PT23

Is a Large Language Model (LLM) More Accurate Than Human Researchers in Correctly Identifying Diseases in Biomedical Abstracts? A Pilot Study

Edema C [1], Rutherford L [1], Martin A [1], Martin C [1], Bertuzzi A [1], King E [1], Letton W [1]
[1] Crystallise Ltd, Stanford le Hope, Essex, UK

Objectives

- Artificial intelligence (AI) such as LLM is being applied to biomedical research to automate processes and improve efficiency. However, ascertaining the precision and reliability of AI in research tasks requires further investigation. We compared the accuracy and speed of an LLM versus human researchers in correctly identifying diseases in biomedical abstracts.

Methodology

- A targeted literature search was conducted to generate a list of 500 biomedical abstracts. Using an Evidence Mapper tool (www.evidencemapper.co.uk), each abstract was indexed separately by researchers and the LLM to nine predefined disease categories. The time taken for each method was recorded
- The OpenAI Python library was used to create a suitable prompt for the LLM's output. A gold-standard disease category was created independently against which the researcher and LLM disease indexing were compared.

Results

- The mean sensitivity and specificity across the disease categories for the LLM was 70.7% and 97.2% versus 66.6% and 98.8% for the researchers. The range of sensitivity and specificity for the LLM was 16.67% to 100% and 87.6% to 100% respectively. For researchers, sensitivity ranged between 0% to 91.4% and specificity from 92.7% to 100%. Overall, indexing by researchers was more accurate with a mean accuracy of 98% compared to 96% using the LLM.
- The initial indexing was more than three times faster using the LLM (3 hours) compared to researchers (10.4 hours). However, it took a longer time to check the LLM indexing (9.5 hours vs 6.25 hours), resulting in a net time saved of 4.15 hours.

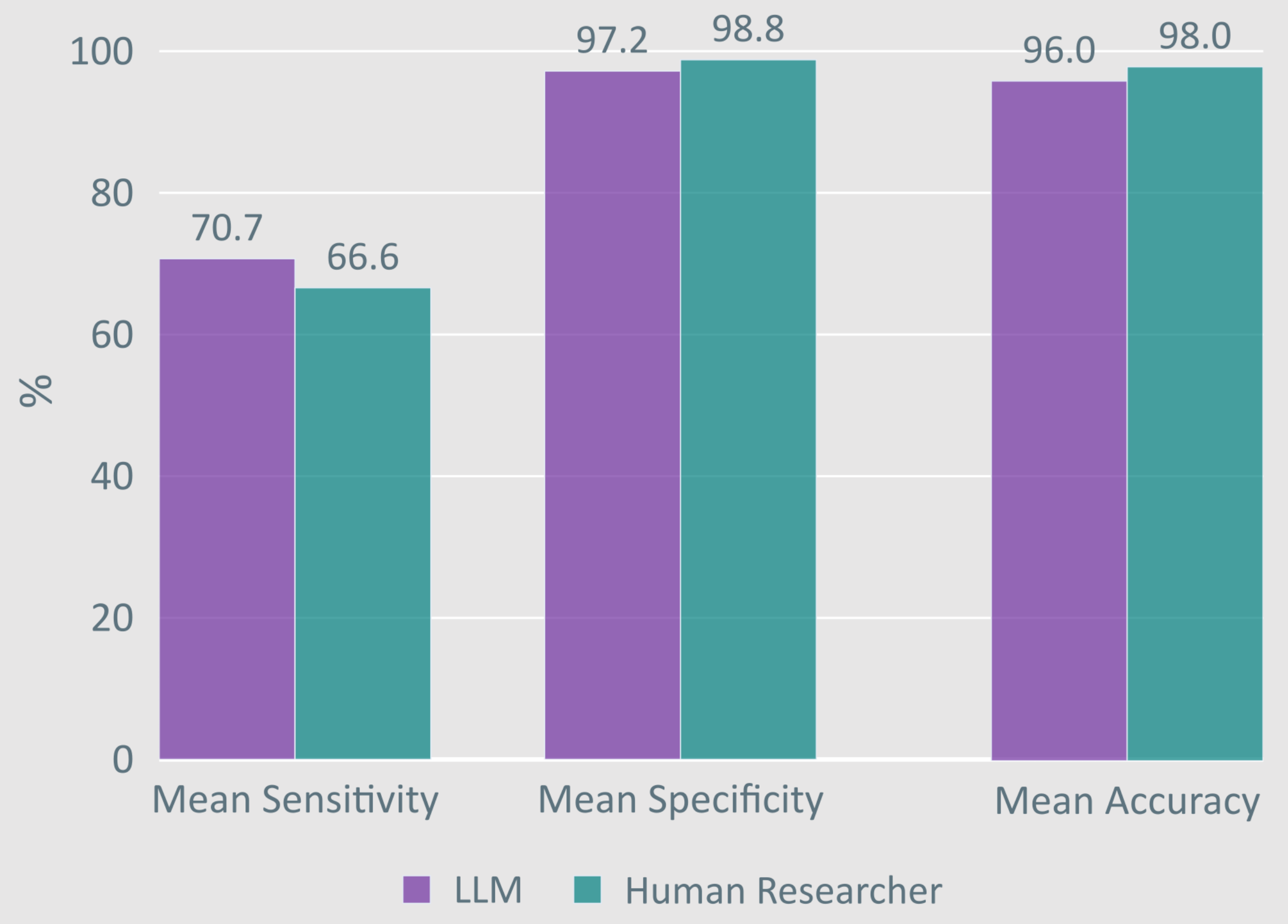

Figure 1: Comparison of the Mean Sensitivity, Mean Specificity, and Mean Accuracy of Researcher vs LLM for Disease Indexing

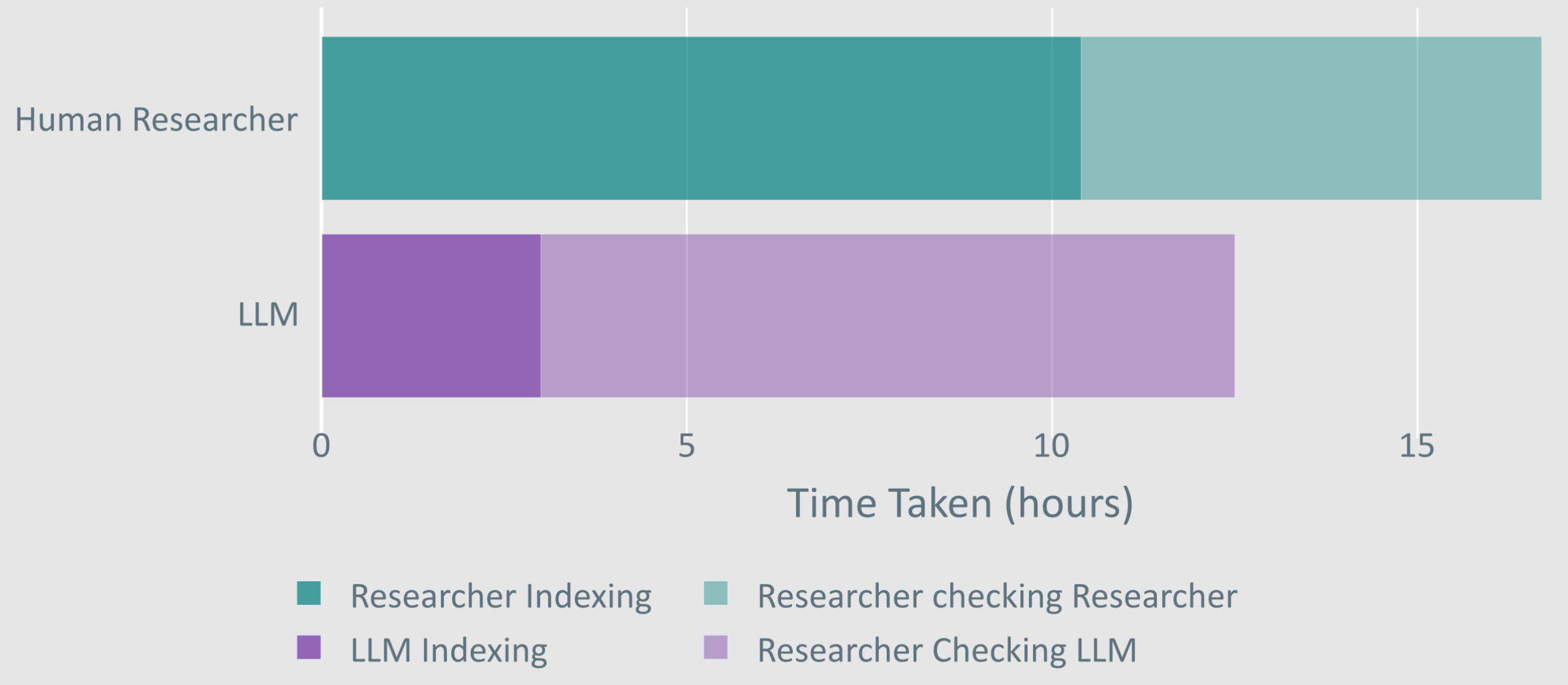

Figure 2: Comparison of the Time Taken for Disease Indexing

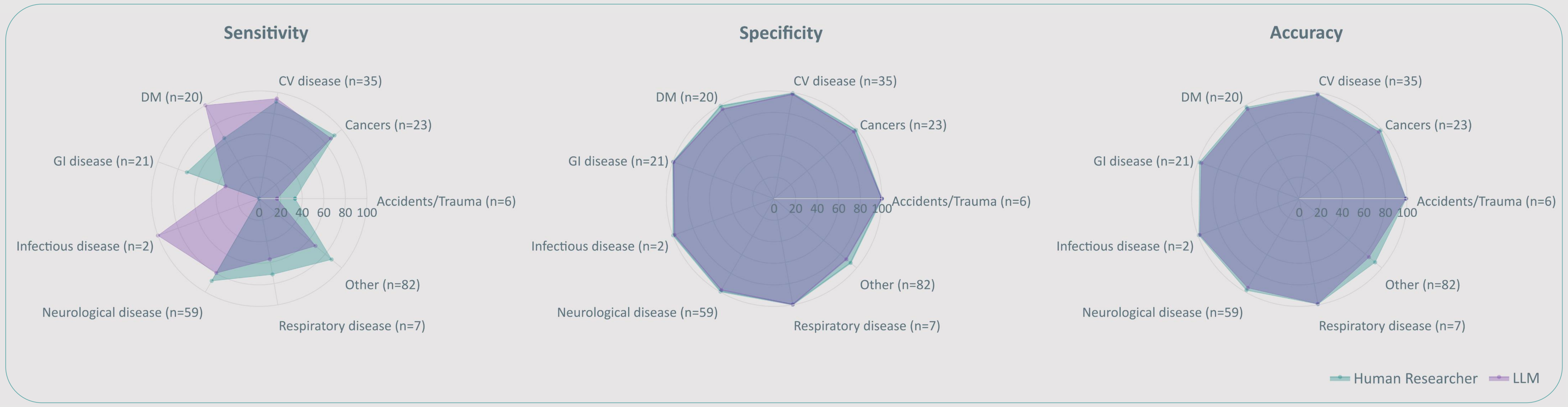

Figure 3: Comparison of the Sensitivity, Specificity, and Accuracy of Researcher vs LLM for Disease Indexing

Conclusion

- Given that verifying the LLM's indexing takes at least as much time as verifying the researcher's indexing, the overall time saved is less than what may be inferred solely from the initial indexing speed. Nonetheless, utilizing LLMs in the evidence synthesis process can be time-saving with an acceptable degree of accuracy compared to humans.
- Human checking of AI-suggested indexing might be the most cost-effective approach for database indexing and abstract screening for literature reviews.

Email: contact@crystallise.com

Website: www.crystallise.com

Crystallise

LinkedIn (Crystallise Ltd)
X (@crystalliseL)
YouTube (@crystallise3499)