

Leveraging Cochrane Systematic Literature Reviews for Prospective Evaluation of Large Language Models

Wojciech Kusa¹, Harrison Scells², Moritz Staudinger¹ and Allan Hanbury¹

¹TU Wien, Vienna, Austria

²Leipzig University, Leipzig, Germany

Abstract

While systematic literature reviews are central to evidence-based medicine, creating them takes significant time and effort. As such, numerous efforts have been dedicated to automating various aspects of systematic review creation. One key technology being applied to automating systematic reviews is large language models (LLMs). However, evaluating methods that use LLMs poses one glaring risk: it is often unknown whether the LLM was trained on systematic reviews, the data used to evaluate many areas of systematic review automation. We propose a conceptual framework for constructing a new dataset based on the Cochrane Database of Systematic Reviews. We envision a dataset to enable prospective evaluation of large language models in an end-to-end systematic literature review automation task. In essence, we provide a way to evaluate systematic review automation methods that strictly guarantees no train-test leakage. This paper highlights limitations in current LLM evaluation methodologies by advocating for a real-world, evolving and dynamic dataset. We aim to mitigate data contamination and prompt sensitivity through prospective evaluation.

Keywords

Prospective Evaluation, Large Language Models, Systematic Literature Reviews, Citation Screening, Data Contamination

1. Introduction

Systematic literature reviews (SLRs) are central to evidence-based medicine, informing clinical practice and policy [1]. They have a well-established and rigorous methodology for synthesising and evaluating the evidence on a specific research question. Despite their importance, creating an SLR is slow and labour-intensive, often taking months or years due to the amount of literature that needs assessing and analysing, making them an ideal candidate for automation. Current efforts in automating SLRs have focused primarily on individual processes, such as search query formulation [2, 3, 4], query refinement [5, 6], document screening prioritisation [7, 8], screening cut-off prediction [9, 10], data extraction [11, 12] or evidence summarisation [13, 14].

To this date, prospective evaluation of SLR automation was limited to single reviews conducted frequently by biomedical and healthcare researchers, focusing on commercial tools or small-scale experimental setups. This was due primarily to the cost of preparing and annotating the necessary examples. On the other hand, the typical evaluation of ML algorithms in the tasks mentioned above relies on rather simplified and siloed datasets, with previous research raising concerns about issues such as annotation quality, data overlap, and even data leakage [15, 16].

Large Language Models (LLMs) provide a path towards end-to-end SLR automation, which can significantly speed-up the generation of SLRs. However, evaluating LLMs in this context has several challenges, such as reduced reproducibility, problems with data contamination, the need to adapt to rapid changes in evidence, the occurrence of hallucinations, and the imperative to ensure high Recall. This paper highlights our conceptual framework for evaluating LLMs for SLR automation tasks and

3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (ALTARS 2024): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems. March 28, 2024, Glasgow, UK

✉ wojciech.kusa@tuwien.ac.at (W. Kusa); harrison.scells@uni-leipzig.de (H. Scells); moritz.staudinger@tuwien.ac.at (M. Staudinger); allan.hanbury@tuwien.ac.at (A. Hanbury)

🌐 <https://wojciechkusa.github.io> (W. Kusa); <https://informatics.tuwien.ac.at/people/allan-hanbury> (A. Hanbury)

🆔 0000-0003-4420-4147 (W. Kusa); 0000-0001-9578-7157 (H. Scells); 0000-0002-5164-2690 (M. Staudinger);

0000-0002-7149-5843 (A. Hanbury)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

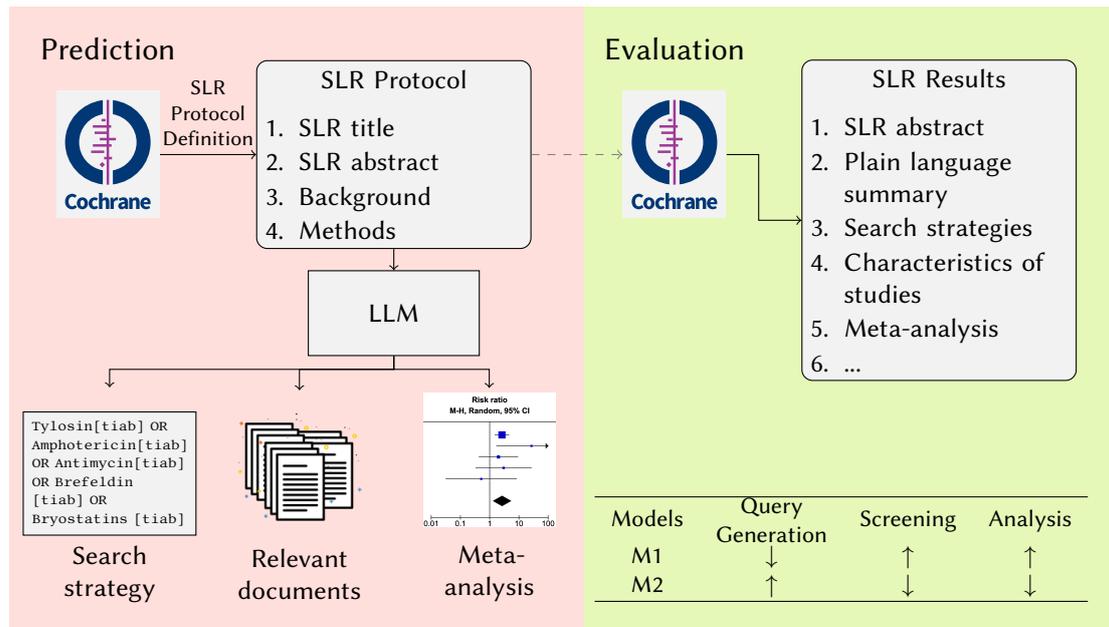


Figure 1: A framework for real-time LLM evaluation using a prospective dataset from Cochrane Systematic Review protocols. Initially, only the SLR protocol (title, abstract, and background) is available. This input informs LLM predictions for each SLR step, including search strategy, identifying relevant documents, and conducting meta-analysis. Results are then compared with manual reviews from the Cochrane Library.

how we aim to extend it for end-to-end automation while also preventing data contamination. This framework provides the groundwork for evaluating LLMs in the biomedical domain, featuring multiple NLP and retrieval tasks and focusing on an evolving dataset. All these features allow for a relatively annotation-free prospective assessment of the effectiveness of LLMs in SLR automation.

2. Design of Prospective Evaluation Dataset

Current LLM evaluation methods often do not adequately reflect the creation of SLRs, mainly due to the reliance on retrospective data, leading to issues such as data contamination. Our dataset creation framework (Figure 1) can address this issue by enabling the evaluation of newly published Cochrane SLR protocols, effectively mitigating contamination. Cochrane is an organisation that manually collects, synthesises, and disseminates medical evidence to aid in making informed decisions regarding health treatments and policies.¹

We envisage creating ‘evaluation sandboxes’ where LLMs can be evaluated in real-time against the gold standard data available after the completion of manual reviews. We plan to use the TIREx platform [17] as the online model submission platform. Additionally, we envision using the CSMed framework [15] and the ReviewManager (RevMan)² format published by the Cochrane Library as the basis for dynamically creating the dataset. With these properties, we intend to publicly release dataset snapshots of the so-called ‘knowledge cutoffs’ on a regular basis. Finally, we aim to include several under-investigated SLR methodologies in the dataset, such as qualitative reviews (analysing based on characteristics such as interviews or focus groups) and prognosis reviews (analysing based on characteristics such as demographic or lifestyle factors) to improve the generalisability of LLMs on different types of SLRs.

Our task begins with the SLR protocol, including the title, abstract, and eligibility criteria, and aims at predicting the entirety of an SLR’s future components. Specifically, the envisioned tasks that our prospective evaluation dataset would encompass includes:

¹<https://www.cochrane.org/>

²<https://training.cochrane.org/online-learning/core-software/revman>

- generating a search strategy through Boolean queries,
- identifying relevant publications,
- extracting PICO (population, intervention, comparison, and outcome) elements, and
- calculating meta-analysis outcomes.

Historically, each of these steps has been evaluated independently. However, the advancement of LLMs enables us to test this as an integrated end-to-end approach, offering a comprehensive solution to automate the SLR process effectively. This holistic approach aims to comprehensively automate the SLR process, addressing the need for efficient evidence synthesis in evidence-based medicine.

Recent studies[18, 19] have found many benchmark datasets already compromised. The CONDA database³ is one recent example of a community effort to try to keep track of the contamination of datasets in LLMs. Our prospective evaluation framework prevents contamination by using recently finished SLR protocols for evaluation, which were only published after the knowledge cutoff date of the LLM it evaluates. Therefore, no test collection data can already be present in the pretraining data of the LLM.

3. Limitations and Future Directions

One of the biggest limitations is that the time between SLR registration and publication is typically around two years for Cochrane SLRs [20]. One way to mitigate this problem could be to use SLR which protocols were registered some time ago but still do not have the review available (as we expect that they are close to the first publication). For instance, 166 registered reviews in 2022 still have not published their results by June 2024, and we can assume that many of them will publish the final review in the next six months (by the end of 2024).⁴

Future directions of research using our prospective dataset include:

1. Extension of the dataset to non-biomedical SLRs, e.g., social science data,⁵ to assess the LLMs' abilities in other contexts;
2. Multi-dimensional evaluation metrics: new metrics that measure aspects beyond Recall, such as contextual understanding and outcomes of systematic reviews [21];
3. Metrics for adaptive learning: new metrics to evaluate how well LLMs adapt to new evidence;
4. Bias detection and mitigation protocols: new methods to identify and address biases in the dataset and the LLM outputs;
5. Finally, this dataset could also be expanded to focus on predicting SLR updates published by Cochrane, and living systematic reviews.

4. Conclusion

Our conceptual framework proposes a way to evaluate LLMs for SLR automation that mitigates data contamination. By leveraging prospective Cochrane reviews for forward prediction, we address key challenges in current evaluation practices. Furthermore, as the prediction task is the end-to-end generation of SLR, we believe that this work could lay the groundwork for more effective IR and NLP systems in the biomedical and healthcare domains. Finally, this framework allows for the creation of a relatively annotation-free evolving test collection.

³<https://huggingface.co/spaces/CONDA-Workshop/Data-Contamination-Database>

⁴<https://www.cochranelibrary.com/cdsr/reviews>

⁵Using SLRs created by the Campbell Collaboration.

Acknowledgments

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

References

- [1] A. Jo, A.-I. Raquel, B. Jane, C. Duncan, E. Alison, F. Debra, H. Susanne, L. Kate, R. Stephen, R. Amber, S. Lesley, S. Christian, W. Paul, W. Nerys, Systematic Reviews: CRD’s guidance for undertaking reviews in health care, CRD, University of York, York, 2009.
- [2] H. Scells, G. Zuccon, B. Koopman, A comparison of automatic boolean query formulation for systematic reviews, *Information Retrieval Journal* 24 (2021) 3–28.
- [3] S. Wang, H. Scells, B. Koopman, G. Zuccon, Can ChatGPT write a good boolean query for systematic review literature search?, *arXiv preprint arXiv:2302.03495* (2023).
- [4] M. Staudinger, W. Kusa, F. Piroi, A. Lipani, A. Hanbury, A reproducibility and generalizability study of large language models for query generation, in: *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP ’24)*, 2024. doi:10.1145/3673791.3698432.
- [5] H. Scells, G. Zuccon, B. Koopman, Automatic boolean query refinement for systematic review literature search, *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019 11* (2019) 1646–1656. URL: <https://doi.org/10.1145/3308558.3313544>. doi:10.1145/3308558.3313544.
- [6] A. Alharbi, M. Stevenson, Refining Boolean queries to identify relevant studies for systematic review updates, *Journal of the American Medical Informatics Association* 27 (2020) 1658–1666. URL: <https://doi.org/10.1093/jamia/ocaa148>. doi:10.1093/jamia/ocaa148. arXiv:<https://academic.oup.com/jamia/article-pdf/27/11/1658/34363868/ocaa148.pdf>.
- [7] W. Kusa, A. Hanbury, P. Knoth, Automation of citation screening for systematic literature reviews using neural networks: A replicability study, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørnvåg, V. Setty (Eds.), *Advances in Information Retrieval*, Springer International Publishing, Cham, 2022, pp. 584–598. URL: https://doi.org/10.1007/978-3-030-99736-6_39.
- [8] A. M. Cohen, K. Ambert, M. McDonagh, A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review, in: *AMIA annual symposium proceedings*, volume 2010, American Medical Informatics Association, 2010.
- [9] M. Stevenson, R. Bin-Hezam, Stopping methods for technology assisted reviews based on point processes, *ACM Transactions on Information Systems* (2023).
- [10] S. Wang, H. Scells, S. Zhuang, M. Potthast, B. Koopman, G. Zuccon, Zero-shot generative large language models for systematic review screening automation, in: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I*, volume 14608 of *Lecture Notes in Computer Science*, 2024, pp. 403–420.
- [11] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, B. C. Wallace, A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature, in: *Proceedings of ACL*, 2018.
- [12] A. Dhrangadhariya, H. Müller, Not so weak PICO: leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation, *JAMIA open* 6 (2023).
- [13] L. L. Wang, J. DeYoung, B. Wallace, Overview of MSLR2022: A shared task on multi-document summarization for literature reviews, in: *Proceedings of the Third Workshop on Scholarly Document Processing*, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 175–180. URL: <https://aclanthology.org/2022.sdp-1.20>.
- [14] C. Shaib, M. L. Li, S. Joseph, I. J. Marshall, J. J. Li, B. C. Wallace, Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success), in: *Proceedings of the*

- 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 1387–1407.
- [15] W. Kusa, Ó. E. Mendoza, M. Samwald, P. Knoth, A. Hanbury, CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews, in: 37th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks, 2023.
- [16] A. Dhrangadhariya, H. Müller, DISTANT-CTO: A zero cost, distantly supervised approach to improve low-resource entity extraction using clinical trials literature, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 345–358. URL: <https://aclanthology.org/2022.bionlp-1.34>. doi:10.18653/v1/2022.bionlp-1.34.
- [17] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The information retrieval experiment platform, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2826–2836.
- [18] S. Balloccu, P. Schmidová, M. Lango, O. Dusek, Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian’s, Malta, 2024, pp. 67–93. URL: <https://aclanthology.org/2024.eacl-long.5>.
- [19] O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, E. Agirre, NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 10776–10787. URL: <https://aclanthology.org/2023.findings-emnlp.722>. doi:10.18653/v1/2023.findings-emnlp.722.
- [20] M. Sampson, K. G. Shojania, C. Garritty, T. Horsley, M. Ocampo, D. Moher, Systematic reviews can be produced and published faster, *Journal of clinical epidemiology* 61 (2008) 531–536.
- [21] W. Kusa, G. Zuccon, P. Knoth, A. Hanbury, Outcome-based evaluation of systematic review automation, in: Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR ’23), ACM, Taipei, Taiwan, 2023, p. 9. URL: <https://doi.org/10.1145/3578337.3605135>. doi:10.1145/3578337.3605135.