MSR12

Evaluating Large Language Models' Performance in Content Generation for Literature Review Reports

E. Finn[1], K. Rtveladze[1], I. Guerra[1], J. Leite[2], S. Caverly[1], V. Shah[1], J. Gallinaro[1], K. Matev[3], A. Lambova[3]

[1]IQVIA, London, UK, [2]IQVIA, Oeiras, Portugal, [3]IQVIA, Sofia, Bulgaria

BACKGROUND AND OBJECTIVE

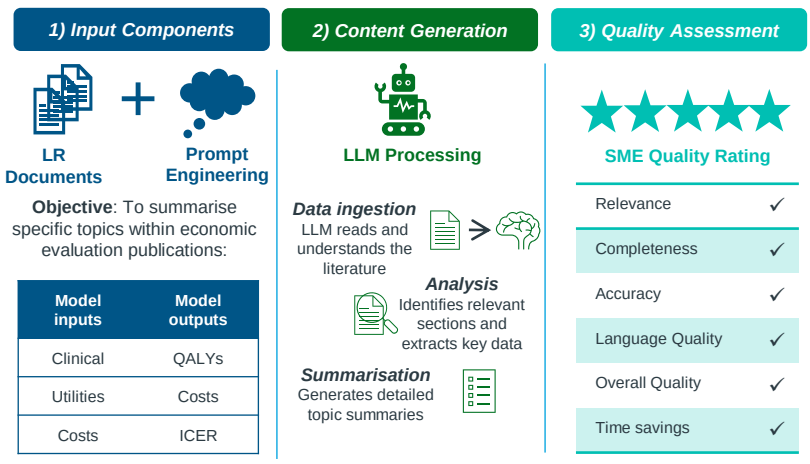
- Literature reviews (LRs) are the cornerstone of health economics workstreams, providing context and enabling the synthesis of information. Writing LR reports is labour-intensive and detail-orientated but is crucial for digesting the literature and achieving successful outcomes.
- Large Language Models (LLMs) are advanced machine learning tools designed to integrate information and utilise this information to generate text as requested by the user. These models have the capability to revolutionize content generation, potentially attaining a degree of expertise that could parallel human proficiency.
- Given this potential, there is significant untapped opportunity for applying LLMs in LR report writing. This study investigated the practicality, efficiency and effectiveness of LLMs in generating content for LR reports, aiming to determine their viability as a tool for enhancing the LR process.

Table 1. Evaluation criteria of LLM generated LR report content

Relevance	Is the response directly related to the question?
Completeness	Does the response provide a full answer to the question?
Accuracy	Is the response accurate?
Language	Is the response well-written?
Overall quality	Is the response of sufficient quality to use directly in a LR report?
Time savings	What percentage time do you estimate would be saved by using this approach compared to human resourcing? (%)

METHODS

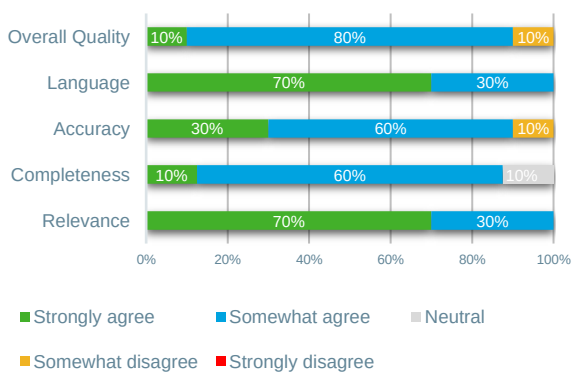
- Subject Matter Experts (SMEs) in LRs utilised a LLM pipeline and prompt engineering to generate content related to economic evaluation publications. These publications covered a range of indications with varying complexities of economic analysis.
- The prompts directed the LLM to summarise details of publications related to specific inputs and outputs of economic models. For inputs, the clinical data, utility values and cost inputs were examined. For model outputs, the focus was on quality-adjusted life years (QALYs), costs, and incremental cost-effectiveness ratios (ICERs).
- Results were analysed based on SME expectations of human-generated outputs when conducting LRs. Performance of the LLM output was evaluated across several dimensions, including relevance, completeness, accuracy, language quality, and overall quality (**Table 1**), using a five-point Likert Scale; 'strongly agree', 'agree', 'neutral', 'disagree' and 'strongly disagree'.
- Additionally, SMEs estimated the time saved when using the LLM, in comparison to manual content generation.



RESULTS

- SMEs strongly or somewhat agreed that all generated content was relevant to the topic under consideration, with 70% strongly agreeing and 30% somewhat agreeing (**Figure 1**). However, SMEs unanimously noted that responses lacked some details, indicating the need for more detailed LLM prompts to elicit more comprehensive responses. It was observed that more detailed prompts not only resulted in more accurate and relevant content, but also reduced instances of hallucinations.
- SMEs strongly or somewhat agreed that 90% of the responses generated were accurate (**Figure 1**), although there were isolated cases where incorrect numbers were retrieved. Despite the inclusion of context in the prompts, hallucinations were still present, highlighting the necessity for human input to carefully review and validate LLM-generated content
- In terms of time savings, the percentage of time saved varied between 20% and 60%, which appeared to correlate with the level of complexity of the economic analysis. More human input was required for studies of high complexity.
- Overall, SMEs strongly or somewhat agreed that they would incorporate the generated responses into their reports, with 90% of responses indicating agreement, and they anticipated significant effort savings compared to manual content generation.
- Additionally, SMEs suggested that future improvements could include refining the LLM's ability to handle complex data and integrating automated tools for identifying and incorporating relevant tables and figures from source documents. This would further enhance the efficiency and accuracy of the LLM in generating comprehensive literature review reports.

Figure 1. SME evaluation of LLM generated LR report content



CONCLUSIONS

- This study demonstrates that LLMs can significantly aid in content generation for LR reports and offer valuable time savings.
- However, their performance is contingent on the complexity of the information and the level of detail provided in the LLM prompts.
- SME review is essential to ensure accuracy and completeness of the outputted information.
- Further testing and refining of prompts, as well as evaluating LLM performance at generating reports for other types of LRs, such as clinical and safety, should be conducted.

> "In general, the LLM output using full questions as prompts is more accurate and useful, with fewer hallucinations. The issue of copying and pasting also seems to be significantly reduced. The LLM was generally good at capturing the relevant details. Improving some prompts to include more necessary details could help to enhance the overall quality."

> "The LLM output is close to what is needed for a first draft report; there remains a need for human review to ensure no important detail is missing. The responses produced by the LLM were largely well-written and of the quality we would expect."

> "While the LLM had some trouble with correctly providing all of the relevant detail on the clinical inputs in the study, everything else was well reported and accurate. The results, in particular, were very clear and easy to transpose to a report."

Abbreviations: ICER: Incremental cost-effectiveness ratio; LLM: Large Language Model; LR: Literature review; QALY: Quality-adjusted life year; SME: Subject Matter Expert.

Correspondence: Inês Guerra (Ines.Guerra@iqvia.com)

International Society for Pharmacoeconomics and Outcomes Research Europe 2024, November 17-20, 2024, Barcelona, Spain