A Comparative Analysis of Large Language Models Utilised in Systematic Literature Review

Hemant R,[1,2] Malik A,[2] Behera D C,[2] Kamboj G[1]

[1]Skyward Analytics, Gurugram, Haryana, India; [2]EasySLR, Gurugram, Haryana, India

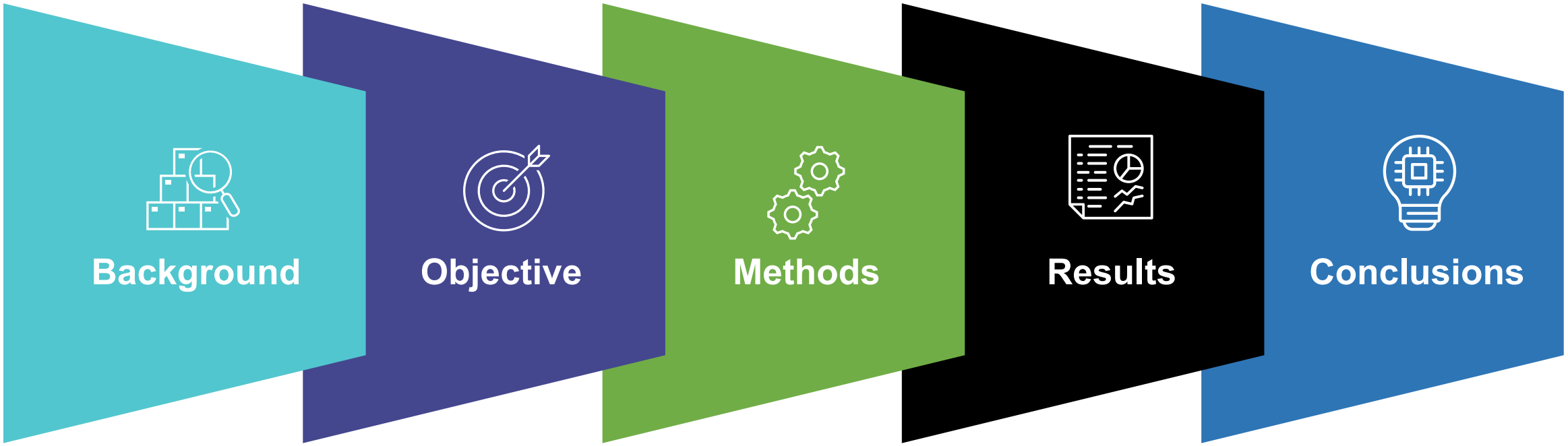
Podium Presentation at the ISPOR Annual European Meeting, 13 Nov 2023, Copenhagen, Denmark
Presentation code: **P21**

Declaration of Interests

All authors are employees of Skyward Analytics (Rathi H, Kamboj G) and EasySLR (Rathi H, Malik A, Behera D C).

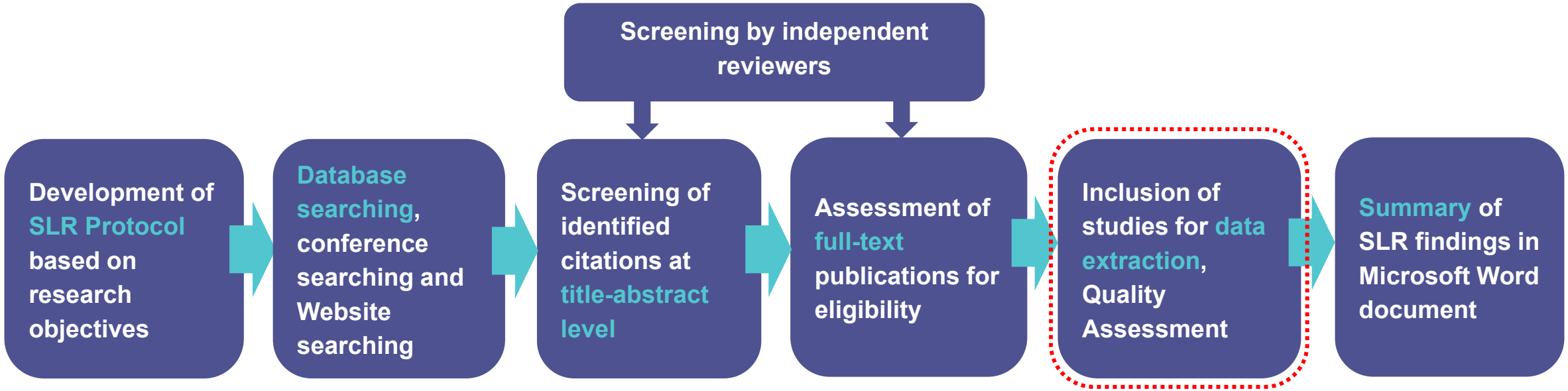
No other funding was received for this study.

Outline



What is a Systematic literature review?

- A SLR is a rigorous research method used to comprehensively gather, assess, and synthesize existing literature on a specific topic or research question, following a predefined and structured protocol[1]
- SLRs aim to provide an evidence-based overview of the existing knowledge in a particular field, serving as a foundation for informed decision-making and further research[1]



1. Paul J, Barari M. Meta-analysis and Traditional Systematic Literature reviews—What, why, when, where, and how? Psychology & Marketing. 2022 Mar 11;39(6).

What are LLMs?

- Large Language Models (LLMs) are advanced artificial intelligence (AI) systems designed to understand and generate human-like text[1]
- They have gained prominence in the field of natural language processing due to their remarkable ability to comprehend and produce human language with an unprecedented level of fluency and context awareness

Why are we talking about this?

- The increasing volume of publications in scientific databases has made it increasingly difficult to conduct a timely literature review[2]
- Within HEOR, these models are increasingly employed for one of the use cases, which is the execution of systematic literature reviews (SLRs)[3]
- LLMs can empower researchers to make more informed decisions and expedite the systematic review process

1. Tustumi F, Andreollo NA, Aguilar-Nascimento JE. Future of the language models in healthcare: The role of ChatGPT. Arq Bras Cir Dig. 2023 May 8;36:e1727.
2. Van Dinter, R., Tekinerdogan, B., Catal, C., 2021. Automation of systematic literature reviews: a systematic literature review. Inf. Software Technol. 136, 106589.
3. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. Systems [Internet]. 2023 Jul 1;11(7):351.

Objective

To conduct a comparison of three LLMs in their application to primary screening during a SLR —

1. AI21 Ultra
2. OpenAI GPT-4
3. Google Vertex AI Model Bison

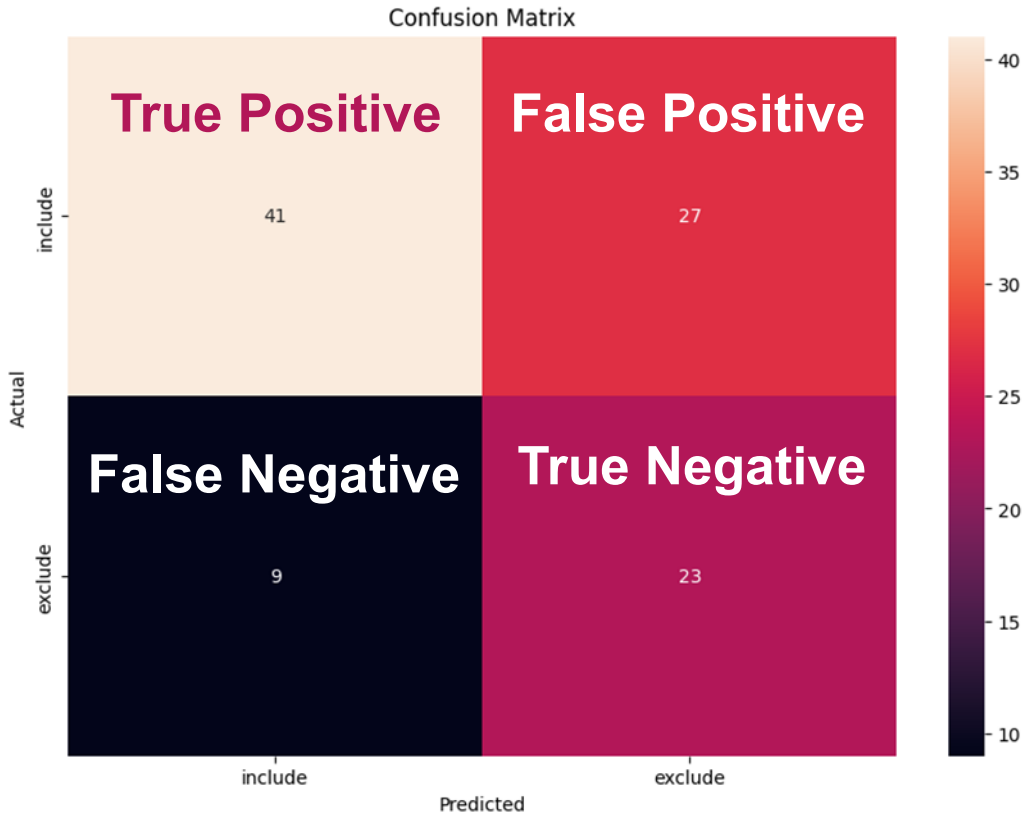
Methods

	Following were fed to all LLMs for primary screening (**title/abstract screening**): ✓ **Comprehensive screening rules** ✓ **Inclusion/Exclusion criteria** ✓ **Three, five, or no sample responses**
	All LLMs screened **100 studies** utilising these screening rules and sample responses
	We compared the decision made by these LLMs to the human reviewer decisions
	Decision made by the **human reviewer** was assumed as reference response to gauge the performance of the LLMs
	LLMs were assessed on **decision match rate**, **precision**, **sensitivity**, **specificity**, **ROC AUC score**, and **F1 score** *(defined on next slide)*

Methods- Evaluation metrics

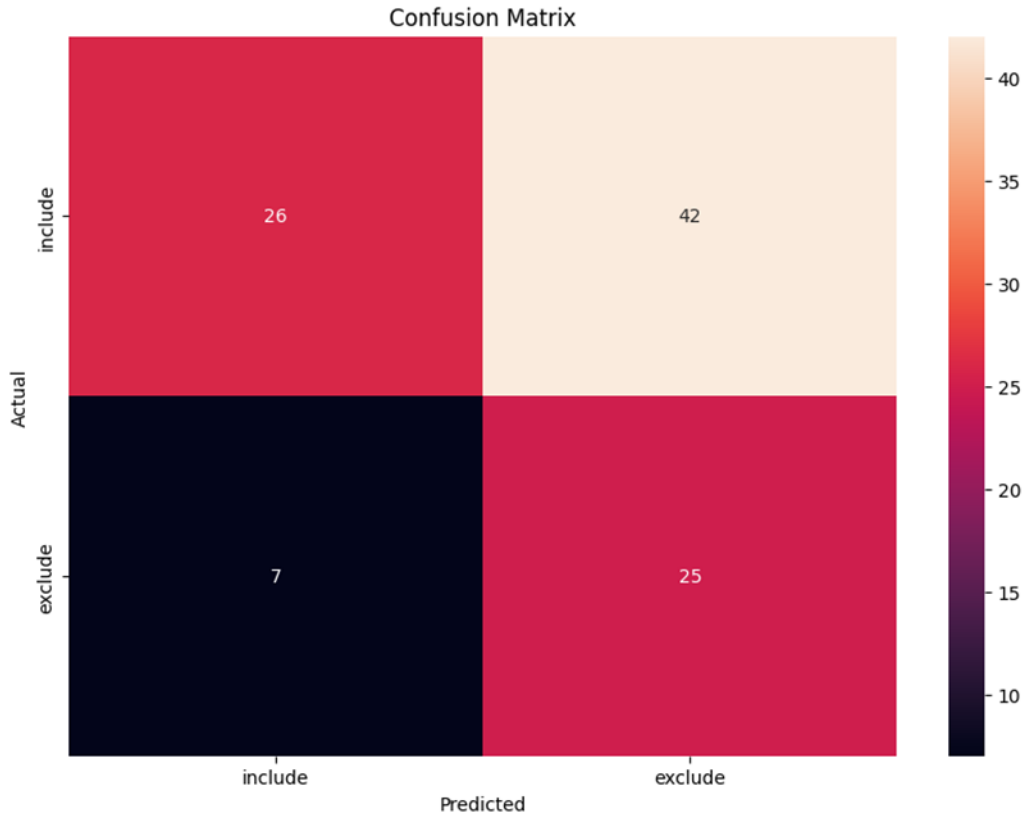
1	Decision match rate	• Cases where inclusion and exclusion decisions were identical between the human reviewer and LLM
2	Precision	• Proportion of predicted 'include' that are actually 'include'
3	Sensitivity	• Proportion of actual 'include' that are predicted 'include'
4	Specificity	• Proportion of actual 'exclude' that are predicted 'exclude'
5	ROC AUC score	• Model's capability of distinguishing between classes
6	F1 score	• Harmonic mean of Precision and Sensitivity

Results: AI21 Ultra – Confusion Matrix



Prompt with 3 sample responses

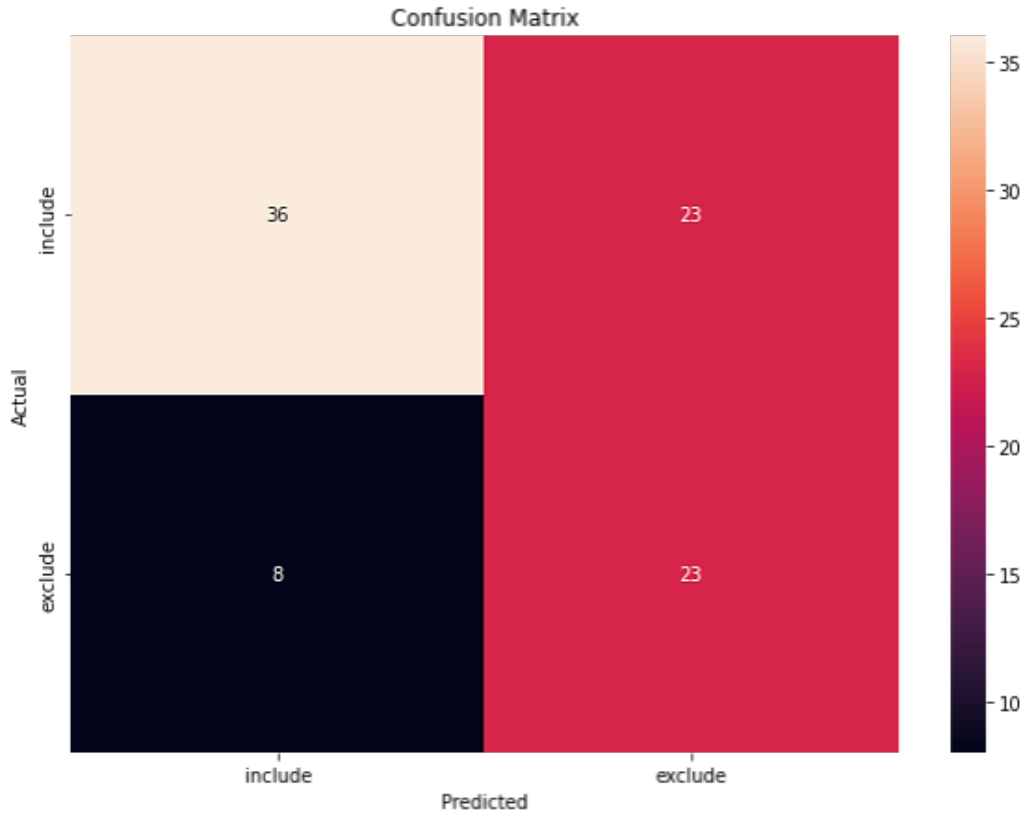
F1 Score: **0.50**
Decision match rate: **64.0%**



Prompt with 5 sample responses

F1 Score: **0.56**
Decision match rate: **51.0%**

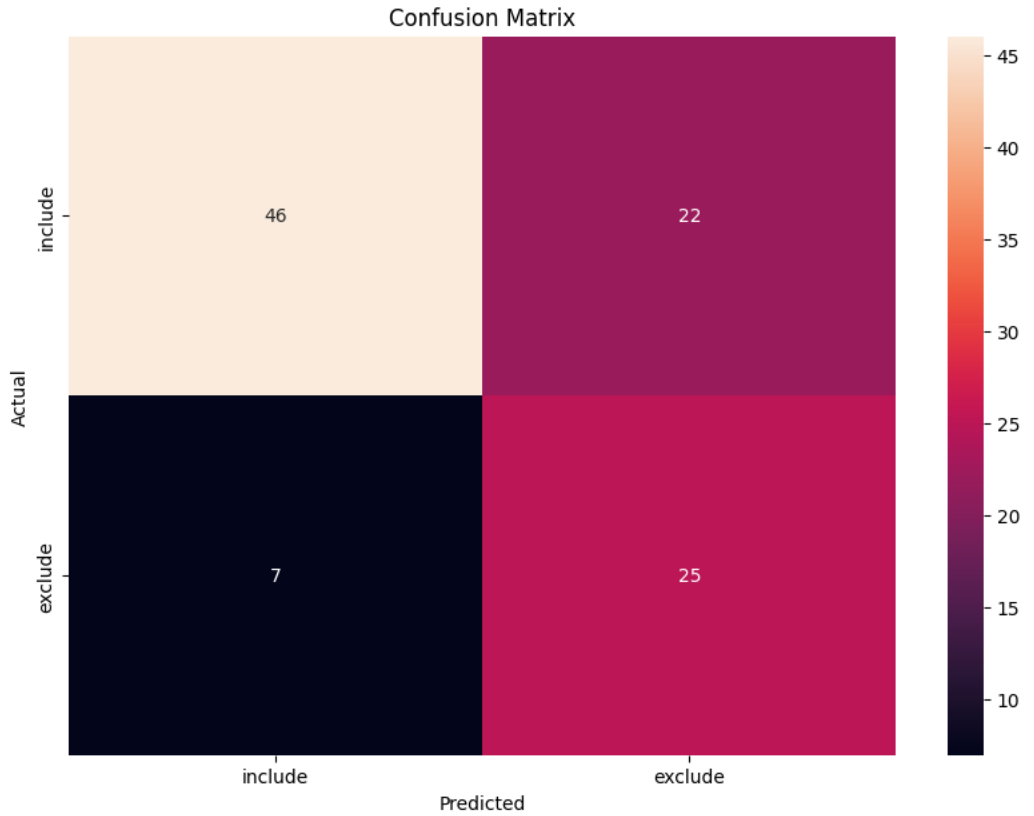
Results: OpenAI GPT-4 – Confusion Matrix



Prompt with 0 sample responses*

F1 Score: **0.60**

Decision match rate: **65.6%**



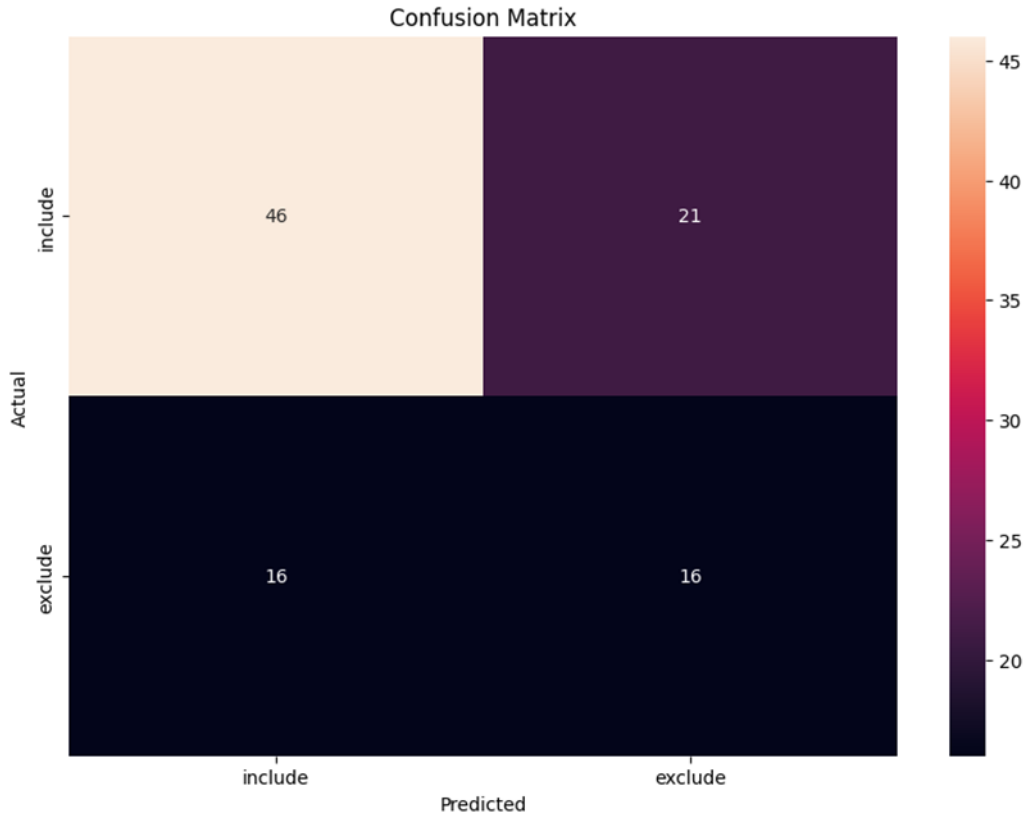
Prompt with 5 sample responses

F1 Score: **0.63**

Decision match rate: **71%**

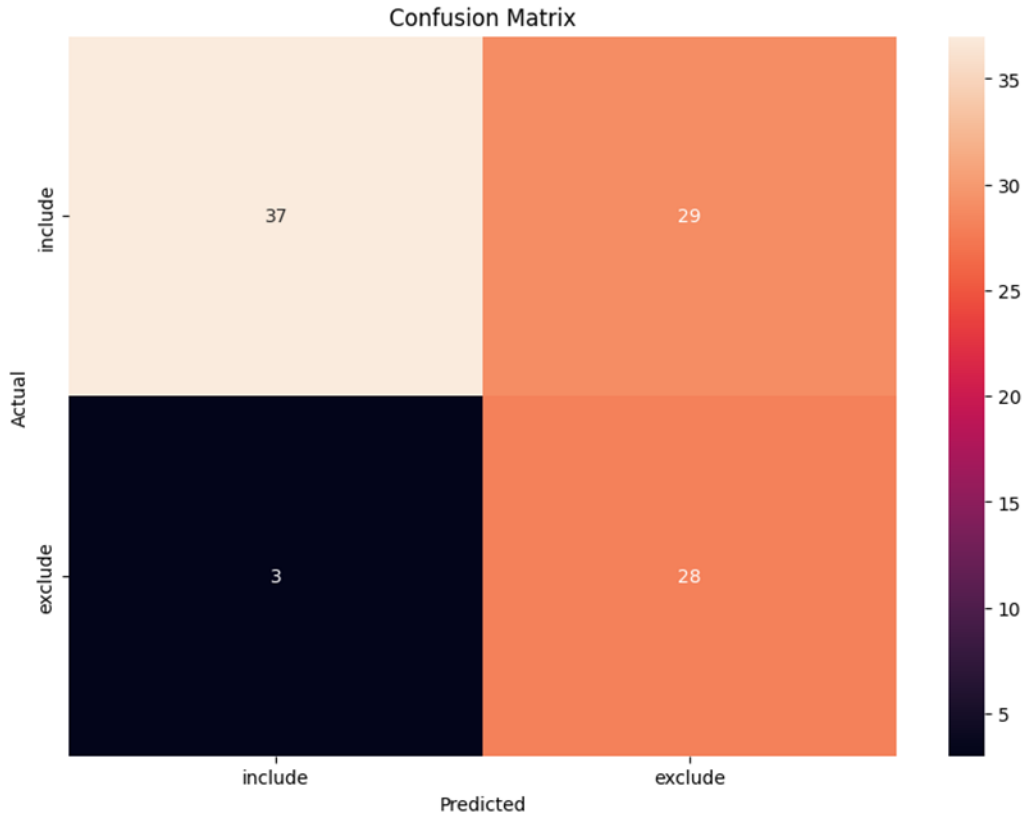
*LLM was unable to take decisions for 10 studies

Results: Model Bison – Confusion Matrix



Prompt with 3 sample responses*

F1 Score: 0.46
Decision match rate: 62.6%



Prompt with 5 sample responses**

F1 Score: 0.63
Decision match rate: 67.7%

*LLM was unable to take decisions for 1 study
**LLM was unable to take decisions for 3 studies

Results: Assessment (0 or 3 sample responses)

Workstream	AI21 Ultra (3)	OpenAI GPT-4 (0)*	Model Bison (3)**
Decision match rate	64.0%	**65.6%**	62.6%
Precision	0.46	**0.50**	0.43
Sensitivity	0.72	**0.74**	0.50
Specificity	0.60	0.61	**0.69**
ROC AUC score	0.66	**0.68**	0.59
F1 score	0.56	**0.60**	0.46

*OpenAI GPT-4 unable to take decisions for 10 studies
**Model Bison unable to take decisions for 1 study

Results: Assessment (5 sample responses)

Workstream	AI21 Ultra	OpenAI GPT-4	Model Bison*
Decision match rate	51.0	**71.0**	67.0
Precision	0.37	**0.53**	0.49
Sensitivity	0.78	0.78	**0.90**
Specificity	0.38	**0.68**	0.56
ROC AUC score	0.58	0.72	**0.73**
F1 score	0.50	0.63	**0.64**

*Model Bison unable to take decisions for 3 studies

Conclusions

Takeaway points:

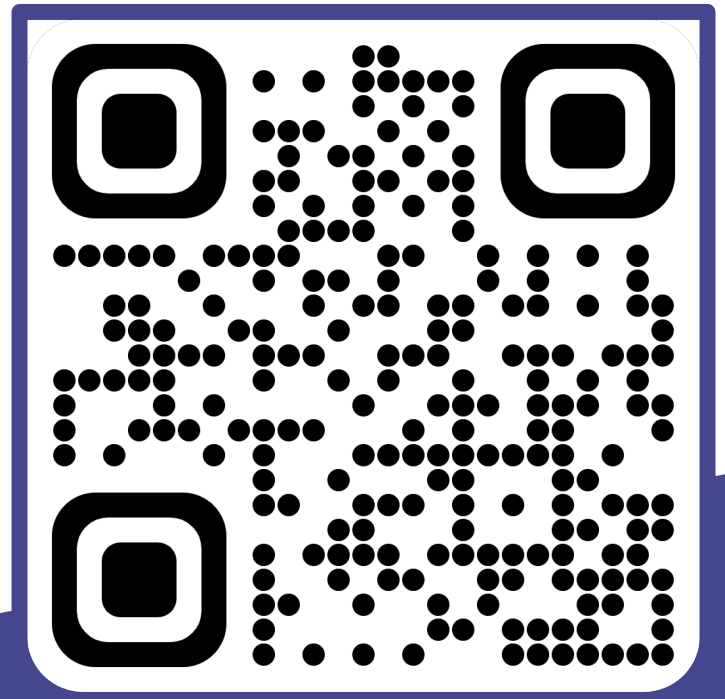
- The results highlight LLMs' potential to assist with the SLR process
- All LLMs were comparable in decision match rate metric
- GPT-4 showed better precision, ROC AUC score, and F1 score than Model Bison and AI21 ultra
- Model Bison showed better specificity than GPT-4 and AI21 Ultra
- AI21 ultra performed better in terms of sensitivity compared to GPT-4 and Model Bison

Limitations and further research:

- Results should be interpreted cautiously as the results may vary with different research questions
- Future research should consider analysing the performance of LLMs on larger datasets, variation in number of sample responses fed, and calibration around framing of screening rules for better understanding by AI
- The upcoming analyses will delve into the utilization of LLMs in the processes of full-text screening and data extraction.

Thank you!

Questions?



hemant.rathi@skywardanalytics.com