5665

Evaluation Of The Use Of ChatGPT In Data Extraction For Systematic Literature Reviews: Friend, Foe Or Fiction?

Roussi, Kalliopi[1]; Rice, Hannah[1]; Bertuzzi, Andrea[1]; Martin, Alison[1]
[1]Crystallise Ltd, Colchester, CO4 9PE

Crystallise

Introduction

Previous research has found inconsistent benefits from large language models (LLMs) for data extraction within systematic literature reviews.

Objective

We aimed to evaluate the accuracy and reproducibility of the latest ChatGPT for data extraction.

Method

We developed a prompt for ChatGPT 4.0 in December 2024 to January 2025 to extract the same specific data items on study methodology and baseline characteristics (BC) from a conference abstract, a clinical trial, a case report, and a retrospective observational study. We evaluated the LLM's accuracy compared to a human researcher-derived gold standard and its consistency of output for each data item when run on the same papers three times a day for one week, then once a day for four weeks.

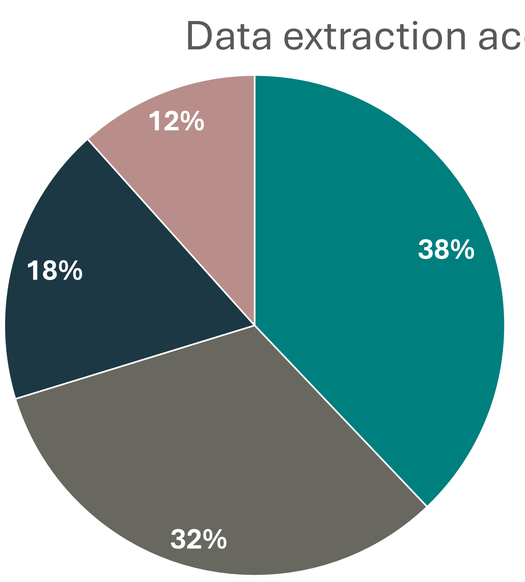
Results

Overall accuracy varied substantially with no clear trend over time. ChatGPT 4.0 inconsistently and unpredictably extracted data: incorrect data in 17.5% of instances, data missing in 9.7%, fabricated data in 1.2% and unclear data in 0.9%.

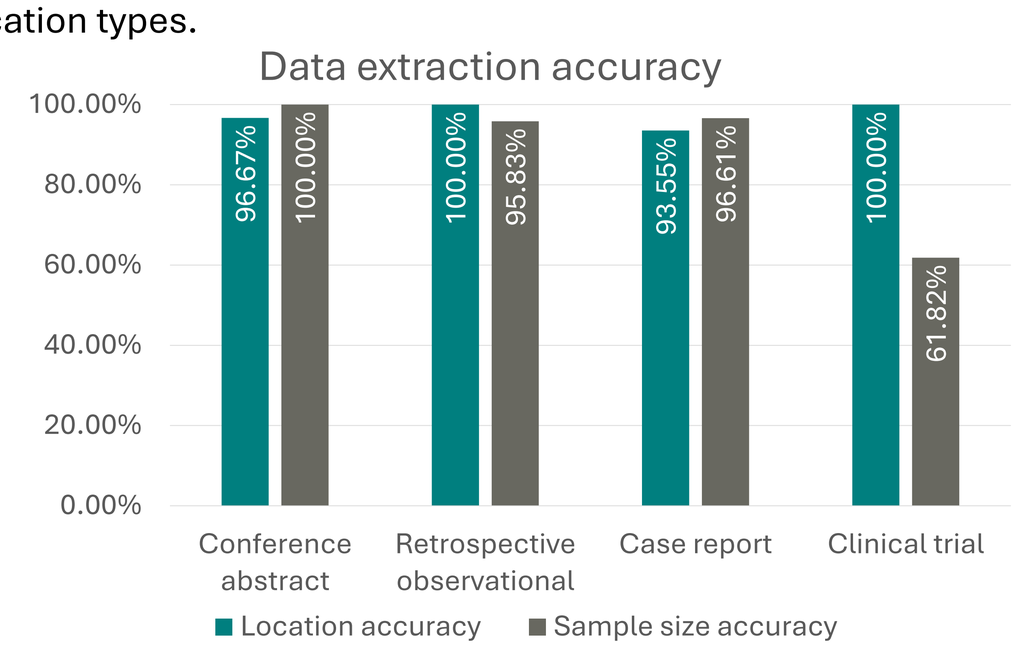
Errors were significantly higher for full-text clinical trial reports (32.4% of instances) than for retrospective observational studies(22.9%, $p < 0.001$).

Simple data types such as number of participants and location were correct in 61.8% to 100% and 93.5% to 100% of runs, respectively, but more complex outcomes such as extraction of the full citation were only correct in 0 to 33.3% of runs.

No data item could be relied on to always be extracted correctly across all publication types.



*Other: Data not found, new data extracted, comment made



GOLD STANDARD

Al-Sweedan. S, Rafat. J, Ali. A, Amal. A, Ibrahim. G, Hawazen. A, Siddiqui. K. Allergic Reaction to Asparaginase in Pediatric Acute Lymphoblastic Leukemia (A.L.L.) – Does It Have Any Prognostic Value?. SIOP 2017

Typographical errors/ symbols

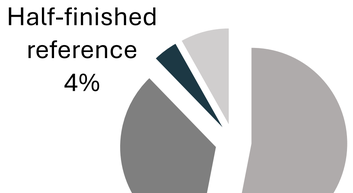
" AlSweedan S, et al. Allergic Reaction to Asparaginase in Pediatric Acute Lymphoblastic Leukemia â€“ Does It Have Any Prognostic Value? SIOP 2017. "

Missing part of the reference

" Al-Sweedan S, et al. SIOP 2017„

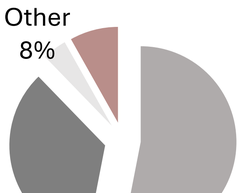
Half-finished reference

" AlSweedan S et al., Allergic reaction to asparaginase in pediatric ALL ... „



Other

" Reference for AlSweedan 2017 „



Conclusion

ChatGPT 4.0 could not be relied on for accurate data extraction but might save time by providing an initial output for human researchers to validate. Accuracy of outputs will depend on the skill of the prompt generator, but varies considerably over time despite using the same prompt. Studies assessing the accuracy of AI for data extraction should take this temporal variation into account. Since this study was concluded ChatGPT 5 has become the most current version, which may show greater improvement to the 4.0 version. Further investigation is required.

Contact Information

Email: contact@crystallise.com
Website: www.crystallise.com

LinkedIn (Crystallise Ltd)

YouTube (@crystallise3499)