En Passant: How Section-Based LLM Processing Unlocks Full-Text Evidence in Systematic Literature Reviews

Clemens Brackmann[1][0009-0002-6667-0165] and Kathrin Nauth[2][0009-0007-3457-102X]

[1] University of Duisburg-Essen, Essen, Germany
clemens.brackmann@uni-due.de
[2] Ruhr University of Bochum, Bochum, Germany
kathrin.nauth@isse.ruhr-uni-bochum.de

Abstract. Systematic literature reviews (SLRs) are essential for synthesizing prior research, identifying trends, gaps, and future directions across various fields. However, SLRs are time-consuming and subject to human cognitive constraints that threaten their rigor, consistency, and replicability. Recent advances in generative artificial intelligence (AI) provide transformative opportunities to enhance the efficiency and scope of literature reviews by automating repetitive tasks, analysing full-text analysis, and uncovering latent patterns. We propose a step-by-step LLM-integrated methodology for SLR that covers two stages currently underserved by existing tools: abstract screening and section-based full-text analysis. The methodology explicitly addresses three known failure modes of prior automated approaches: context-window limitations (resolved through section-based document decomposition), hallucination risks (mitigated through narrow binary prompting and mandatory validation sampling), and replicability deficits caused by opaque tool fine-tuning (resolved by using general-purpose LLMs with fully documented prompts and decision rules). Challenges related to full-text accessibility, PDF parsing, and LLM consistency are acknowledged and addressed with concrete guidance. The methodology is primarily intended for systematic and structured literature reviews where comprehensive, traceable, and replicable coverage is the goal.

Keywords: Literature Review, Artificial Intelligence, Large Language Model, Full-Text Analysis.

1 Introduction

Systematic literature reviews (SLR) are a tool for researchers to structure a research field by synthesizing prior studies to identify trends, gaps, interconnections with other research fields, and future directions within that field. Performing an effective literature review creates a foundation upon which new knowledge can be built [30], that is created through interpreting existing knowledge and combining this knowledge [29]. Thereby the intellectual development and progress of a topic can be captured systematically. Literature reviews are more than mere summaries of prior works. They are critical analyses that prepare research findings, organize and evaluate them and integrate them to

advance theories on the one hand, and practical understanding on the other. Since the amount of research involved in literature reviews is substantial and covers several fields of research, a carefully planned and structured process must be followed. The importance of this comprehensive review process is emphasized by vom Brocke et al. [29] and proposed as a meticulous and replicable approach for literature reviews, ensuring that a broad range of research fields is considered and appropriately contextualized to enable researchers to build up the previously mentioned knowledge base.

Recent developments in AI technology have raised important questions about how to integrate these tools into the systematic literature review process without compromising its defining standards of rigor, consistency, and replicability. Traditional SLR methods, while well-established, are time-consuming and subject to human cognitive constraints that introduce variability across reviewers [24]. Yet integrating AI into SLRs introduces its own challenges: LLM outputs may be inconsistent across runs, model fine-tuning is often opaque to the user, and hallucination can silently corrupt screening and extraction results. Generative AI can also expedite the search, synthesis, and analysis of vast bodies of literature [18]. For instance, AI-driven language models can assist researchers in identifying relevant studies, summarizing key findings, and detecting emerging themes across numerous publications [25]. This technological shift enables researchers to handle larger corpora and supports more iterative review processes. However, the promise of AI-assisted SLRs is undermined by several concrete failure modes in existing approaches. First, most AI-supported tools operate exclusively at the metadata level (titles, abstracts, keywords), leaving the full text of papers unanalysed. Second, purpose-built LLM tools for literature reviews (e.g., Elicit, Consensus) introduce replicability risks because their model fine-tuning and internal logic are not disclosed to the user, making it impossible to verify or reproduce results. Third, when general LLMs are applied to full-text inputs without structured decomposition, attention loss in long documents (the “lost in the middle” problem) and hallucination risk undermine reliability. Existing research mainly focuses on mapping available AI and automation tools to the individual steps of the SLR process while evaluating their usefulness and pointing out potential pitfalls with respect to compliance with the SLR quality criteria (e.g., 26, 7). Specified LLM-based tools frequently exhibit a disadvantage in terms of transparency regarding the models’ fine-tuning, which can have a detrimental effect on the quality criteria of replicability, as software versions are usually not traceable for the user. The application of general LLM models offers the potential to address these limitations by aligning more closely with the specific requirements of researchers and the demands of replicability in SLR. The present research focuses specifically on two steps of the SLR process: (1) title/abstract screening and (2) section-based full-text analysis. These steps exhibit the highest risk of replicability loss and the highest potential for further automation [7, 26]. The proposed method is primarily intended for systematic and structured literature reviews [19] where the goal is comprehensive, traceable, and replicable coverage of a research field. It may be less suited to interpretive or narrative review types where subjective synthesis is central. Based on prior evaluations of automation for literature reviews (e.g., 7, 26), we see the highest risk of replicability loss, as well as the highest potential to further automate SLR in these steps. Against this drawback, we propose the following research question:

How can LLMs be used to support title/abstract screening and section-based full-text analysis in systematic literature reviews while preserving the consistency and replicability of the SLR process?

To address this question, an LLM-enhanced workflow for SLR will be provided, which can be readily extended by scholars with additional AI-Tools selected and evaluated in previous research, such as by Tingelhoff et al.

Building on the rigorous processes proposed by vom Brocke et al. [29] and Webster and Watson [30], we make three distinct contributions. First, we present a step-by-step LLM-integrated workflow for SLR that explicitly covers both title/abstract screening and section-based full-text analysis; a combination not addressed by prior tooling. Second, we propose a section-based document decomposition strategy that directly mitigates the "lost in the middle" attention problem of LLMs, improving both reliability and traceability of extraction results. Third, by using general-purpose LLMs (rather than opaque fine-tuned tools), the workflow preserves replicability: model identifiers, prompts, temperature settings, and decision rules can all be documented and reproduced. We propose a procedure that enables researchers to accelerate their SLRs by automating full-text analysis through partitioning articles into sections before LLM processing, thereby providing a higher degree of traceability and preventing quality losses due to the lost-in-the-middle problem [15, 31]. The remainder of this research article is structured as follows: chapter two provides an overview of relevant research towards AI in literature reviews including application examples of generative AI in these. Chapter three introduces our research methodology that proposes processing full-text articles with LLMs by answering predefined questions. We explain in detail how the full-texts are pre-processed before being fed into the LLM. Finally, the methodology is discussed in chapter four, and the article closes with a conclusion and an outlook.

2 AI in Literature Reviews

Literature reviews serve as a cornerstone of academic research, providing a systematic synthesis of prior research to identify gaps, establish theoretical frameworks and inform future investigations. They must follow a strict process to ensure comprehensibility and replicability. Thus, several traditional methods have been established [13, 20, 27]. However, recent advances in AI technology have significantly enhanced the scope and efficiency of literature reviews, allowing researchers to handle larger datasets and uncover patterns beyond human cognitive limits.

AI-driven tools increasingly assist in automating labour-intensive aspects of literature reviews, such as data extraction, categorization, and summarization. These tasks are often time-consuming and prone to subjective perception due to the involvement of human decisions. The integration of AI-driven tools for literature reviews aims to automate and accelerate the process, but also to reduce the possible source of errors. These tools often rely on natural language processing (NLP) techniques to process titles, abstracts, and keywords of academic articles. However, despite their potential, most AI-supported tools fall short of incorporating full-text analysis due to computational and licensing constraints. Instead, they focus on metadata-level insights, leveraging

bibliometric and scientometric methods to map research landscapes (i.e., co-citation networks or author collaborations) and to identify influential works or emerging trends [2, 3].

Bibliometric and scientometric techniques form a foundation for AI-supported literature reviews, utilizing citation analysis, co-authorship patterns, and keyword clustering to evaluate research impact and connectivity. By utilizing these methods researchers can identify topics of interest within a research domain and propose future research agendas resulting from identified gaps. Tools like VOSviewer [28], CiteSpace [1], and pyBibX [21] employ these approaches, enabling researchers to visualize relationships between academic contributions. Such methods are crucial for understanding the evolution of research fields and for identifying underexplored areas [28]. Most bibliometric or scientometric methods rely on statistical evaluation and analysis of the underlying metadata to gain valuable insights.

Further statistical methods, particularly Latent Dirichlet Allocation (LDA) and other topic modelling techniques, have gained prominence in structuring literature reviews and gaining an overview of a research field [5]. LDA identifies thematic patterns in large datasets by grouping words into topics, allowing researchers to explore overarching trends across disciplines [5]. When combined with AI, topic modelling enhances the ability to synthesise disparate research fields, uncover latent connections, and predict future research trajectories [32].

Most literature review tools prioritize titles, abstracts, and keywords for analysis due to their accessibility and relevance. These metadata elements capture the essence of academic articles while circumventing the need for full-text access, which is often restricted by copyright or technical barriers. While this limitation narrows the scope of insight, the integration of AI mitigates some challenges by providing high-level summaries and visualizations that remain robust even in the absence of full-text data. When it comes to analysing full-texts, recent research on LLMs seems promising [22]. Models like GPT or Llama are capable of processing given input texts and providing reasonable answers to prompted questions, indistinguishable from human behaviour [6]. However, the wording of the questions is important for generating high-quality answers [9, 14]. Asking an LLM to provide an answer referring to a whole article faces the challenge of hallucination and important information within the article might be ignored by the LLM or even unintended answers are given despite the users' expectations [11]. Thus, integrating AI or generative AI into full-text analysis is a challenge especially in the automation of the literature review process. Since full-texts are mostly not considered by tools supporting literature review processes, we want to address this gap and propose a method that enables AI-supported literature reviews to be automated to a certain point while covering full-text analysis instead of only relying on metadata evaluation.

3 Methodology

The proposed method integrates LLMs into the established SLR process of vom Brocke et al. [29] at two specific stages: (1) title and abstract screening, and (2) section-based full-text analysis. Together, these two stages represent the highest-effort and highest-

variability steps in a manual SLR and therefore offer the greatest potential for automation while preserving replicability. All other steps follow established SLR practice. Throughout this section, steps are marked [Automated] where no human intervention is required, and [Manual / Review] where human oversight is necessary. Figure 1 provides an overview of the complete workflow.

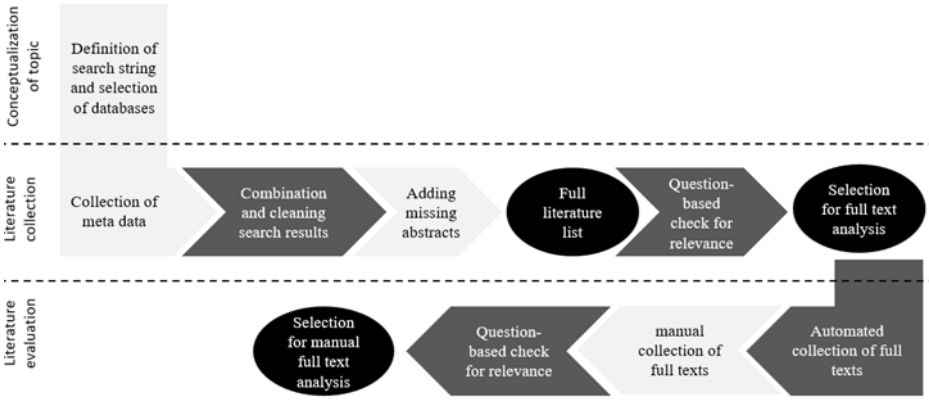

Fig. 1. LLM-enhanced SLR process with automated steps as dark grey arrows, and interim results as black circles.

3.1 Conceptualization and Metadata Collection

The first stage [Manual] follows conventional SLR practice: the researcher defines the research question, formulates the search string, and selects the databases to query. To enable automated downstream processing, search results should be exported as .bib files, which guarantee consistent metadata field names across databases. The relevant fields to retain are: author, year, title, doi, url, abstract, journal, keywords, and source database.

Duplicate removal [Automated] is performed by matching on doi as the primary key, or on title + year + first author when no doi is available. Entries that cannot be further processed (e.g., conference recordings, book chapters without abstracts) should be excluded at this stage. Where abstracts are missing, they can be automatically retrieved via the Crossref API [Automated]; any remaining gaps should be filled manually via publisher websites [Manual]. In cases where no abstract can be obtained, the title may be used as a substitute for the subsequent screening step [17], though this introduces additional uncertainty that should be reported.

3.2 Stage 2: LLM-Based Abstract Screening

Conventional SLRs employ a two-stage content screening process: first filtering by title, then by abstract [13]. If conducted manually, abstract screening is the most time-consuming of these steps. Integrating an LLM at this stage [Automated] can

substantially accelerate the process and renders a separate title screening step redundant, as the LLM can evaluate both simultaneously from the abstract text.

The proposed approach differs from both manual screening and from purpose-built AI screening tools such as ASReview or Elicit. Unlike active-learning tools (e.g., ASReview), the LLM-based approach requires no pre-labelled training set and can be deployed immediately with custom, domain-specific questions. Unlike opaque specialized tools (e.g., Elicit, Consensus), the use of a general-purpose LLM with fully documented prompts and settings ensures that the screening process is transparent and reproducible. Compared to asking the model a single open-ended relevance question, we recommend a multi-question approach: a set of narrow, independently answerable yes/no questions is posed against the abstract. This preserves context-specific significance [26] and reduces the risk that a single ambiguous judgment drives the inclusion or exclusion decision.

Prompt design. Each prompt should follow the template: "{abstract} — based on the abstract, please answer the following question with 'yes' or 'no': [question]." The temperature setting should be kept low (e.g., 0.2) to maximise output consistency. All LLM responses and screening decisions must be logged for auditability. Example questions for a review on AI methods in literature reviews might include: (A) "Does the article discuss the use of AI or machine learning methods?" and (B) "Does the article relate to literature review methodology or bibliometric analysis?" [10].

Decision rules. Exclusion rules must be defined explicitly before running the pipeline and applied consistently to all papers. For example: "Exclude if answer to A is 'no' OR answer to B is 'no'." If the model does not return a valid binary response [Manual / Review] — for instance returning "Sorry, I cannot answer the question" due to insufficient information in the abstract — these cases should be flagged and manually reviewed or carried forward to the full-text stage to avoid excluding potentially relevant papers. Both the defined decision rules and the handling of non-response cases must be documented and reported to enable replication.

Validation [Manual / Review]. To establish confidence in the LLM screening step, a stratified sample of approximately 50 papers (varying by publication year, venue, and discipline) should be independently evaluated by a human reviewer applying the same inclusion criteria [16]. Basic agreement metrics, such as Cohen's kappa or raw percentage agreement, and a brief characterization of disagreement modes (e.g., hallucination-driven errors, ambiguous abstracts) should be reported. A simple stability check e.g., re-running the same prompt set on a random subset of 20–30 papers, is also recommended to assess output consistency across LLM runs, since stochastic variation can affect results even at low temperature settings [12, 33].

3.3 Stage 3: Full-Text Retrieval

Once the abstract screening stage has produced a reduced candidate set, full texts must be retrieved for in-depth analysis. Publishers and databases such as Elsevier, Springer, Wiley, and IEEE provide full-text access via APIs for research purposes [Automated]; the Crossref API can be used to extract download links not included in the initial metadata export. Papers not accessible via an automated route must be downloaded

manually [Manual]. Automated web scraping is technically feasible but raises legal concerns under most database terms of service and is therefore not recommended.

A non-trivial proportion of papers in any corpus will be inaccessible due to paywalls, API coverage limits, or the absence of a digital full text. These technically-driven exclusions must be clearly distinguished from research-driven exclusions (i.e., papers excluded for lack of relevance) in the PRISMA-style reporting of the review process [4]. Researchers should comment on the likely structural biases introduced by inaccessibility e.g., paywalled or older publications may be underrepresented, and assess how this affects the generalizability of the review findings.

3.4 Stage 4: Document Parsing and Section Extraction

Full-text analysis requires the content of individual paper sections rather than the full document to be provided to the LLM. This is necessary to address the "lost in the middle" problem: even advanced LLMs such as GPT-4 show degraded attention when processing long input texts exceeding approximately 1,000 tokens, causing relevant information in the middle of the input to be overlooked [31]. Academic papers follow a consistent structure (introduction, methodology, results, discussion), making section-level decomposition a natural and reliable strategy.

Full texts should be retrieved as .pdf files to preserve document structure metadata. We recommend using the open-source nlm-ingestor tool from NLmatics, which parses .pdf files into structured sections using layout-based indexing and can be deployed locally via Docker [Automated]. The parsed output identifies section boundaries and labels, enabling targeted extraction of the sections relevant to the research question (e.g., the methods section for a review of research methodologies).

Parsing limitations [Manual / Review]. nlm-ingestor may fail to extract the correct section when: (a) the PDF contains scanned content without an OCR layer, (b) the document uses a multi-column layout that disrupts the reading order, or (c) the authors have used non-standard section labels (e.g., "Approach" instead of "Methods"). In such cases, manual identification of the relevant section is required. Papers where no parseable version of the target section can be obtained must be excluded as technically-driven exclusions and reported accordingly, separately from relevance-based exclusions. Researchers should note that technically-driven exclusions may disproportionately affect conference papers, older publications, and papers from certain geographic regions or publishers, potentially introducing selection bias into the corpus.

3.5 Stage 5: Section-Based Full-Text Analysis with LLMs

With section texts extracted, the LLM is applied to answer predefined questions about each section [Automated]. This is the core novel contribution of the proposed method: rather than asking the model to process and interpret an entire paper, each question is answered with reference to a specific, bounded section of the text. This targeted approach improves both the reliability of the LLM's responses and the traceability of the output. Hence, each answer can be linked to a specific section of a specific paper.

Prompt design. For extractive tasks (e.g., identifying the method used), we recommend a structured prompt template: "{section_content} - based on the [section name] section, please answer the following question using numbered bullet points: [question]." If multiple questions are to be answered from the same section, they should be grouped within a single prompt (e.g., A) and B)) to reduce API calls and avoid context fragmentation. The temperature setting should again be kept low (0.2) for consistency. For complex information extraction tasks, where the expected output is not binary but descriptive, few-shot examples of expected answer formats can be included in the system instructions to improve precision and reduce over-generalization [23].

Decision rules. Inclusion and exclusion criteria for the full-text stage must be defined before running the pipeline, independently of the abstract screening criteria. For example: "Include if the method section identifies at least one method from category X and at least one method from category Y." Cases where the model produces no response or an ambiguous answer should be flagged for [Manual / Review] rather than automatically excluded. Depending on the research objectives, exclusion decisions can be made section by section or based on aggregated responses across all queried sections.

Validation [Manual / Review]. As at the abstract screening stage, a validation sample should be manually reviewed. Because full-text extraction is a more complex task than binary screening, a sample of 20–30 papers is recommended as a minimum. The human reviewer should apply the same extraction questions independently and compute agreement with the LLM output. Disagreement modes should be categorized, for instance, distinguishing between hallucinated content not present in the source text, over-generalized labels, and genuine ambiguity in the source material. This categorization informs whether the disagreements reflect a systematic LLM limitation (which would need to be addressed through prompt revision) or acceptable noise in a well-performing pipeline.

After the final exclusion round, the remaining papers are available for content synthesis or, if necessary, for further in-depth manual analysis. The full sequence of automated and manual decisions, together with all prompts, decision rules, temperature settings, and toolchain version details, should be documented to ensure complete replicability of the review [8].

3.6 Reporting and Replicability Requirements

For the LLM-enhanced SLR to meet the replicability standards of the SLR tradition [29, 30], a minimum documentation set must accompany any published review that applies this methodology. This should include: (1) the complete prompt set for each pipeline stage, including system instructions if used; (2) the LLM model identifier and version (e.g., gpt-4o-mini); (3) the temperature setting and any other sampling parameters; (4) all decision rules for inclusion and exclusion, including the handling of non-response cases; (5) the toolchain used for parsing (e.g., nlm-ingestor version, Docker image identifier, Tika version); (6) a PRISMA-style attrition table distinguishing research-driven from technically-driven exclusions at each stage; and (7) the results of the validation sample, including the agreement metric and a description of disagreement modes. This documentation can be provided as a supplementary appendix to the

published review. Researchers should also consult the editorial policies of their target publication venue, as disclosure requirements for AI-assisted methods vary across IS outlets.

4 Demonstration

To demonstrate the workflow in practice, we applied it end-to-end to a single guiding research question: *What AI methods are currently used in academic literature reviews?* The aim of this application is illustrative rather than substantive. It is intended to show that the pipeline executes from metadata collection through to section-based extraction and produces a traceable, auditable decision trail at every stage. It is therefore presented as a worked example of the method, not as a definitive systematic review of AI methods in literature reviews, and its findings should be read in that light.

Seven databases were queried on 17 November 2024 (see Fig. 2). In line with the reporting and replicability requirements set out in section 3.6, this query date is treated as a documented parameter of the run rather than as an incidental detail: together with the model identifier (GPT-4o-mini), the temperature setting (0.2), the prompt set, and the toolchain versions, it fully specifies the conditions under which the result was produced. It allows any reader either to reproduce the run or to re-execute it from the same starting point forward. A fixed and reported search date is thus not a weakness of the demonstration but a direct instance of the transparency the method prescribes.

The initial result of this database query contained 5,218 articles from which duplicates and article types other than conference proceedings or journals were removed to retain 4,376 for the pipeline processing. In 32 cases the model returned no valid binary response; following the procedure in Section 3.2, these were manually reviewed and two were judged relevant and retained. Net of this manual step, abstract screening excluded 3,306 articles and retained 1,070. To validate this large exclusion sample, a stratified sample of 50 abstracts (varying by publication year and source database) was independently screened by a human reviewer applying the same inclusion question. Agreement with the LLM decision was 95%. Articles with available full-text and successful parsing were then further processed for section-based full-text analysis regarding the demonstrative research question. This full-text analysis was conducted based on the method section with the following prompt: “Name the method used to analyze literature in the article. Do not describe the method.” We also manually validated the results of this prompt by evaluating a randomly selected sample of 20 articles (approximately 3% of the total). The purpose of the analysis was to identify a final selection of relevant articles, which use AI methods for literature reviews.

All methods mentioned in response were catalogued and categorized, resulting in five categories: Literature Review Methods, NLP Methods, Bibliometric Methods, Machine Learning Methods, and Statistical Methods. We defined the inclusion criteria that the methodological approach of the articles includes at least one method from the literature review category and at least one method from the NLP or machine learning category. This led to a final selection of 36 articles which were then manually verified resulting in a final exclusion of four articles and a resulting set of 32 (see Fig. 2).

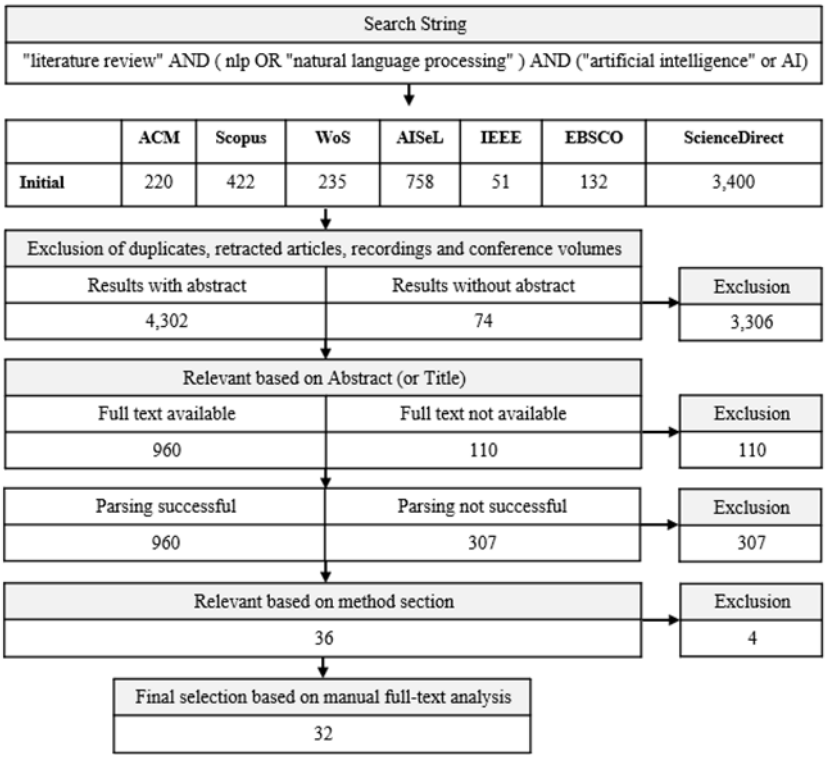

Fig. 2. Literature selection process based on Bandara et al. [4]

5 Discussion and Conclusion

The proposed methodology makes a focused methodological contribution to SLR literature by directly addressing three known failure modes of existing AI-assisted approaches. First, the section-based decomposition strategy resolves the full-text blindness of metadata-only tools: by partitioning papers into individual sections before LLM processing, the method enables structured extraction from the full text of each paper. Second, the reliance on a general-purpose LLM with fully documented prompts, decision rules, temperature settings, and model identifiers directly addresses the replicability deficit of purpose-built, opaque tools such as Elicit or Consensus. Any researcher with API access to the same model can reproduce the screening and extraction results. Third, section-based input directly mitigates the “lost in the middle” attention degradation that affects LLMs when processing long documents: by limiting the input of each LLM call to one bounded section, important content is less likely to be ignored. Together, these three design choices allow the methodology to achieve a level of transparency and replicability that existing full-text AI tools do not offer. The multi-question

abstract screening approach is a further distinguishing feature of the methodology. Rather than asking the model a single broad relevance question, decomposing the screening criteria into several narrow binary questions reduces the risk that a single ambiguous LLM judgment drives an inclusion or exclusion decision. This also produces a more auditable log of the screening process: each question-and-answer pair is a discrete, traceable unit of evidence. The section-based full-text analysis extends this logic to the extraction stage, where targeted questions about specific sections yield more precise and verifiable outputs than whole-paper queries.

Despite its contributions, the proposed methodology has important limitations that must be acknowledged. First, the abstract screening stage relies on the quality of available metadata. Titles and abstracts may lack the granularity of full texts and missing or poorly formatted abstracts introduce noise into the early screening stage. Second, full-text unavailability and parsing failures are unavoidable features of large-scale SLRs. Technically-driven exclusions, caused by paywalled content, API coverage limits, or complex PDF layouts, disproportionately affect conference papers, older publications, and articles from smaller publishers, and may introduce systematic selection bias. Researchers applying this method should report these exclusions separately from relevance-based exclusions in a PRISMA-style attrition table and should discuss the likely direction of any resulting corpus bias.

A further limitation of the methodology is its dependence on the chosen LLM. General-purpose models without domain-specific fine-tuning may produce outputs that lack domain depth or that over-generalise specific terms. Even at a low temperature setting, stochastic variation means that identical prompts may produce slightly different outputs across runs, making a stability check an essential component of validation. Additionally, API-based full-text retrieval creates accessibility barriers: institutional access to publisher APIs varies considerably, which may restrict replicability across geographical and institutional contexts. Future work should investigate the performance of the methodology with open-source, locally deployable LLMs (e.g., Llama) to mitigate API dependency and further strengthen replicability.

The proposed method offers a replicable and transparent starting point for LLM-assisted SLRs. Its robustness across different research domains, corpus sizes, and LLM models remains an important open question for future research. The primary technical challenge is the correct and complete extraction of sections from diverse PDF formats. Parsing quality could be further improved by incorporating AI or ML-based layout recognition, or by complementing nlm-ingestor with OCR pipelines for scanned documents. Future work should also investigate the application of few-shot prompt engineering and domain-adapted LLMs to improve extraction precision for specialized fields. Most importantly, a direct empirical comparison between LLM-supported and fully manual SLR processes, measuring both screening time and decision overlap, is needed to rigorously quantify the methodology's added value over established manual and semi-automated approaches. Such a comparison would also clarify for which review types and corpus sizes the LLM-based approach offers the greatest practical benefit.

References

1. (2016) CiteSpace: a practical guide for mapping scientific literature
2. Abbasova S (2023) Trends in scientific and public interest in managerial accounting: Bibliometric analysis. Problems and Perspectives in Management 21:650–664. doi: 10.21511/ppm.21(4).2023.49
3. Ardianto A, Anridho N (2018) Bibliometric Analysis of Digital Accounting Research. IJDAR:141–159. doi: 10.4192/1577-8517-v18_6
4. Bandara W, Furtmueller E, Gorbacheva E et al. (2015) Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support. Communications of the Association for Information Systems
5. Blei DM, Ng AY, Jordan MI (2002) Latent Dirichlet Allocation. In: Advances in neural information processing systems 14. Proceedings of the 2001 conference. MIT Press, Cambridge, Mass., pp 601–608
6. Burger B, Kanbach DK, Kraus S et al. (2023) On the use of AI-based tools like ChatGPT to support management research 26:233–241. doi: 10.1108/ejim-02-2023-0156
7. Da Silva Júnior EM, Dutra ML (2021) A roadmap toward the automatic composition of systematic literature reviews 1:1–22. doi: 10.47909/ijsmc.52
8. Galli C, Gavrilova AV, Calciolari E (2025) Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations 16:378. doi: 10.3390/info16050378
9. Giray L (2023) Prompt Engineering with ChatGPT: A Guide for Academic Writers 51:2629–2633. doi: 10.1007/s10439-023-03272-4
10. Homiar A, Thomas J, Ostinelli EG et al. (2025) Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review 28. doi: 10.1136/bmjment-2025-301762
11. Ji Z, Lee N, Frieske R et al. (2023) Survey of Hallucination in Natural Language Generation 55:1–38. doi: 10.1145/3571730
12. Khraisha Q, Put S, Kappenberg J et al. (2024) 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 15:616–626. doi: 10.1002/jrsm.1715
13. Kitchenham B, Brereton P (2013) A systematic review of systematic review process research in software engineering. Information and Software Technology 55:2049–2075. doi: 10.1016/j.infsof.2013.07.010
14. Lieberum J-L, Toews M, Metzendorf M-I et al. (2025) Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review 181:111746. doi: 10.1016/j.jclinepi.2025.111746
15. Liu NF, Lin K, Hewitt J et al. (2024) Lost in the Middle: How Language Models Use Long Contexts 12:157–173. doi: 10.1162/tacl_a_00638
16. Manning CD, Raghavan P, Schütze H (2022) An introduction to information retrieval, Reprinted. Cambridge University Press, Cambridge

17. Mateen FJ, Oh J, Tergas AI et al. (2013) Titles versus titles and abstracts for initial screening of articles for systematic reviews 5:89–95. doi: 10.2147/CLEP.S43118
18. Mozelius P, Humble N (2024) On the Use of Generative AI for Literature Reviews: An Exploration of Tools and Techniques. ecrm 23:161–168. doi: 10.34190/ecrm.23.1.2528
19. Paré G, Trudel M-C, Jaana M et al. (2015) Synthesizing information systems knowledge: A typology of literature reviews 52:183–199. doi: 10.1016/j.im.2014.08.008
20. Paul J, Lim WM, O'Cass A et al. (2021) Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). Int J Consumer Studies 45. doi: 10.1111/ijcs.12695
21. Pereira V, Basilio MP, Santos CHT (2023) pyBibX -- A Python Library for Bibliometric and Scientometric Analysis Powered with Artificial Intelligence Tools
22. Scherbakov D, Hubig N, Jansari V et al. (2025) The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review 32:1071–1086. doi: 10.1093/jamia/ocaf063
23. Schulhoff, S [Sander], Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S [Sevien], Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F [Feileen], Tao, H., Srivastava, A., . . . Resnik, P. (2024). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*.
24. Snyder H (2019) Literature review as a research methodology: An overview and guidelines. Journal of Business Research 104:333–339. doi: 10.1016/j.jbusres.2019.07.039
25. Susnjak T (2023) PRISMA-DFLLM: An Extension of PRISMA for Systematic Literature Reviews using Domain-specific Finetuned Large Language Models
26. Tingelhoff F, Brugger M, Leimeister JM (2025) A guide for structured literature reviews in business research: The state-of-the-art and how to integrate generative artificial intelligence 40:77–99. doi: 10.1177/02683962241304105
27. Tranfield D, Denyer D, Smart P (2003) Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. British J of Management 14:207–222. doi: 10.1111/1467-8551.00375
28. van Eck NJ, Waltman L (2011) Text mining and visualization using VOSviewer
29. vom Brocke J, Simons A, Niehaves B et al. (2009) RECONSTRUCTING THE GIANT: ON THE IMPORTANCE OF RIGOUR IN DOCUMENTING THE LITERATURE SEARCH PROCESS. ECIS 2009 Proceedings
30. Webster J, Watson RT (2002) Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly 26:xiii–xxiii
31. Wu, Y., Iso, H., Pezeshkpour, P., Bhutani, N., & Hruschka, E. (2023, September 14). *Less is More for Long Document Summary Evaluation by LLMs*.
32. Yu D, Bo X (2023) Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling

33. Zhao M, Li F, Cai F et al. (2025) Can we trust LLMs to help us? An examination of the potential use of GPT-4 in generating quality literature reviews 16:128–142. doi: 10.1108/NBRI-12-2023-0115