Advancing Systematic Literature Reviews: A Comparative Analysis of Large Language Models (Claude Sonnet 3.5, Gemini Flash 1.5, and GPT-4) in the Automation Era of Generative AI

MSR78

Rai P[1], Pandey S[1], Attri S[1], Singh B[2], **Kaur R**[1]
[1]Pharmacoevidence, Mohali, India, [2]Pharmacoevidence,London,UK

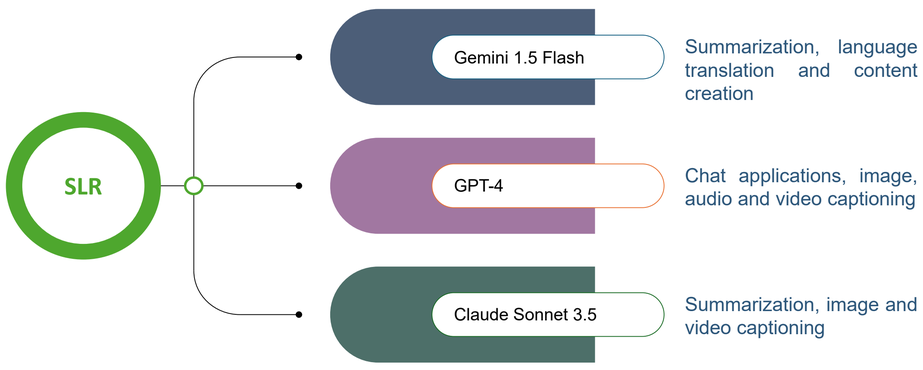
CONCLUSION

The study highlights the comparative effectiveness of the three AI models for automating literature reviews. The automated SLR tool achieved 96.02% accuracy with a two-review human process. Future investigations should further explore these capabilities and their application across diverse research domains.

INTRODUCTION

- A Systematic Literature Review (SLR) is a research methodology that employs a systematic approach to gather, identify, and critically evaluate the available research studies
- This process is inherently time-intensive and complex, as it requires precise search strategies, the searching of large volumes of literature
- With the development of generative AI, these issues are addressed by automated, intelligent systems capable of doing many of the challenging tasks required in SLRs
- SLRs can be conducted quickly and efficiently using the capabilities of these generative AI models, while simultaneously adhering to the rigorous standards
- The LLM and generative AI approaches allow the automation of screening tasks
- Advanced LLMs like Gemini 1.5 Flash, GPT-4, and Claude Sonnet 3.5 are powerful because of their outstanding capacity to perform tasks associated with SLRs

Figure 1: SLR using various LLM model



OBJECTIVE

- In recent years, the advent of LLMs has revolutionized the traditional approach of conducting SLRs. These models exhibit diverse capabilities in comprehending and synthesizing the vast volumes of literature, offering potential efficiency gains and novel insights. Understanding their comparative efficiency is essential for discovering the optimal tool in the evolving landscape of AI-driven literature analysis.
- This research investigates the relative efficiency of the generative AI models (Claude Sonnet 3.5, Gemini Flash 1.5, and GPT-4) in data collection phase of SLRs

METHODS

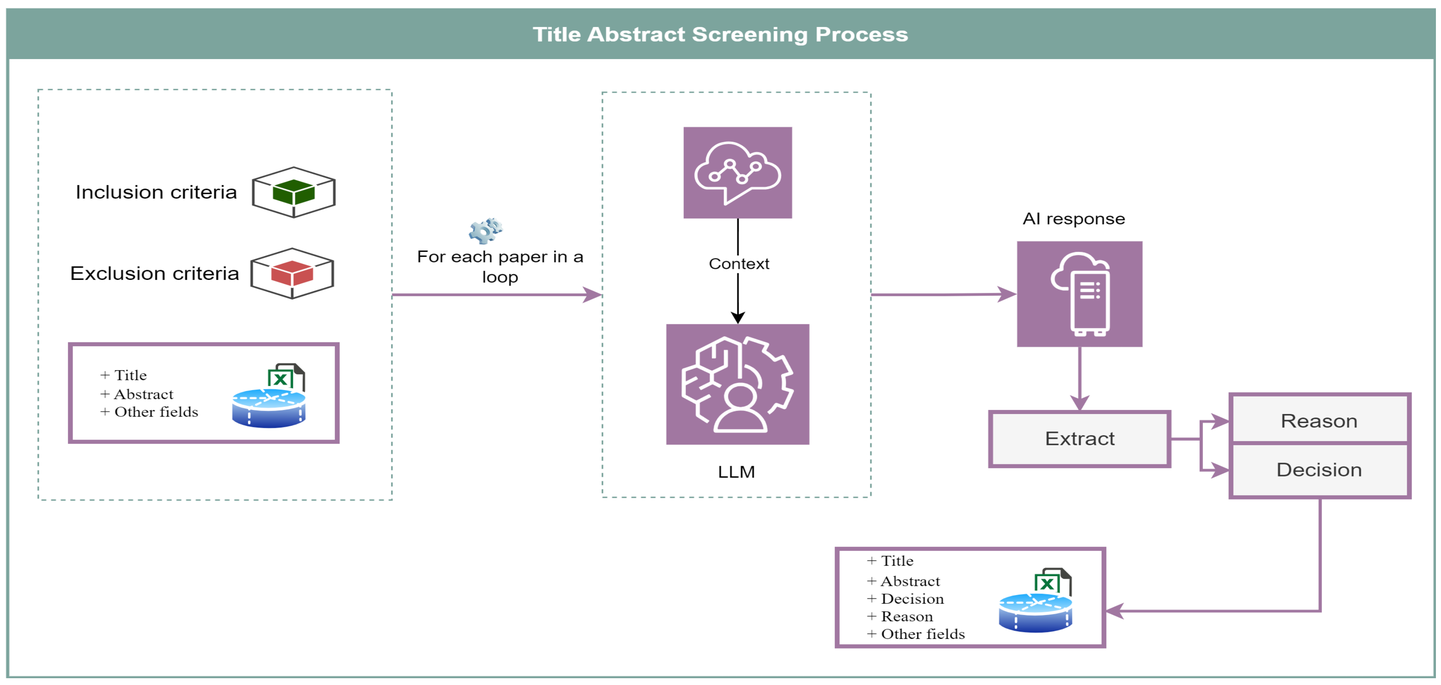
- Embase®, Medline®, and Cochrane databases were searched to identify relevant randomized controlled trials (RCTs) in the disease area of interest
- A Python script was developed to simplify the interaction between the input data sheets and LLMs
- A subject matter expert with over a decade of domain knowledge optimized and fine-tuned the final prompt to identify evidence meeting the eligibility criteria
- Title and abstract-based screening was conducted to identify eligible publications
- Three LLMS were used to screen the data, which included titles, abstracts, and other relevant fields
- The LLM models are pre-trained models that understand complex language structures and provide decisions as per the eligibility criteria
- The Python script iterated through each entry in the spreadsheet, passing the title, abstract, inclusion and exclusion criteria, and context to the LLMs for evaluation, as shown in Figure 2
- This iterative procedure ensured that every article's information was thoroughly reviewed using the stated criteria and contextual information
- A parser function was built to extract the final decision of the LLM models as "Included" or "Excluded"
- The final decision was analysed by the SME to check the performance of the different LLM models
- This technique combined AI-driven automated screening with human validation to make systematic literature reviews far more effective, and efficient

METHODS

Prompt used:

```
Mark the Citation as "Included" only if it strictly and exactly passes all Inclusion Criteria statements; otherwise, mark the citation as "Excluded".
## Criteria:
Inclusion Criteria (All should be met for Include):
{inclusion_criteria}
Exclusion Criteria (Anyone met will result in Exclude):
{exclusion_criteria}
## Paper Information:
Title: {title}
Abstract: {abstract}
## Your Response:
Provide your response in the following format:
Classification: [Include/Exclude]
Reason: [Provide a brief, clear explanation for your decision, referencing specific criteria]
Remember: Be thorough and objective and base your decision strictly on the provided criteria and information.
```

Figure 2: Systematic workflow diagram for literature classification



RESULTS

- The models were evaluated using the performance measures such as accuracy, precision, recall, and specificity, F1 score
- Overall, all three AI models performed exceptionally well in screening based on titles and abstracts. While there were no significant differences in accuracy rates, Gemini Flash 1.5 exhibited the highest accuracy rate at 96.02%, followed by GPT-4 at 95.00%, and Claude Sonnet 3.5 at 94.69%. In terms of sensitivity, GPT-4 suggested better results attaining 95.97% of sensitivity followed by 94.63% with Gemini Flash 1.5 and 88.59% with Claude Sonnet 3.5.

Figure 3: Comparison of all LLM model for SLR screening phase

References

1. Mahuli et al. *Br Dent J*. 2023;235(2):90-92
2. Issaiy et al. *BMC Med Res Methodol*. 2024;24(1):78
3. Winberg et al. *Value Health*. 2023;26(12):S410

Disclosure

PR, SP, SA, BS and RK authors declare that they have no conflict of interest

Need a poster copy?

Scan here

Hands-on experience with AI tools?

Contact: rshiny@pharmacoevidence.com