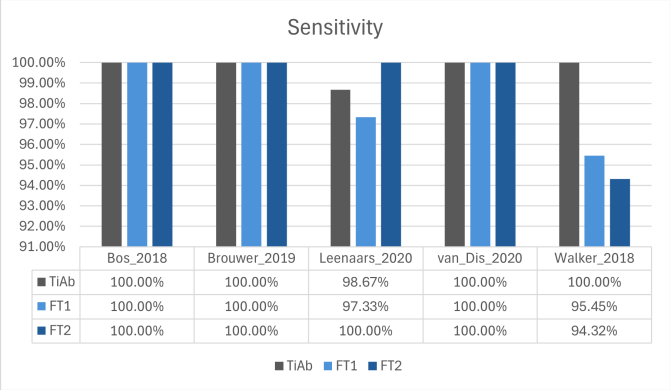
where n represents the total number of papers

IV. RESULTS

As in [3], alpha = 4 was used to transform the LLM response into a binary classification for further analysis, meaning that all

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

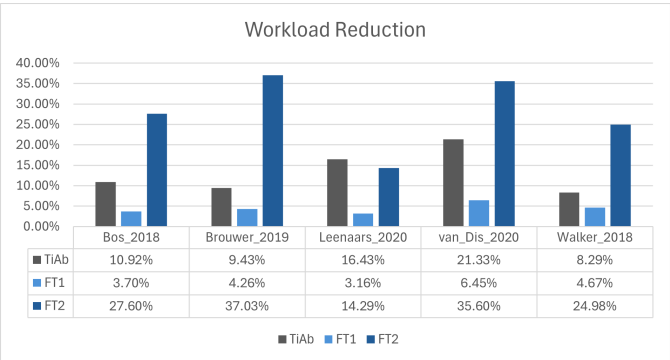
records with a tag value smaller or equal to 4 got the decision 'include' and entries with tag value 5 got the decision 'exclude'. Figure 2 shows the sensitivity per dataset for each prompting approach. FT1 was on par with TiAb on three datasets, while FT2 outperformed TiAb in Leenaars_2020, where TiAb had 98.67% and FT 100% sensitivity. Overall, in only 4 cases, a sensitivity of less than 100% was achieved. For Walkers_2018 both FT approaches and Leenaars_2020 TiAb and FT1, the sensitivity was less than 100%. The lowest sensitivity score had FT2 for Walkers_2018 with only 94.32%.



	Bos_2018	Brouwer_2019	Leenaars_2020	van_Dis_2020	Walker_2018
TiAb	100.00%	100.00%	98.67%	100.00%	100.00%
FT1	100.00%	100.00%	97.33%	100.00%	95.45%
FT2	100.00%	100.00%	100.00%	100.00%	94.32%

Figure 2. Sensitivity per Dataset for each Prompting Approach.

In Figure 3, the workload reduction per dataset is visualized. FT2 outperformed TiAb in 4 out of 5 datasets, whereas FT1 showed the weakest workload reduction of all three approaches in every dataset. In the most extensive dataset, Brouwer_2019 with 6482 records, FT2 outperformed TiAb in terms of workload reduction, achieving 37.03% compared to 9.43%, whereas TiAb outperformed FT2 on the smallest dataset, Leenaars_2020 with 791 records, with 16.43% compared to 14.29%.



	Bos_2018	Brouwer_2019	Leenaars_2020	van_Dis_2020	Walker_2018
TiAb	10.92%	9.43%	16.43%	21.33%	8.29%
FT1	3.70%	4.26%	3.16%	6.45%	4.67%
FT2	27.60%	37.03%	14.29%	35.60%	24.98%

Figure 3. Workload Reduction per Dataset for each Prompting Approach.

Figure 4 and Figure 5 give valuable information when comparing the three different approaches by visualizing the weighted averages of sensitivity and workload reduction for TiAb, FT1, and FT2. Given the varying sizes of the datasets used in the evaluation, the results have been weighted by the size of each dataset to obtain weighted averages for sensitivity and workload reduction. In this way, the weighting prevents the largest dataset from disproportionately influencing the average performance of each screening method.

On average, TiAb screening achieved a sensitivity of 99.73%, while FT1 and FT2 had 98.56% and 98.86% respectively.

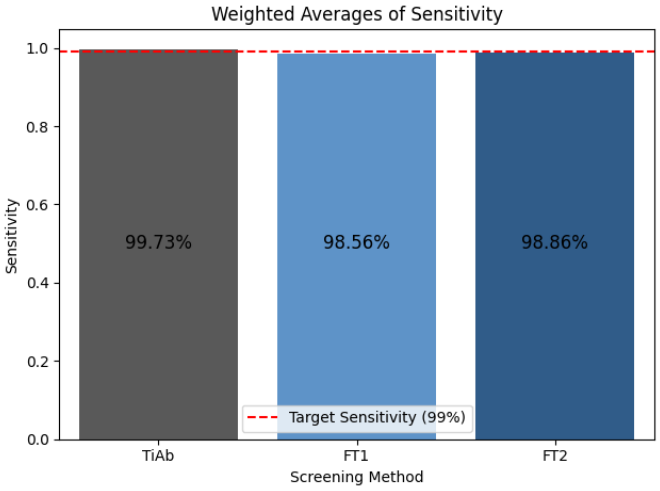


Figure 4. Weighted Average of Sensitivity for each Prompting Approach.

However, when comparing the workload reduction of each approach, FT2 achieved the highest result with 27.9%, almost 15% higher than TiAb, which had a 13.28% reduction. FT1 turned out to perform weakest in terms of workload reduction with only 4.45%.

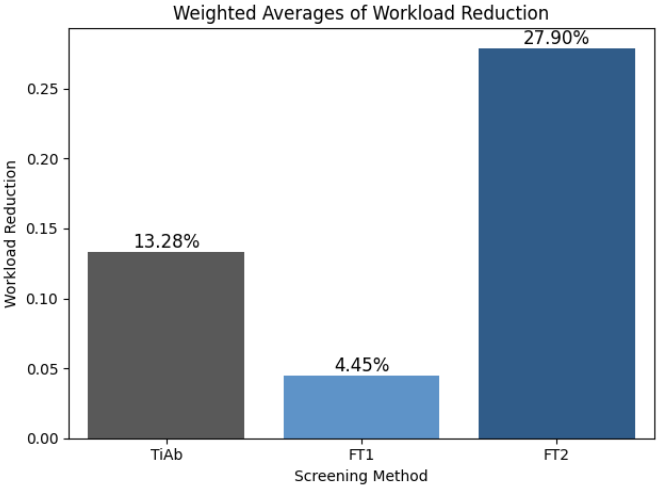


Figure 5. Weighted Average of Workload Reduction for each Prompting Approach.

V. Discussion

The results showed that when using LLaMA 3.1 8B, only LLM-powered TiAb screening meets Cochrane's sensitivity requirement (sensitivity $\geq 99\%$).

The experiments using FT were based on the assumption that additional context might increase the workload reduction by reducing the number of irrelevant papers included by the LLM, while maintaining a high sensitivity.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

FT1 turned out to be the least favorable setting, achieving the lowest sensitivity and workload reduction among all three approaches. The results were counterintuitive to the assumption that additional context might help the LLM to make better screening decisions. Hence, FT2 was introduced to check whether the task was too complex for the model or, due to the long context of the instruction and paper, the LLM got lost in the middle of processing and 'forgot' about eligibility criteria. The only difference between FT1 and FT2 was the positioning of the eligibility criteria. By setting the eligibility criteria before and after the FT in the prompt, sensitivity increased slightly, while workload reduction improved significantly. [10] also reported performance increases when positioning criteria before and after FT to address the 'lost-in-the-middle' phenomenon, confirming the plausibility of the FT2 approach.

The lower sensitivities of the FT approaches compared to TiAb mean that the LLM wrongly classified certain papers as 'exclude' while the ground truth was 'include'. When comparing papers against the eligibility criteria, humans naturally focus more on methodological chapters and results. When providing the FT of a paper to an LLM, chapters such as the related work and future work might be misleading and could influence the screening decision.

VI. Conclusion and Future Work

The conducted experiments gave valuable insights into LLM-powered TiAb and FT screening. Additional context, provided by passing the FT instead of only the TiAb, showed no further improvements in terms of sensitivity. Other approaches need to be considered to verify whether additional context might help to constantly reach over 99% of sensitivity. However, the increase in workload reduction in the FT2 setting indicates that additional context provided enhances the decision making of LLMs during screening.

A more extensive evaluation dataset, with SRs from different topics, could confirm the robustness of the 5-Tier-Prompting approach for FT screening. Given the rapid evolution of LLMs, an evaluation with newer models could provide further insights. As this study focused on the use of an open-weight model, a comparison of results between LLaMA 3.1 8B and newer LLaMA models could give further insights into whether the currently false negative classifications are occurring due to the high complexity of the task. Lastly, not all parts of a paper are likely to be relevant during the screening process. New experiments, where LLMs are enhanced to focus more on relevant chapters by either changing the current prompt or introducing pre-processing of FTs, might further improve performance.

In summary, this study confirmed the significant potential of the 5-Tier-Prompting approach. Although the extension of the approach by considering the FTs requires further evaluation, the results with LLaMA 3.1 8B are promising and potentially open up even better results with newer open-weight models. Nonetheless the small decrease in terms of sensitivity when using FT needs to be further investigated.

Acknowledgements

The joint CERN and WHO ARIA [25] project is funding the PhD project of Elias Sandner and the research project and stay at CERN of Marko Zeba.

References

[1] B. Nussbaumer-Streit *et al.*, "Resource use during systematic review production varies widely: A scoping review", *Journal of clinical epidemiology*, vol. 139, pp. 287–296, 2021.

[2] K. G. Shojania *et al.*, "How quickly do systematic reviews go out of date? a survival analysis", *Annals of internal medicine*, vol. 147, no. 4, pp. 224–233, 2007.

[3] E. Sandner *et al.*, "Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach Meeting Cochrane's Sensitivity Requirement", in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, Dubai, United Arab Emirates: IEEE, Nov. 2024, pp. 150–159, ISBN: 979-8-3503-5479-9. DOI: 10.1109/FLLM63129.2024.10852425.

[4] A. Pollock and E. Berge, "How to do a systematic review", *International Journal of Stroke*, vol. 13, no. 2, pp. 138–156, 2018, PMID: 29148960. DOI: 10.1177/1747493017743796. eprint: https://doi.org/10.1177/1747493017743796.

[5] E. Calderon Martinez *et al.*, "Ten Steps to Conduct a Systematic Review", *Cureus*, vol. 15, no. 12, e51422, 2023, ISSN: 2168-8184. DOI: 10.7759/cureus.51422.

[6] J. P. Higgins *et al.*, *Cochrane Handbook for Systematic Reviews of Interventions Version 6.5 (updated August 2024)*. Cochrane, 2024.

[7] Centre for Reviews and Dissemination, University of York, "Prospero: International prospective register of systematic reviews", 2025, [Online]. Available: https://www.crd.york.ac.uk/prospero/ (visited on 09/05/2025).

[8] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry", eng, *BMJ open*, vol. 7, no. 2, e012545, Feb. 2017, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-012545.

[9] M. R. Demetres, D. N. Wright, A. Hickner, C. Jedlicka, and D. Delgado, "A decade of systematic reviews: An assessment of Weill Cornell Medicine's systematic review service", *Journal of the Medical Library Association : JMLA*, vol. 111, no. 3, pp. 728–732, Jul. 2023, ISSN: 1536-5050. DOI: 10.5195/jmla.2023.1628.

[10] C. Cao *et al.*, *Prompting is all you need: LLMs for systematic review screening*, Jun. 2024. DOI: 10.1101/2024.06.01.24308323.

[11] F. Trad *et al.*, *Streamlining Systematic Reviews: A Novel Application of Large Language Models*, Dec. 2024. DOI: 10.48550/arXiv.2412.15247. arXiv: 2412.15247 [cs].

[12] OpenAI, "Gpt-4 system card", 2023, [Online]. Available: https://cdn.openai.com/papers/gpt-4-system-card.pdf (visited on 09/05/2025).

[13] X. Chen and X. Zhang, *Large language models streamline automated systematic review: A preliminary study*, 2025. arXiv: 2502.15702 [cs.IR].

[14] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield, "Can large language models replace humans in systematic reviews? Evaluating GPT -4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages", *Research Synthesis Methods*, vol. 15, no. 4, pp. 616–626, Jul. 2024, ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1715.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

[15] A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, *The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews*, May 2024. DOI: 10.48550/arXiv.2404.15667. arXiv: 2404.15667 [cs].

[16] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain", *Systematic Reviews*, vol. 13, no. 1, p. 158, Jun. 2024, ISSN: 2046-4053. DOI: 10.1186/s13643-024-02575-4.

[17] L. Affengruber *et al.*, "An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review", *BMC Medical Research Methodology*, vol. 24, no. 1, p. 210, Sep. 2024, ISSN: 1471-2288. DOI: 10.1186/s12874-024-02320-4.

[18] J. Thomas *et al.*, "Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for cochrane reviews", *Journal of Clinical Epidemiology*, vol. 133, pp. 140–151, 2021.

[19] E. Sandner *et al.*, *From abstracts to full texts: The impact of context positioning in llm-based screening automation*, Accessed: 2025-10-21, 2025. DOI: 10.5281/zenodo.16419874.

[20] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot, *SYNERGY - Open machine learning dataset on study selection in systematic reviews*, version V1, 2023. DOI: 10.34894/HE6NAQ.

[21] National Center for Biotechnology Information, "Bioc-pmc: Biomedical natural language processing apis", 2025, [Online]. Available: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/ (visited on 09/05/2025).

[22] S. An *et al.*, "Make your llm fully utilize the context", *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 160–62 188, 2024.

[23] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts", *arXiv preprint arXiv:2307.03172*, 2023.

[24] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks", *arXiv preprint arXiv:2309.17453*, 2023.

[25] World Health Organization, "Aria tool on who partners platform", 2025, [Online]. Available: https://partnersplatform.who.int/tools/aria (visited on 09/05/2025).

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org