The Use of Large Language Models for Systematic Literature Review Automation: An Evaluation of Quality and Time Savings

Poster #MSR41

Ryan Thaliffdeen[1], Meelis Lootus[2], Iradj Reza[3], Carrie Nielson[4], Lulu Zhao Beatson[2], Harriet Dickinson[5]

[1] Global Health Economics and Outcomes Research, Gilead Sciences, Foster City, CA, USA
[2] Tehistark, London, UK
[3] Library & Information Services, Gilead Sciences, Uxbridge, UK
[4] Real World Evidence, Gilead Sciences, Foster City, USA
[5] Real World Evidence, Gilead Sciences, Uxbridge, UK

Key Takeaways

- This study performed full end-to-end automation of a human-made systematic literature review (SLR) with LLMs and compared their results step by step. It was possible to use LLMs to complete a SLR fully end-to-end (including writing).
- Human input had a substantial effect on the quality of the LLM-SLR, with key points for providing feedback to the model. It is possible to direct the model to include specific sections in the report.
- In the extraction tasks, the LLMs were most proficient at extracting treatment characteristics and were less effective at extracting patient characteristics.
- The use of the LLMs to complete a SLR resulted in considerable time savings – 20.3% for title-abstract screening, 61.8% for full-text screening, and 55.6% for extraction.

Introduction

- Systematic literature reviews (SLRs) are a type of study design that can be used to support drug development and evaluation activities.
- SLRs involve identifying, selecting and critically appraising published scientific research to address a well-defined research question. The SLR process should be transparent, rigorous and reproducible. Whilst valuable, SLRs are costly, time-consuming, and their results can become quickly outdated.
- This research sought to understand if Large Language Models (LLMs) could be used to automate the entire end-to-end SLR process.
- LLMs have previously demonstrated potential in supporting isolated components of systematic literature reviews (SLRs)[1,2], but a comprehensive assessment of their capabilities across the entire process is lacking (e.g., evidence is still emerging around the value of LLMs in SLR report writing).

References 1. Saied, Mohamed, et al. "AI in Literature Reviews: a survey of current and emerging methods." 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). IEEE, 2024.**2**. Bolanos, Francisco, et al. "Artificial intelligence for literature reviews: Opportunities and challenges." Artificial Intelligence Review 57.10 (2024): 259. **3.** Susnjak, Teo, et al. "Automating research synthesis with domain-specific large language model fine-tuning." ACM Transactions on Knowledge Discovery from Data (2024).

Correspondence: Harriet Dickinson, harriet.dickinson@gilead.com.

Disclosures: This work was fully funded by Gilead Sciences, Inc.
HD, CN, RT, IR, are employees of Gilead Sciences, Inc., and may own stock in Gilead Sciences, Inc. **ML** is a director and shareholder of Tehistark. **LZB** is a former employee of Tehistark and may own stock in Tehistark.

Methods

We replicated an existing human-made SLR using the AutoSLR platform (**Figure 1**) in three different ways (**Table 1**), using the gpt-4o-2024-08-06 LLM in all the steps:

1. A fully automated SLR (**FullAuto**)
2. A SLR with a human-defined search approach (**PartAuto1**)
3. A SLR with a human-defined search and screening process with writing feedback (**PartAuto2**)

Evaluation

- The performance of AutoSLR was evaluated across four SLR activities (search, screening, data extraction, report writing) and compared to the human-made SLR.

Evaluation sets

- The evaluation sets were constructed as follows:
  - **TA100**: consisting 100 title-abstract pairs (36 manually labelled as included and 64 excluded).
  - **FT77**: consisting 77 full-texts (15 manually labelled as included and 62 as excluded).
  - **E15**:15 articles with 31 columns selected for extraction, with the gold standard being the human SLR.

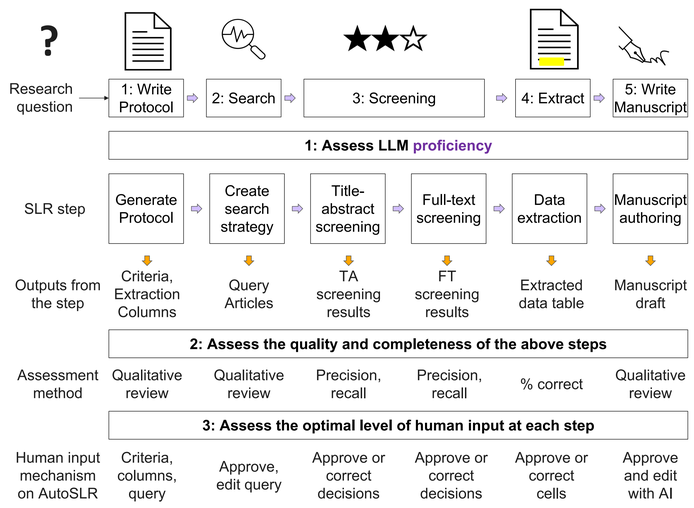

Figure 1. The SLR steps automated in this study, and assessment methods used.

	(1) Original human-made SLR	(2) **FullAuto** (29th Oct 2024)	(3) **PartAuto1** (11th Nov 2024)	(4) **PartAuto2**: (21st Nov 2024)
1: Protocol	Human	AI	Human	Human
2: Search	Human	AI	Human	Human
3: Screen	Human	AI	AI	Human
4: Extract	Human	AI	AI	AI
5: Write	Human	AI	AI	AI + Human

Table 1. Comparison of pipelines: (1) original SLR, (2) **FullAuto**, (3) **PartAuto1**, (4) **PartAuto2**, along with completion dates.

- The quality of reports was qualitatively assessed by both an SLR expert and an HEOR expert.
- Labour savings were measured between fully manual screening and extraction into a spreadsheet vs. reviewing and fixing the AI results in the AutoSLR app.
- For FullAuto, only the general research question was provided to AutoSLR. AutoSLR generated the search strategy and used PubMed API to find and retrieve articles.
- For PartAuto1, the AutoSLR was provided a search string for use with searching scientific literature.
- In PartAuto2, report writing was iterated using direct human-in-the-loop feedback.

Results

Evaluation of Fully Automated SLR

- AutoSLR was able to produce an end-to-end SLR fully automatically.
- The search query was evaluated as basic and lacked some relevant synonyms and controlled vocabulary terms.
- Title-abstract screening achieved 72% accuracy (62.5% precision, 55.6% recall, 58.8% F1-score) compared to a human-made review on TA100, while full-text screening achieved 70.1% accuracy (33.3% precision, 53.3% recall, 41.0% F1-score) compared to a human-made review on FT77.
- AutoSLR was able to extract data relating to 117 different columns. It achieved good/perfect extraction in 71.6% of extractions.
- Performance was best for treatment characteristics (86.7% good/perfect) and worst for patient characteristics (54.7% good/perfect) (see **Table 2**)].
- The initial FullAutoSLR written report lacked some common SLR sections, a PRISMA diagram, and overall depth.
- The final report partially met the needs of the initial research question, with strong introduction and methods section, but too high-level results and discussion sections.

	Manual vs Fully Automated SLR				
	Perfect	Good	Partial	Wrong / Missing	Extra from LLM
Overall (31 columns)	57.2%	14.4%	7.7%	14.6%	6.1%
Study Characteristics (17 columns)	55.2%	20.8%	8.6%	13.1%	2.3%
Treatment Characteristics (4 columns)	68.8%	17.9%	10.7%	2.7%	0.0%
Patient Characteristics (6 columns)	50.0%	4.7%	5.3%	28.0%	12.0%
Outcomes (4 columns)	66.7%	0.0%	0.0%	0.0%	33.3%

Table 2. Data extraction quality by column type.

Human in the loop iterations:

- By specifying search strategy, screening strategy, and writing input, the quality of the report could be substantially improved.
- Human input was particularly impactful in creating a more comprehensive abstract summary, more structured discussion (basic to thematic organization covering efficacy, safety and limitations), and more robust introduction (2-4 paragraphs) whilst maintaining consistent conclusion quality.
- Compared to human-only SLRs, the synergistic AI-human approach was 20.3% faster for title-abstract screening, 61.8% for full-text screening, and 55.6% for data extraction (**Figure 2**).
- AutoSLR's audit trail functionality and fluid correction mechanisms proved valuable for ensuring high review quality.

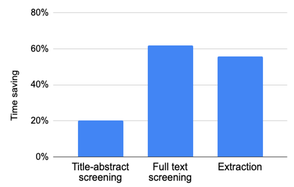

Figure 2. Time savings associated with the use of LLMs in a SLR study (compared to a human-made SLR), where human time spent on human-in-the-loop feedback was included.

Limitations

- This study replicated a human-made SLR covering an oncology research question. It should be noted that SLRs covering other therapy areas may produce different results.
- A human-made SLR does not constitute a true ground truth, and this work measures closeness to the human review in quantitative assessments.
- Cost savings associated with LLM use are challenging to assess, but due to the scalable nature of this tool and the speed to completion, costs are likely to be substantially cheaper.
- It was not investigated why the LLM was better at completing some extraction tasks than others, and further research is needed in this area.

Conclusion

- LLMs are able to conduct SLRs end-to-end but require human input to achieve high quality. This finding is consistent with other research[3], which highlights the value of human and domain expertise input in generating content that meets stakeholder needs and is consistent with PRISMA reporting guidelines.
- The inclusion of LLMs shifted human tasks from execution-based tasks to review-based activities. E.g., a human-only review would entail a significant proportion of time spent on extraction, while an LLM-enhanced review could entail more human time spent on review and feedback.
- This work demonstrated substantial time savings when using LLMs for SLRs, and the savings at each step combined will overall result in shorter project timelines. Using this approach a SLR was completed in 30 days.