A Collection of Systematic Reviews in Computer Science

Pierre Achkar[1], Tim Gollub[2] and Martin Potthast[3]

[1]Leipzig University, Fraunhofer ISI Leipzig
[2]Bauhaus-Universität Weimar
[3]Kassel University, hessian.AI, ScaDS.AI

Abstract

Systematic reviews are the standard method for synthesizing scientific evidence, but their creation requires substantial manual effort, particularly during retrieval and screening. While recent work has explored automating these steps, evaluation resources remain largely confined to the biomedical domain, limiting reproducible experimentation in other domains. This paper introduces SR4CS, a large-scale collection of systematic reviews in computer science, designed to support reproducible research on Boolean query generation, retrieval, and screening. The corpus comprises 1,212 systematic reviews with their original expert-designed Boolean search queries, 104,316 resolved references, and structured methodological metadata. For controlled evaluation, the original Boolean queries are additionally provided in a normalized, approximated form operating over titles and abstracts. To illustrate the intended use of the collection, baseline experiments compare the approximated expert Boolean queries with zero-shot LLM-generated Boolean queries, BM25, and dense retrieval under a unified evaluation setting. The results highlight systematic differences in precision, recall, and ranking behavior across retrieval paradigms and expose limitations of naive zero-shot Boolean generation. SR4CS is released under an open license on Zenodo, [1] together with documentation and code,[2] to enable reproducible evaluation and future research on scaling systematic review automation.

Keywords

Systematic reviews, Boolean queries, Information retrieval, Computer science

1. Introduction

Systematic reviews are the standard method for synthesizing evidence on focused research questions. They are valued for their completeness, transparency, and reproducibility, and their results often serve as a guide for future research, policy, and practice [1, 2]. Conducting a systematic review is time-consuming and involves steps such as defining research questions, retrieving candidate papers, screening for relevance, and synthesizing findings [3]. Retrieving candidate papers is a vital step, typically accomplished by translating the review topic into complex Boolean queries to support interpretability and reproducibility. These queries define the pool of eligible candidate studies and directly influence the outcome of the review because any loss in recall during retrieval can hardly be recovered [4].

In recent years, increasing efforts have been made to automate systematic reviews to reduce the cost of their creation, both end-to-end and in individual phases such as query formulation, screening, and synthesis. Work on automatic Boolean query generation ranges from computational adaptations of expert-designed strategies [5] to recent investigations using large language models (LLMs), which show potential but also exhibit substantial trade-offs in precision and recall [6, 7]. LLMs have also been explored for screening, where recall-oriented calibration can yield promising zero-shot performance [7], and agent-based systems are emerging to orchestrate multiple stages of the systematic review workflow [8]. Despite this progress, the reliability and reproducibility of automated methods remain open research questions that require controlled evaluation.

However, evaluation resources for systematic review automation remain heavily concentrated in the biomedical domain. Datasets such as the SIGIR 2017 SysRev Query Collection [9], the CLEF TAR 2019 corpus [10], and the Seed Studies collection [11] provide valuable test collections for medical

[1] https://doi.org/10.5281/zenodo.17163932
[2] https://github.com/webis-de/scolia26-sr4cs

SCOLIA 2026, the Second International Workshop on Scholarly Information Access, ECIR 2026, 2nd April 2026, Delft, Netherland
0009-0007-0791-9078 (P. Achkar); 0000-0003-1737-6517 (T. Gollub); 0000-0003-2451-0665 (M. Potthast)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-03-26

Table 1
Benchmark datasets for systematic review retrieval.

Dataset	Domain	Topics	Reference pool	Tasks
SIGIR 2017 SysRev[1]	Biomedicine	94	26M MEDLINE records	Retrieval; screening
CLEF TAR 2019[2]	Biomedicine	123	PubMed baseline (2019)	Retrieval; screening
Seed-Studies 2022[3]	Biomedicine	40	PubMed retrieved sets + snowballing	Retrieval (seed-based); screening; citation chasing
SR4CS	Computer science	1,212	104k references (89k with abstracts)	Retrieval; screening

systematic reviews, but there is no comparable large-scale resource for other domains. In the absence of domain-specific benchmarks, evaluating retrieval and screening methods often requires costly expert involvement, constraining scale and reproducibility.

To address this gap, we introduce SR4CS, the first large-scale test collection of systematic reviews in computer science, paired with the original Boolean search strategies reported by the authors and curated reference pools. Each review includes the original Boolean query, alongside an approximation normalized to a unified title-and-abstract-only retrieval setting for controlled evaluation. In addition, each review is associated with structured methodological metadata, including research objectives, inclusion and exclusion criteria, databases searched, and temporal constraints. In total, SR4CS contains 1,212 systematic reviews with 104,316 resolved references, of which 89,447 include abstracts. The dataset is publicly available together with documentation and code, enabling reproducible evaluation of Boolean query translation, retrieval effectiveness, and screening methods in computer science.

2. Related Work

Existing evaluation resources for retrieval and screening in systematic reviews have largely focused on the biomedical domain. While these resources vary in scale, design, and supported tasks, they all provide reusable test collections for evaluating retrieval effectiveness, screening, and related automation methods. Table 1 summarizes the most relevant datasets and contrasts them with SR4CS.

The SIGIR 2017 SysRev Query Collection [9] is based on ∼26 million MEDLINE records derived from 94 Cochrane reviews published between 2014 and 2016. The reported search strategies were converted into executable queries, and included and excluded references were mapped to MEDLINE identifiers, enabling controlled experiments on retrieval and screening prioritization in the biomedical domain.

The Seed Studies Collection [11] was developed to support research on seed-driven search strategies. It comprises 40 medical systematic review topics curated by information specialists and includes the original PubMed Boolean queries, seed studies, retrieved records, and final included references. By explicitly distinguishing genuine seed studies from pseudo-seeds, the collection enables more realistic evaluation of automatic query formulation, screening prioritization, and citation chasing.

The CLEF TAR 2019 corpus [10], developed as part of the CLEF lab on technology-assisted reviews, is also grounded in Cochrane reviews and PubMed. It includes expert queries, retrieved records, and relevance judgments for 123 topics across several medical review types, and supports evaluation of protocol querying and screening prioritization.

SR4CS extends these efforts beyond biomedicine by providing the first large-scale test collection of systematic reviews in computer science. It preserves the original Boolean search strategies as reported in the reviews, while also providing normalized, approximated variants for controlled evaluation. With 1,212 systematic reviews and resolved reference pools, SR4CS supports reproducible evaluation of

[3] https://github.com/ielab/SIGIR2017-SysRev-Collection
[4] https://github.com/ielab/sysrev-seed-collection
[5] https://github.com/CLEF-TAR/tar

Boolean query translation, retrieval effectiveness across paradigms, and screening methods in a domain where comparable evaluation resources have previously been lacking.

3. The SR4CS Corpus

The corpus construction comprises three main stages: data collection and filtering, parsing and information extraction, and resolving references. Figure 1 provides an overview of the pipeline.

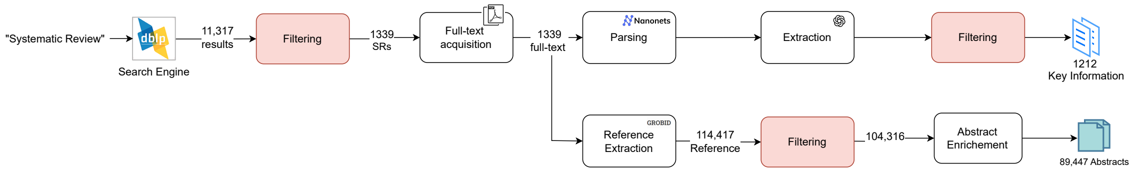

Figure 1: SR4CS-25 construction pipeline: collection, filtering, parsing, extraction, and reference resolution.

3.1. Data Collection

A set of candidate systematic reviews was retrieved from DBLP by searching titles for “systematic review”, yielding 11,317 results. We applied filter rules to retain only genuine systematic reviews that required clear wording (e.g., “systematic review of/on”), peer review procedures, open access, and valid DOIs. Outliers due to extreme page counts were removed. After filtering, 1,339 candidate reviews remained. In the following steps, only reviews that explicitly stated their Boolean queries were considered for inclusion in the final dataset.

3.2. Parsing and Extraction

The PDF files of the systematic reviews were downloaded and converted into Markdown format for further processing. To achieve the conversion, `Nanonets-OCR-s`,[6] a modern OCR model that produces semantically enriched Markdown text rather than plain text, was used. Unlike traditional OCR tools, it can recognize LaTeX equations, describe figures and tables, and preserve document structure. This functionality was particularly important because many reviews contain Boolean queries and other important information in figures or tables.

The documents were then processed in a zero-shot settings using GPT-4.1 Mini to extract a series of structured fields: (i) databases used, (ii) Boolean queries reported, (iii) year range, (iv) language restrictions, (v) inclusion and exclusion criteria, (vi) main topic, (vii) objective, (viii) research questions, and (ix) whether snowballing or citation chasing was performed. This approach follows recent work demonstrating the feasibility of LLM-based zero-shot data extraction in scientific documents [12, 13].

To assess extraction quality, the extracted fields were manually inspected for errors and inconsistencies. Additionally, we randomly selected 10% of the reviews and manually extracted the same information independently. In 85% of the sample cases, the fields extracted by the LLM matched the manual extraction at the field level, and the errors found were minimal, such as minor discrepancies in the wording of the inclusion/exclusion criteria or the specification of certain field values (e.g., language restrictions) based on indirect clues rather than explicit statements. These findings indicate that zero-shot LLM-based extraction provides a sufficiently reliable basis for large-scale corpus construction, while still leaving room for more robust, feedback-driven extraction methods in future extensions. Finally, we excluded reviews that did not explicitly state their Boolean queries, resulting in a final dataset of 1,212 systematic reviews.

[6] https://huggingface.co/nanonets/Nanonets-OCR-s

Table 2
Core statistics of the SR4CS corpus.

Statistic	Value
Systematic reviews	1,212
Total resolved references	104,316
References with abstracts	89,447
Average references per review	∼83
Median references per review	74
Minimum references per review	4
Maximum references per review	459

Table 3
Most frequently used bibliographic databases in SR4CS.

Database	Count
Scopus	649
ScienceDirect	478
Web of Science	418
ACM Digital Library	439
IEEE Xplore	387
Google Scholar	346
PubMed	150

3.3. Reference Resolution

References were extracted from each systematic review via a hybrid pipeline combining AnyStyle[7] and GROBID,[8] a configuration proven to be highly effective for structured citation analysis [14]. First, non-scientific items such as reports, interviews, and patents were removed, excluding 10,101 entries. The remaining references were enriched with abstracts from public bibliographic APIs, including Crossref, OpenAlex, Semantic Scholar, PubMed, Europe PMC, and the arXiv API. During resolution, DOI matches were prioritized, while references without a DOI were handled through a high-threshold title-author match and consistency checks for the publication year to minimize false positives. Overall, this process produced a consolidated reference pool on which the SR4CS corpus is based.

3.4. Corpus Statistics

Table 2 summarizes the size and structural properties of the SR4CS corpus, providing context for the retrieval and screening experiments, while Table 3 reports the bibliographic databases most frequently cited in the reviewed systematic reviews, reflecting common data sources used in computer science literature reviews.

4. Experiments and Intended Use

The SR4CS corpus is released to support reproducible research on retrieval and screening for systematic reviews in computer science. The dataset is designed for three main tasks: (i) evaluating automatic Boolean query generation against approximations of expert-designed Boolean queries, (ii) comparing retrieval paradigms under realistic systematic review topics, and (iii) investigating screening behavior using curated reference pools. To illustrate its use, we present baseline experiments for Tasks (i) and (ii), evaluating Boolean, probabilistic (BM25), and dense retrieval methods using titles and abstracts only.

[7] https://github.com/inukshuk/anystyle
[8] https://github.com/kermitt2/grobid

Table 4
Screening effectiveness of approximated expert-designed and automated retrieval approaches. Non-Boolean methods are capped at the top 1,000 retrieved documents.

Method	Precision	Recall	F_1	F_3
Expert Boolean (Approx.)	**0.352**	0.342	**0.224**	0.241
GPT4.1-Mini Boolean (ZS)	0.298	0.099	0.095	0.085
BM25 (Title+Obj, Top-1k)	0.041	0.512	0.074	0.223
Dense (Title+Obj, MiniLM)	0.057	**0.702**	0.104	**0.309**

Table 5
Ranking effectiveness of approximated expert-designed and automated retrieval approaches on SR4CS-25. Boolean result sets are deterministically ordered using BM25 for ranking-based evaluation.

Method	MAP	P@10	R@100
Expert Boolean (Approx.)	0.173	0.490	0.238
GPT4.1-Mini Boolean (ZS)	0.054	0.256	0.083
BM25 (Title+Obj)	0.175	0.432	0.262
Dense (Title+Obj)	**0.281**	**0.545**	**0.373**

4.1. Approximating Expert-Designed Boolean Queries

A Boolean retrieval baseline is constructed by approximating the expert-designed queries reported in the original reviews. Since these queries rely on database-specific syntax and metadata fields, normalization is required to obtain a unified Boolean formulation operating over document titles and abstracts.

Query rewriting. Reported Boolean queries are converted into valid SQLite FTS5 `MATCH` syntax using `GPT-4.1-mini` in a zero-shot setting. The query structure is preserved, restricted to the `title` and `abstract` fields, and stripped of unsupported metadata filters. A random 25% sample is manually verified, with only seven queries requiring minor corrections, primarily for verbatim phrases.

Index and execution. A SQLite FTS5 index is created using the `unicode61` tokenizer with diacritics folded and prefix indexing (2–10) to account for common spelling and formatting variations. All refined queries are executed against this index; when multiple Boolean queries are associated with a review, their result sets are merged via union.

4.2. Alternative Retrieval Baselines

Using the Boolean approximation as a reference point, we evaluate three additional baselines operating on the same corpus and fields: (i) zero-shot Boolean queries generated from the review title and objective using `GPT-4.1-mini`, (ii) a keyword-based BM25 baseline, and (iii) a dense semantic retriever based on the `all-MiniLM-L6-v2` Sentence-BERT model.

For both BM25 and dense retrieval, queries are constructed from the review title, the objective, and their concatenation. Results are reported for the best-performing variant (title + objective). Retrieval is capped at 1,000 documents per query to reflect a fixed screening budget and ensure comparability across retrieval methods.

4.3. Evaluation and Results

All retrieval outputs are compared against the curated reference pools to compute precision, recall, F_1, and F_3, macro-averaged across reviews. For ranked retrieval methods, we additionally report MAP, P@10, and R@100.

Although Boolean retrieval is inherently unranked, Boolean result sets are deterministically ranked using SQLite's internal BM25 scoring function for the purpose of ranking-based evaluation.

Table 4 reports set-based screening effectiveness, while Table 5 reports ranking effectiveness across all evaluated methods. Together, these results provide a consistent and reproducible reference point for evaluating Boolean query generation and alternative retrieval paradigms within the SR4CS corpus.

5. Discussion

The development of SR4CS builds on recent advances in document processing and natural language technologies for large-scale extraction from scientific publications. OCR parsers based on vision–language models improve handling of complex PDF layouts [15], LLMs support the extraction of methodological information from scientific texts [13, 16], and GROBID continues to perform strongly in reference extraction [14]. By integrating these components into a unified pipeline, SR4CS demonstrates a reproducible approach to constructing systematic review corpora and provides a transferable methodology for building evaluation resources across scientific domains.

The retrieval results clarify how different paradigms behave when evaluated under the same title-and-abstract-only constraints imposed during corpus construction. Approximated versions of expert-designed Boolean queries achieve the highest precision (0.352), confirming that Boolean logic remains the most effective means of controlling screening effort; the moderate recall (0.342) primarily reflects the approximation process, which strips database-specific operators, controlled vocabularies, and citation-based expansion techniques rather than shortcomings of the original strategies. Dense retrieval, in contrast, delivers substantially higher recall (0.702) and the strongest ranking effectiveness (MAP 0.281, P@10 0.545), demonstrating robustness to vocabulary mismatch and terminological variation in metadata-constrained settings. The zero-shot GPT-4.1- Mini Boolean baseline performs poorly across all metrics (recall 0.099), providing a clear negative result: naive zero-shot Boolean generation with an off-the-shelf model fails to recover either the coverage or selectivity of expert-designed queries, indicating the need for more structured and adaptive query construction approaches. Overall, these results establish SR4CS as a targeted diagnostic benchmark for exposing precision–recall trade-offs and failure modes in retrieval and automated query generation.

Several directions for future extensions of SR4CS are possible. First, deeper integration with open scholarly infrastructures such as OpenAlex would support reference coverage, metadata enrichment, and query execution within an open and reproducible ecosystem, enabling end-to-end studies that rely on open identifiers and transparent metadata. Second, extending the corpus-construction workflow to additional domains beyond computer science (e.g., education and social science) would broaden the benchmark's scope and help assess how retrieval and query-generation methods transfer across differences in terminology and reporting conventions. Third, future work could explore agentic LLM-based information extraction frameworks that use iterative feedback and explicit grounding in source text to improve extraction fidelity and consistency [17, 18]; such approaches may also make it easier to attach verifiable evidence to extracted fields and to diagnose error modes in downstream evaluation.

6. Conclusion

We introduced SR4CS, a large-scale test collection of systematic reviews in computer science, designed to support reproducible research on Boolean query generation, retrieval, and screening. By providing approximated versions of expert-designed Boolean queries, curated reference pools, and a controlled evaluation setting over titles and abstracts, SR4CS enables systematic comparisons of retrieval paradigms and query formulation strategies beyond the biomedical domain. The dataset is released with documentation and code to facilitate reuse, benchmarking, and future work on systematic review automation, retrieval evaluation, and query generation in computer science and related fields.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL Write for sentence polishing and grammar correction.

References

[1] A. Liberati, D. Altman, J. Tetzlaff, C. Mulrow, P. Gøtzsche, J. Ioannidis, M. Clarke, M. Clarke, P. Devereaux, J. Kleijnen, D. Moher, The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration, PLoS Med. (2009). doi:doi:10.1371/journal.pmed.100010.

[2] G. Lamé, Systematic literature reviews: An introduction, Proc. of Design Soc.: Int. Conf. on Engineering Design (2019). doi:10.1017/dsi.2019.169.

[3] C. Lefebvre, J. Glanville, S. Briscoe, A. Littlewood, C. Marshall, M.-I. Metzendorf, A. Noel-Storr, T. Rader, F. Shokraneh, J. Thomas, L. S. Wieland, Searching for and selecting studies, John Wiley & Sons, Ltd, 2019. doi:https://doi.org/10.1002/9781119536604.ch4. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119536604.ch4.

[4] A. MacFarlane, T. Russell-Rose, F. Shokraneh, Search strategy formulation for systematic reviews: Issues, challenges and opportunities, Intel. Sys. with Applications (2022). doi:https://doi.org/10.1016/j.iswa.2022.200091.

[5] H. Scells, G. Zuccon, B. Koopman, A comparison of automatic boolean query formulation for systematic reviews, Inf. Retr. J. (2021). doi:10.1007/S10791-020-09381-1.

[6] S. Wang, H. Scells, B. Koopman, G. Zuccon, Can chatgpt write a good boolean query for systematic review literature search?, in: Proc. of SIGIR 2023, ACM, 2023. doi:10.1145/3539618.3591703.

[7] S. Wang, H. Scells, S. Zhuang, M. Potthast, B. Koopman, G. Zuccon, Zero-shot generative large language models for systematic review screening automation, in: Proc. of ECIR 2024, LNCS, Springer, 2024. doi:10.1007/978-3-031-56027-9_25.

[8] M. A. Sami, Z. Rasheed, K. Kemell, M. Waseem, T. Kilamo, M. Saari, A. Nguyen-Duc, K. Systä, P. Abrahamsson, System for systematic literature review using multiple AI agents: Concept and an empirical evaluation, CoRR (2024). doi:10.48550/ARXIV.2403.08399. arXiv:2403.08399.

[9] H. Scells, G. Zuccon, B. Koopman, A. Deacon, L. Azzopardi, S. Geva, A test collection for evaluating retrieval of studies for inclusion in systematic reviews, in: Proc. of SIGIR 2017, ACM, 2017. doi:10.1145/3077136.3080707.

[10] E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, CLEF 2019 technology assisted reviews in empirical medicine overview, in: W.N. of CLEF 2019, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_250.pdf.

[11] S. Wang, H. Scells, J. Clark, B. Koopman, G. Zuccon, From little things big things grow: A collection with seed studies for medical systematic review literature search, in: Proc. of SIGIR 2022, ACM, 2022. doi:10.1145/3477495.3531748.

[12] M. P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering - example of chatgpt, CoRR (2023). doi:10.48550/ARXIV.2303.05352. arXiv:2303.05352.

[13] G. Gartlehner, L. Kahwati, R. Hilscher, I. Thomas, S. Kugley, K. Crotty, M. Viswanathan, B. Nussbaumer-Streit, G. Booth, N. Erskine, A. Konet, R. Chew, Data extraction for evidence synthesis using a large language model: A proof-of-concept study, Research Synthesis Methods (2024). doi:https://doi.org/10.1002/jrsm.1710. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1710.

[14] T. Backes, A. Iurshina, M. A. Shahid, P. Mayr, Comparing free reference extraction pipelines, Int. J. Digit. Libr. (2024). doi:10.1007/S00799-024-00404-6.

[15] J. Poznanski, A. Rangapur, J. Borchardt, J. Dunkelberger, R. Huff, D. Lin, A. Rangapur, C. Wilhelm, K. Lo, L. Soldaini, olmocr: Unlocking trillions of tokens in pdfs with vision language models, 2025. doi:10.48550/ARXIV.2502.18443. arXiv:2502.18443.

[16] H. Lai, J. Liu, C. Bai, H. Liu, B. Pan, X. Luo, L. Hou, W. Zhao, D. Xia, J. Tian, Y. Chen, L. Zhang, J. Estill, J. Liu, X. Liao, N. Shi, X. Sun, H. Shang, Z. Bian, K. Yang, L. Huang, L. Ge, H. Li, Y. Wang, H. Zhang, D. Zhu, D. Peng, F. Wang, Y. Li, S. Tang, H. Liu, Z. Li, Z. Yang, X. Yu, Y. Qin, Language models for data extraction and risk of bias assessment in complementary medicine, npj Digit. Medicine (2025). doi:10.1038/S41746-025-01457-W.

[17] Z. Li, Y. Yu, W. Gu, T. Zhu, H. Song, W. Guo, X. Yang, Z. Zhu, Dual-llm adversarial framework for information extraction from research literature, bioRxiv (2025). URL: https://www.biorxiv.org/content/early/2025/09/16/2025.09.11.675507. doi:10.1101/2025.09.11.675507. arXiv:https://www.biorxiv.org/content/early/2025/09/16/2025.09.11.675507.full.pdf.

[18] J. Barrow, R. Patel, M. Kharkovski, B. Davies, R. Schmitt, Safepassage: High-fidelity information extraction with black box llms, CoRR abs/2510.00276 (2025). URL: https://doi.org/10.48550/arXiv.2510.00276. doi:10.48550/ARXIV.2510.00276. arXiv:2510.00276.