ISPOR 2025
May 13-16 Montreal, QC, Canada

EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

Evaluating the performance of Claude 3.7 Sonnet in data extraction automation for systematic literature reviews (SLRs)

Chow C, Kasireddy E, Pourrahmat MM, Collet JP, **Fazeli MS**

Evidinno Outcomes Research Inc., Vancouver, BC, Canada

Contact: MFazeli@Evidinno.com

Disclosures

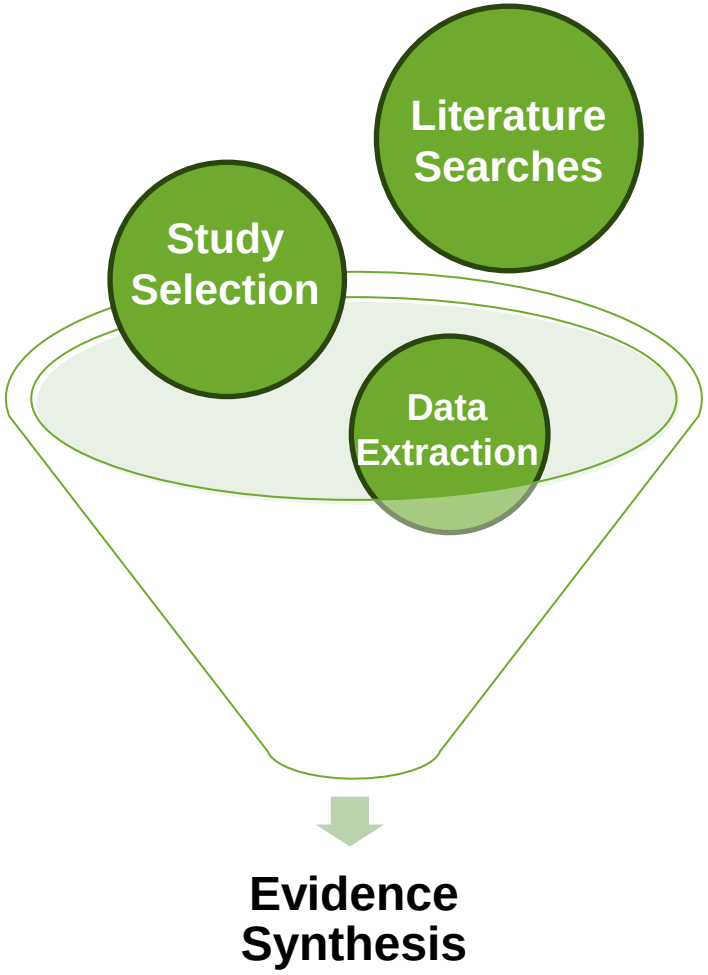
- All authors report employment with Evidinno Outcomes Research Inc. (Vancouver, BC, Canada).

Background

Why automate data extraction in systematic reviews?

- Systematic literature reviews (SLRs) are crucial for evidence synthesis
- Manual data extraction is resource-intensive, requiring high accuracy and consistency[1,2]
- For text-based tasks, large language models (LLMs) such as the Claude 3.7 Sonnet LLM, offer potential to[3-12]:
  - Increase efficiency
  - Reduce human error
  - Enhance scalability

Rigorous performance evaluation is essential to ensure reliability and uphold methodological standards in SLRs



Objective

To evaluate the performance of Claude 3.7 Sonnet for automating data extraction in systematic literature reviews

Methods – Model Overview and Setup

Model Overview

- Custom artificial intelligence (AI) extraction system built on **Claude 3.7 Sonnet**, which offers:
  - **A large reading capacity (200,000 tokens)**, allowing it to process full articles at once
  - **The ability to read PDFs directly**, including visual content such as figures and charts, without needing conversion
  - **Strong performance on language tasks** (top-ranked in LiveBench 2024)

System Setup

- The system includes:
  - **LangChain**, a framework to help manage and organize AI prompts dynamically
  - **Chunkr.ai**, a platform to convert PDFs into clean, structured text
  - **Zod schemas**, validation and typing tools to keep extracted data organized and consistent
- All extracted information is mapped to a **predefined database-ready format** to support further analysis
- The **prompting strategy** used structured, step-by-step instructions to guide consistent extraction of specific data types from publications.

Methods – Master Variable List and Datasets

Previously completed SLRs (treatment efficacy and safety in oncology)
(9 SLRs)

Master Variable List
(all 9 SLRs – 767 publications)

- Created by two senior researchers
- Includes variables and definitions to provide contextual understanding
  - Study characteristics
  - Participant characteristics
  - Intervention characteristics
  - Outcomes

Training Dataset
(4 SLRs – 7 publications)

- Prompt refinements were made based on discrepancies between model outputs and human extractions
- Over 10 iterative rounds of refinements

Validation Dataset
(4 SLRs – 20 publications)

- Model extraction was compared with human data extraction by a senior researcher
- Performance metrics were calculated

Methods – Data Extraction Model Workflow Summary

Input:
Full-text PDF articles included in SLR

Preprocessing by Claude 3.7 Sonnet LLM

- Processing PDFs
- Organizing PDF content into meaningful sections

Guided Data Extraction

- Structured instructions (prompts) tell the model **what** information to find and **how** to format it

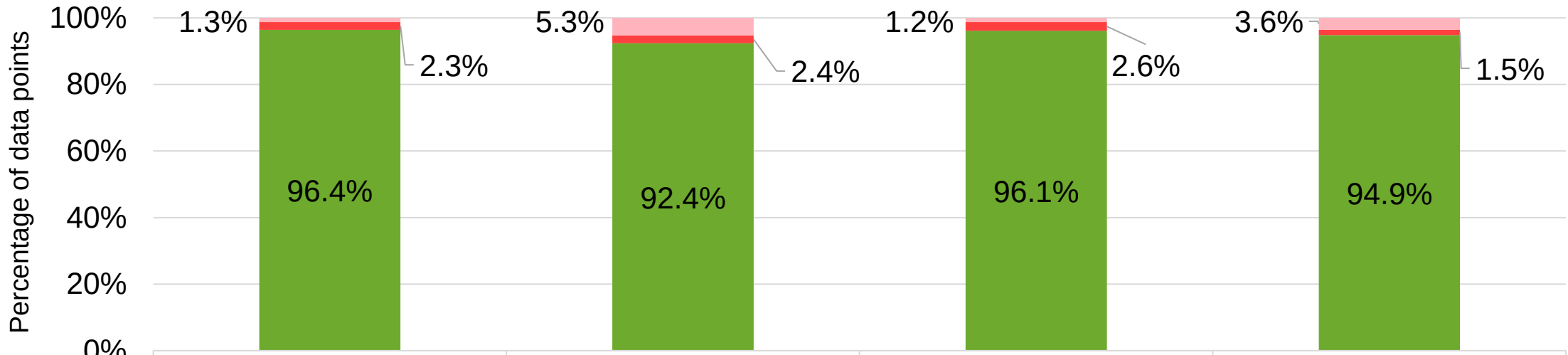
Two-Stage Extraction Process

1. Identify relevant sections in PDF to extract data from
2. Extract specific data

Output:
Extracted data organized in table format

Results – Extraction Performance by Data Domain

- A total of **117,889 data points** were extracted across four data domains



	Study Characteristics	Intervention Characteristics	Participant Characteristics	Outcomes
False negative (n)	9	131	192	3,558
False positive (n)	16	59	406	1,522
True positive (n)	675	2,296	14,874	94,151

- **False negatives** = Relevant data missed by the model
- **False positives** = Incorrect or misclassified data extracted by the model
- **True positives** = Correct data extracted by the model, matching human extraction

Results – Model Performance Metrics

Overall Performance

- **Precision (Positive Predictive Value):** 98.2%

  = Proportion of data points extracted by the model that were correct

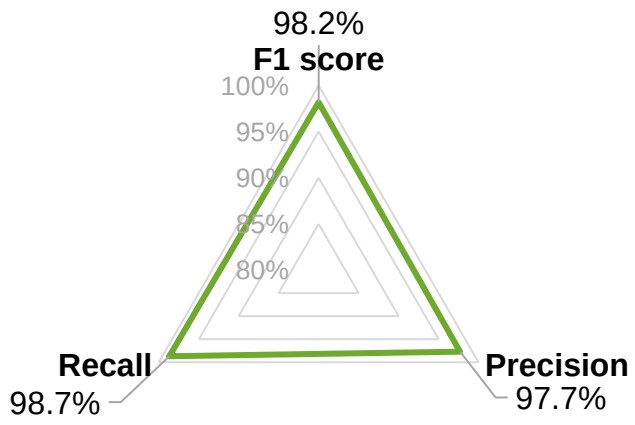
- **Recall (Sensitivity):** 96.6%

  = Proportion of all relevant data points that the model successfully extracted

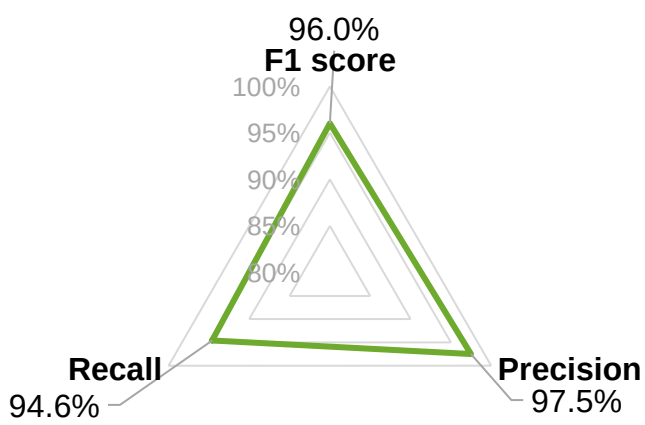
- **F1-score:** 97.4%

  = Harmonic mean (balance) of precision and recall

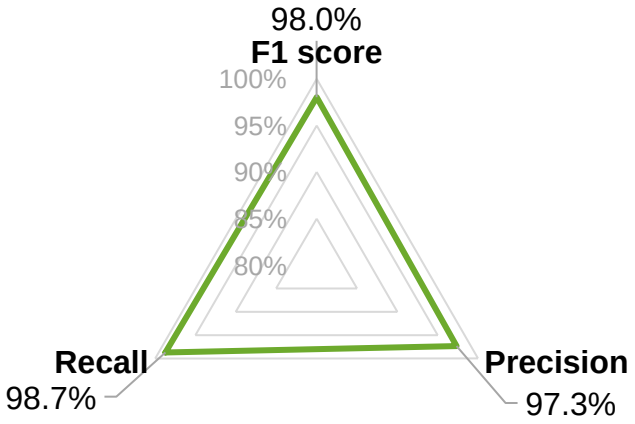
Study Characteristics



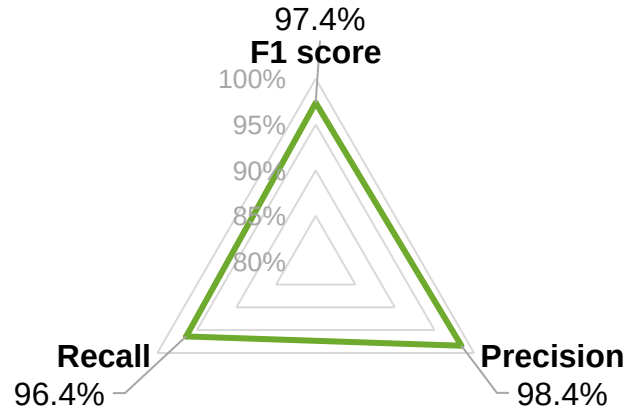
Intervention Characteristics



Participant Characteristics



Outcomes



Results – Error Analysis

Extraction Errors (False Positives)

- **Most common cause:**
  The model extracted information that was not explicitly reported in the publication (e.g., assumption or hallucination)

Extraction Error examples

- Incorrect assumption of open-label studies
- Mismatch between actual vs. planned treatment
- Model extracted treatment-arm level data when only population-level data was reported
- Mislabeling or over-standardization of outcome names

Omission Errors (False Negatives)

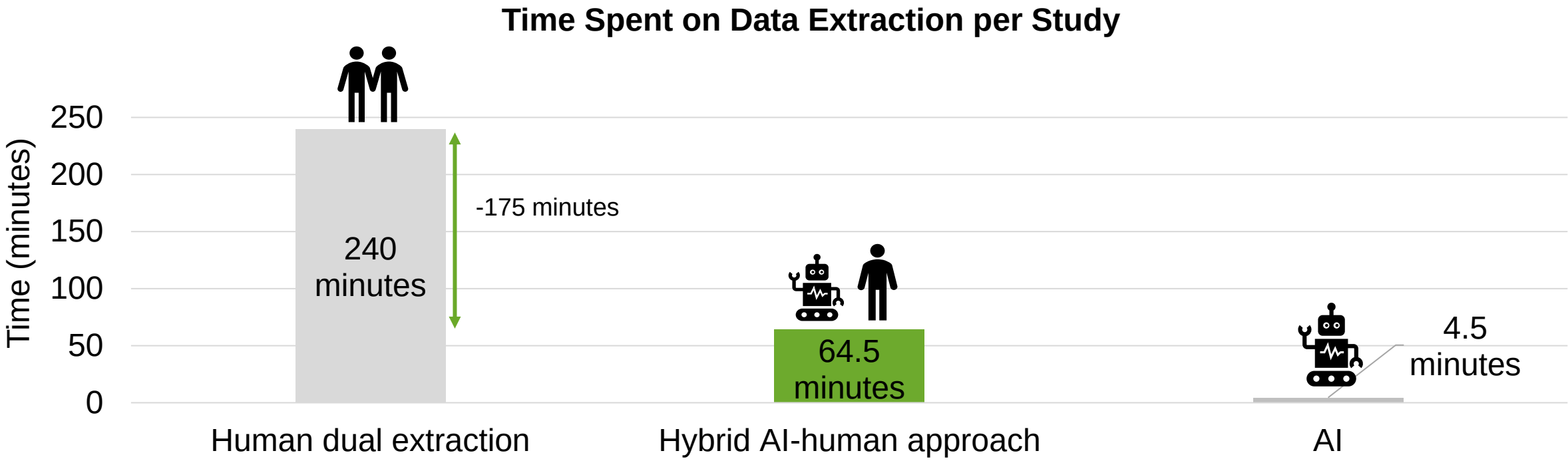
- **Most common cause:**
  The model missed implicitly stated or indirectly reported information and data with inconsistent terminology

Omission Error examples

- Safety assessment method often implied in publication (e.g., number of participants in safety analysis reported)
- Background therapy and line of treatment missed due to varying terminology across studies

These errors highlight areas for further improvement, particularly concerning handling implicit or inconsistent data.

Results – Efficiency Comparison



- Hybrid AI-driven approach with human oversight saves **almost 3 hours per study** compared to traditional dual extraction
- For example, for an **average-sized SLR with 50 studies** included for data extraction, this translates to **~145 hours of time saved** (**or roughly 3.5 weeks of 1 FTE**)

Discussion and Conclusions

- **First known study to comprehensively evaluate AI-based data extraction in SLRs at this scale** (117,889 data points across 106 variables)
- **High performance across domains**
  - Slightly lower recall in *Outcomes* and *Intervention Characteristics* data, indicating **room for refinement**
- **Hybrid AI-driven approach improves efficiency and data quality**
  - Saves **almost 3 hours of work time per study** from 1 FTE
  - **Helps reduce errors** from reviewer fatigue, variability between reviewers, or missed information
- In conclusion, the Claude 3.7 Sonnet-based AI model **demonstrated robust precision and recall** in oncology SLRs, **enabling faster data extraction while maintaining data quality**
- Through **ongoing efforts** to refine terminology, reduce errors, and improve generalizability across therapeutic areas, we aim to **further strengthen the model's overall performance**

Acknowledgments

- The authors would like to thank Jun Collet from Evidinno Outcomes Research Inc. for contributing to the presentation development.

Presenter Contact Information
Mir Sohail Fazeli, MD, PhD
Evidinno Outcomes Research Inc.
MFazeli@Evidinno.com

References

1. Reference Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545.

2. Pham B, Bagheri E, Rios P, et al. Improving the conduct of systematic reviews: a process mining perspective. Journal of clinical epidemiology. 2018;103:101-111.

3. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15(4):616-626.

4. Alqahtani T, Badreldin HA, Alrashed M, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Research in Social and Administrative Pharmacy. 2023;19(8):1236-1242.

5. Schmidt L, Hair K, Graziozi S, et al. Exploring the use of a Large Language Model for data extraction in systematic reviews: a rapid feasibility study. 2024.

6. Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of biomedical informatics. 2016;64:265-272.

7. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. JCPP Advances. 2024;4(2):e12234.

8. Jaspers S, De Troyer E, Aerts M. Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA. EFSA Supporting Publications. 2018;15(6):1427E.

9. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods. 2024;15(5):818-824.

10. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. Journal of clinical epidemiology. 2022;144:22-42.

11. Santos Á OD, da Silva ES, Couto LM, Reis GVL, Belo VS. The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review. Journal of biomedical informatics. 2023;142:104389.

12. Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022;13(3):353-362.

Abbreviations

AI, artificial intelligence

LLM, large language model

SLR, systematic literature review