ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/vhri

Methodology

Evaluating the Performance of Claude 3.7 Sonnet in Data Extraction Automation for Systematic Literature Reviews

Ellen Kasireddy, MHSc, Cuthbert Chow, MSc, Jun Collet, MSc, Mir-Masoud Pourrahmat, MSc, Mir Sohail Fazeli, PhD

ABSTRACT

Objectives: To evaluate the performance of Claude 3.7 Sonnet in automating data extraction for systematic literature reviews (SLRs).

Methods: An artificial intelligence (AI) extraction model based on the Claude 3.7 Sonnet large language model was developed through a structured process, including targeted training using a master data list and selected full-text articles. The master data list enhanced the model's contextual knowledge, guiding data extraction. Seven full-text articles from 4 oncology-focused treatment efficacy and safety SLRs were used for early testing and iterative refinement through error analysis. Model performance was then evaluated using 20 full-text articles, drawn from the same SLRs but not used for model development, and benchmarked against human extractions. Evaluation metrics included precision, recall, and F1 score. Extraction time was also compared across 3 different approaches: AI model-only, hybrid (AI model with human verification), and traditional human extraction.

Results: The AI model extracted 117 889 data points across 106 variables, achieving an overall precision of 98.2%, recall of 96.6%, and F1-score of 97.4%. Extraction performance was highest for Study Characteristics (precision: 97.7%, recall: 98.7%) and Participant Characteristics (precision: 97.3%, recall: 98.7%). Outcome data showed 96.4% recall and 98.7% precision. Intervention Characteristics achieved 97.5% precision and 94.6% recall. Extraction using the AI model alone averaged 4.5 minutes per article, compared with 64.5 minutes with the hybrid approach and approximately 240 minutes with traditional human extraction.

Conclusions: The Claude 3.7 Sonnet-based model demonstrated strong performance, supporting efficient and reliable AI-driven data extraction in oncology SLRs, with potential for broader applicability.

Keywords: artificial intelligence (AI), data extraction, large language models (LLMs), performance evaluation, systematic reviews.

VALUE HEALTH REG ISSUES. 2026; 53:101539

Introduction

Systematic literature reviews (SLRs) are foundational to evidence synthesis in health economics and outcomes research (HEOR), supporting key deliverables, such as cost-effectiveness analyses, global value dossiers, and health technology assessment (HTA) submissions. However, SLRs are time consuming and resource intensive, with manual extraction of data from large volumes of literature remaining a significant bottleneck and prone to errors and inconsistencies.[1,2] The accelerating growth of biomedical literature further underscores the need to enhance the efficiency, accuracy, and scalability of SLR workflows.

Advances in artificial intelligence (AI) offer promising solutions. By automating labor-intensive tasks, such as data extraction, AI tools can streamline SLRs, accelerating review timelines while potentially reducing associated human error.[3–12] In particular, large language models (LLMs) have shown strong potential in this space because of their ability to interpret complex medical language, reason contextually, and adapt to varied reporting formats.[13,14] These models can be applied to extract patient populations, intervention characteristics, outcome measures, and study characteristics, tasks traditionally performed by human reviewers.

These capabilities are especially relevant in HEOR, in which structured data on clinical endpoints, health resource utilization, and health state utilities are relevant to payer and reimbursement decisions. LLM-assisted extraction can enhance consistency and reduce the workload associated with synthesizing evidence for economic modeling and reimbursement submissions.

However, significant challenges remain. Biomedical literature exhibits considerable heterogeneity in format, terminology, and reporting structure, and LLMs often struggle with ambiguous phrasing, missing information, or interpreting sophisticated analytical outputs, such as subgroup analyses, adjusted estimates,

or time-to-event outcomes. Additionally, ensuring consistency across documents, accurately processing numerical data, and distinguishing closely related variables (eg, progression-free versus overall survival) require specialized, domain-specific tuning. These complexities underscore the need for rigorously validated AI models that are tailored to the needs of HEOR.

Although several tools have emerged, few of these tools have undergone robust validation across a comprehensive set of data elements. Published studies on LLM-assisted data extraction have typically focused on feasibility or evaluation of limited data elements, rather than comprehensive performance across diverse data categories.[5,15] This knowledge gap is particularly critical because data extraction challenges vary considerably, including those reported in complex and nonstandardized ways. This gap limits the broader adoption of LLMs in SLR workflows.

The application of LLMs in SLRs represents a globally relevant innovation. The need to streamline evidence synthesis is particularly pressing in low- and middle-income countries (LMICs), in which HEOR capacity may be constrained by limited funding, infrastructure, and trained personnel. LLM-based automation has the potential to support more equitable access to high-quality evidence synthesis in resource-constrained settings by reducing reliance on manual resources. To address persistent health system challenges, including workforce shortages, data collection, and decision support, generative AI has been recognized as a vital tool particularly in LMICs.[16]

To investigate the potential of AI-driven data extraction, we developed a customized AI model using Claude 3.7 Sonnet, an advanced LLM by Anthropic. Claude 3.7 Sonnet was selected for its superior capability in handling complex natural language tasks, including document summarization, question answering, and entity extraction, as well as its enhanced reasoning clarity and reduced biases compared with earlier models.[17]

This study aimed to evaluate the performance and efficiency of the customized Claude 3.7 Sonnet-based AI model for automating data extraction in SLRs. We assessed the model's performance by comparing its output with verified human-extracted data. Additionally, we evaluated efficiency by measuring the time required across 3 approaches: fully automated AI model extraction, a hybrid method (AI model extraction with human oversight), and traditional human extraction. These findings offer insights into the current capabilities and limitations of AI-assisted data extraction and inform future improvements to ensure methodological rigor, practical utility, and accessibility of evidence synthesis in HEOR.

Methods

AI Model Overview

A custom-built AI data extraction model (hereafter called "AI model") was developed to automate the retrieval of information from published articles utilizing Claude 3.7 Sonnet (nonextended thinking model run). Claude natively supports direct PDF uploads and features an expansive context window of 200 000 tokens (individual units of text such as words, characters, or subwords), allowing it to process entire articles within a single prompt. It also offers multimodal capabilities for handling diverse types of input.[17] At the time of writing, this model had the best average performance across public LLM benchmarks and evaluations,[18,19] and within our own prior anecdotal experience, it produced the most favorable results and was thus chosen as the language model for our data extraction model.

The extraction pipeline integrates the LangChain framework for structured model interaction, Zod schema validation to enforce consistent data structures, and structured output parsing to ensure compatibility with database formats. The AI model, powered by Chunkr.ai, processes PDF documents and converts unstructured text into structured data essential for SLRs, improving accuracy and accelerating the review process.[17]

AI Model Training Set

We reviewed 9 previously completed and validated SLRs on therapeutic efficacy and safety in oncology and compiled study, intervention, participant, and outcome data elements from their data extraction sheets. These elements were consolidated in a master data list, accompanied by definitions, to provide domain-specific contextual knowledge to the AI model, enhancing its understanding of oncology treatment efficacy and safety-related SLRs.

To simulate real-world data extraction during model training and early testing, a training set of 7 full-text articles, randomly selected from 4 of the 9 SLRs, was used to refine the AI model iteratively.

AI Model Prompt Development and Instruction

We implemented a structured prompting approach that combined role-based system instructions, domain-specific context, detailed extraction schemas, and in-context examples (Fig. 1). Prompts and schemas were tailored to specific study types and therapeutic areas, enabling the AI model to interpret specialized terminology and generate outputs in predefined formats.

Structured output parsing was a central feature of this design, in which prompts specified the targeted information and the required extraction schema, defined in JavaScript Object Notation (JSON) format. The schema included annotations for data types (eg, string and integer), optionality (ie, whether it was required to be extracted), and field descriptions (see Appendix Fig. 1 in Supplemental Materials found at https://doi.org/10.1016/j.vhri.2025.101539). Both the prompts and code pipeline were designed to ensure schema compliance (see example in Appendix Fig. 2 in Supplemental Materials found at https://doi.org/10.1016/j.vhri.2025.101539).

For the extraction of outcomes and participant characteristics, therapeutic area-specific examples were provided to the model—essentially a master list of the potential values that may be present in the study—to help the model run in a "few-shot learning" configuration.

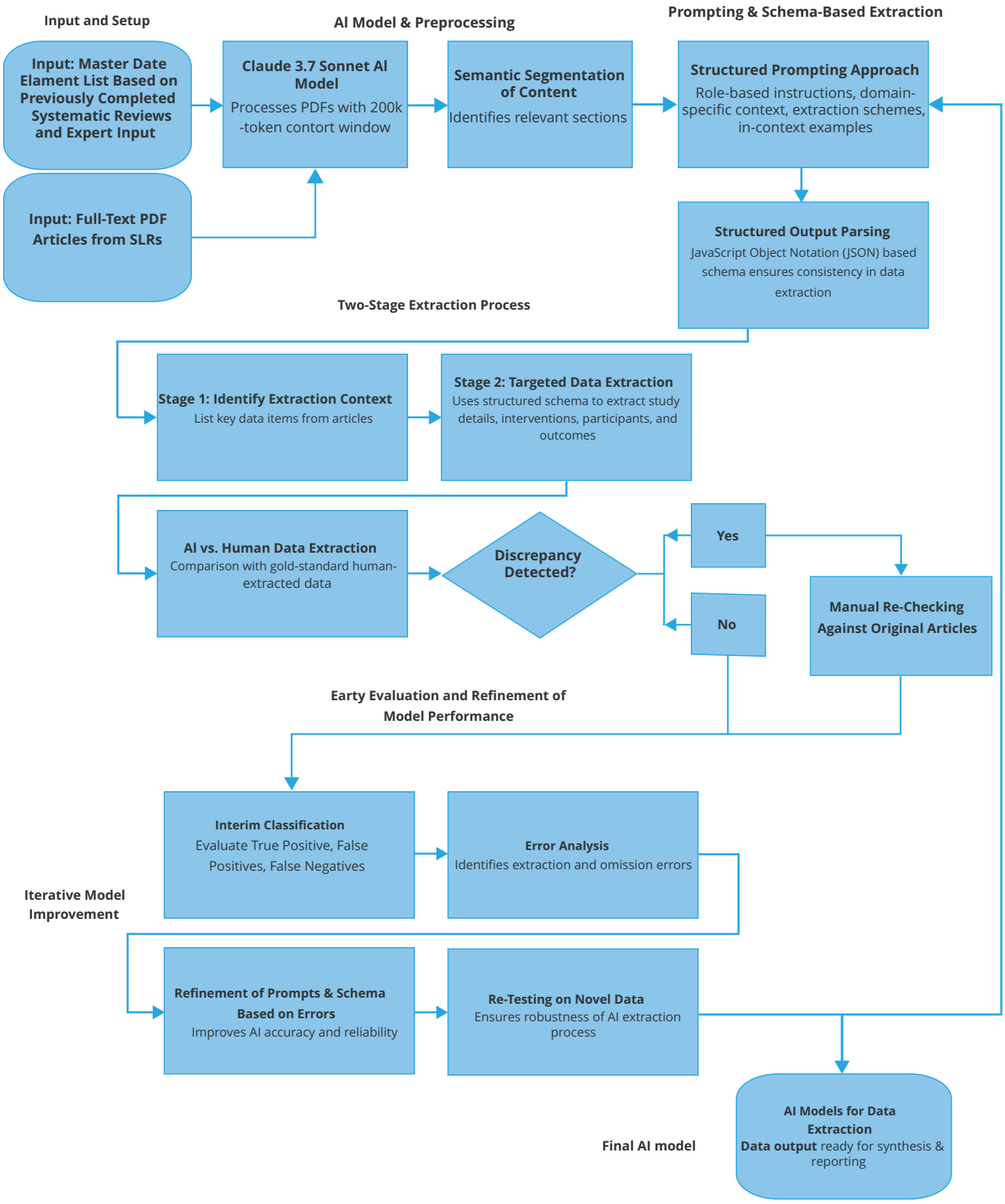
Data extraction followed a 2-stage process. First, an "extraction context" was established using the master data list, covering study, intervention, participant, and outcome elements. Then, extraction was performed on smaller segments within this context.

To mitigate bias and enhance performance, we refined prompts and schemas using the training set, maintaining a strict training/testing split. The AI model was run at a "Zero Temperature" setting to ensure deterministic outputs for identical inputs and was instructed to return verbatim source text for key variables to ensure traceability to the original content.

AI Model Extraction and Output

The AI model extracted data for multiple variables and recorded them in an Excel extraction sheet using various data formats, including numerical values, structured data elements, and free-text entries. In this context, variables refer to characteristics or attributes measured or reported in the included studies. Each variable may correspond to a single data point or multiple data points, with each data point representing a specific

Figure 1. Model prompt development and instructions.



piece of information captured in a single cell of the extraction sheet.

Single-data-point variables are often fixed study characteristics that yield only 1 value per study. Examples include study design (eg, randomized controlled trial and cohort study), total sample size (eg, 500 participants), and follow-up duration (eg, 12 months). In contrast, multi-data-point variables contain multiple values extracted from a single study. These typically arise from subgroup analyses, multiple time points, or distinct outcome measures. Examples include age distributions (eg, mean age reported separately for males and females), treatment arms (eg, different dosages or drug combinations tested), and outcome measures (eg, adverse events stratified by age group or disease severity). Unlike single-data-point variables, multi-data-point variables were given a more structured extraction instruction to ensure consistent and comprehensive capture across studies.

Model Validation

The performance of the AI model was evaluated using 20 randomly selected publications across 4 SLRs. These articles had not been used in training. Extracted data included study characteristics, intervention characteristics, patient characteristics, and outcomes (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.vhri.2025.101539) and were compared with verified human-extracted data, which served as the gold standard. Human-extracted data were obtained from 2 reviewers who independently performed data extraction following the Cochrane Handbook for Systematic Reviews of Interventions.[20] A third reviewer resolved discrepancies between the 2 sets of extracted data by verifying the correct values. Additionally, a fourth senior methodologist independently reviewed and verified the results to ensure the accuracy and reliability of the extracted data for use as the gold standard.

Performance Metrics

Data points were defined as follows:

- True Positives: Correctly extracted data points.
- False Positives: Instances in which the AI extracted incorrect, misclassified, or unnecessary data.
- False Negatives: Instances in which the AI failed to capture relevant information that was present in both the publication and the gold standard.

True Negatives refer to irrelevant data points that were correctly not extracted. These could not be systematically counted because the gold standard contained only relevant data.

The AI model's performance metric was evaluated using the following metrics:

- Precision (positive predictive value): The proportion of correctly extracted data points among all AI-extracted data points, reflecting the correctness of the AI-extracted data, is calculated as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall (sensitivity): The proportion of correctly extracted data points among all relevant data points in the gold standard, indicating the AI's ability to capture all relevant data, is calculated as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- F1-score: The harmonic mean of precision and recall, providing a balanced measure of accuracy, is calculated as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Error Analysis

Errors were classified into extraction errors (false positives) and omission errors (false negatives). A systematic analysis was conducted to understand the root causes of these errors, evaluate their impact on evidence synthesis, and identify areas for improvement. This process included an error rate assessment at both the individual data point and the broader variable levels. Errors were also examined across key data domains, including study characteristics, intervention characteristics, participant characteristics, and outcomes, to identify patterns within different sections of publications. In addition, we explored the presence of systematic issues, such as complex sentence structures, varying terminologies, and specific data formats that posed challenges for the AI.

Efficiency Comparison

To evaluate efficiency, we compared the time required for data extraction across 3 approaches: (1) manual extraction by 2 independent human reviewers, (2) automated extraction by the AI model, and (3) a hybrid approach, in which a human reviewer verified and refined the AI model's outputs. Time was recorded for each approach to assess potential time savings associated with automation and human-AI collaboration.

Results

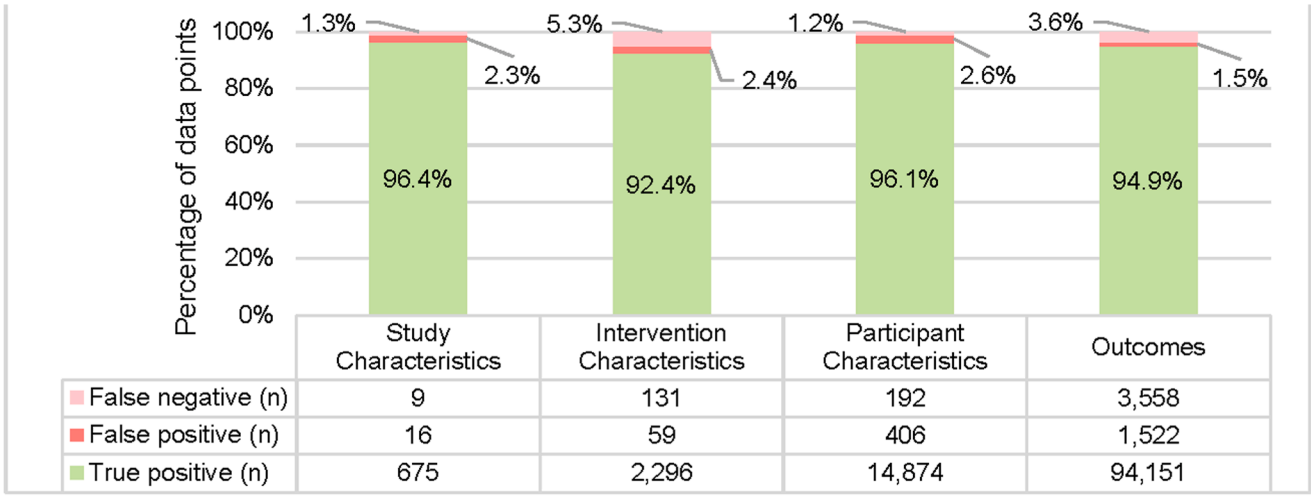
AI Model Data Extraction Outputs

The AI model was finalized through 10 iterative refinements. A total of 117 889 data points were extracted (Fig. 2). The Outcomes category accounted for the largest number of extracted data points (99 231), resulting from the detailed and multidimensional nature of outcome reporting, including multiple measures, multiple time points, and multiple treatment arms with different dosages, combinations, or intervention strategies.

The model demonstrated a high level of correctly extracted data points across all data domains, with true-positive rates ranging from 92.4% in Intervention Characteristics to 96.4% in Study Characteristics. A small proportion of data were missed, including 5.3% (131 data points) in Intervention Characteristics, 3.6% (3558 data points) in Outcomes, 1.3% (9 data points) in Study Characteristics, and 1.2% (192 data points) in Participant Characteristics. Incorrect extractions were also minimal, with error rates of 2.6% (406) in Participant Characteristics, 2.4% (59) in Intervention Characteristics, 2.3% (16) in Study Characteristics, and 1.5% (1,522) in Outcomes.

Of the 106 variables across 4 data domains (Fig. 3), Study Characteristics had the highest proportion of perfectly extracted variables, with 20 out of 35 (57.1%) variables achieving 100% correctness and 13 variables achieving 90% to 95% correctness. However, the variable Safety Assessment Method was extracted with 80% to 89%

Figure 2. Classification of all AI extracted data points (n = 117 889) by data domain.



	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

n = number of data points.

AI indicates artificial intelligence.

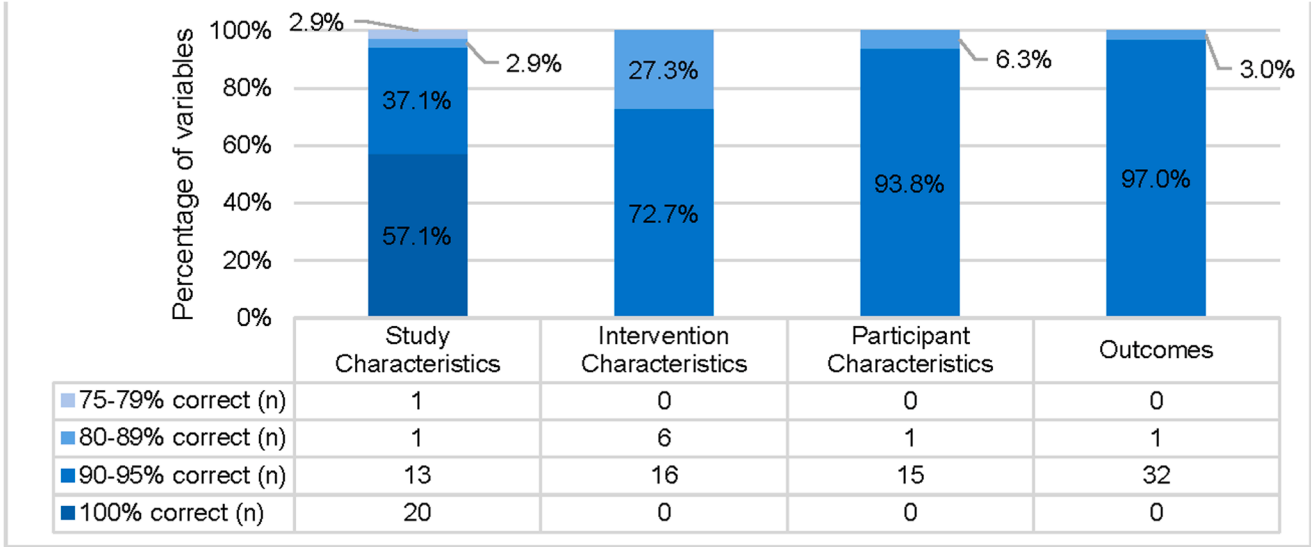
correctness, and Follow-up Estimate Type with 75% to 79% correctness. For the Intervention Characteristics domain, 16 variables reached 90% to 95% correctness, whereas 6 fell within the 80% to 89% range, including Crossover Criteria, Background Therapy, Treatment Duration Value, Treatment Duration Unit, Maintenance Dose Value, and Maintenance Dose Unit. Notably, no variables in this domain fell below the 80% correctness threshold. In the Participant Characteristics domain, Estimate Value was the only variable extracted with 80% to 89% correctness, whereas all other variables exceeded 90% accuracy. For the Outcomes domain, Standardized Outcome Name was extracted with 80% to 89% correctness, but no variables fell below this range.

An overview of variables that demonstrated less than 90% correctness and require further improvements is presented in Table 1.

AI Model Performance Metrics

The AI model achieved an overall precision of 98.2%, meaning that most extracted data points for all 106 variables were extracted correctly. Recall was 96.6%, indicating that the majority of relevant data points present in the source were successfully captured. The F1-score, which balances precision and recall, was 97.4%, demonstrating the model's strong overall performance (Fig. 4).

Figure 3. Correctness of AI-extracted variables (n = 106) by data domain.



	Study Characteristics	Intervention Characteristics	Participant Characteristics	Outcomes
75-79% correct (n)	1	0	0	0
80-89% correct (n)	1	6	1	1
90-95% correct (n)	13	16	15	32
100% correct (n)	20	0	0	0

n = number of variables.

AI indicates artificial intelligence.

Table 1. Variables with less than 90% correctness by data domain.

Correctness	Study characteristics (n = 2)	Intervention characteristics (n = 6)	Participant characteristics (n = 1)	Outcomes (n = 1)
80%-89%	• Safety assessment method	• Crossover criteria • Background therapy • Treatment duration value • Treatment duration unit • Maintenance dose value • Maintenance dose unit	• Estimate value	• Standardized outcome name
75%-79%	• Follow-up estimate type	None	None	None

Performance remained consistently high across all data domains (Fig. 5). For Study Characteristics, the model achieved an F1-score of 98.2%, with 97.7% precision and 98.7% recall, suggesting a well-balanced extraction process with minimal errors. Similarly, for Participant Characteristics, the model performed strongly, with an F1-score of 98.0%, precision of 97.3%, and recall of 98.7%, indicating that nearly all relevant participant-related data points were successfully captured.

In the Outcomes domain, the model maintained a high performance, with an F1-score of 97.4%, precision of 98.4%, and recall of 96.4%, demonstrating reliable outcome extraction with a low error rate. The Intervention Characteristics domain had the lowest recall at 94.6%, despite a high precision of 97.5%, resulting in an F1-score of 96.0%. This suggests that although most extracted intervention-related data were correct, some relevant data points were not captured.

Error Analysis

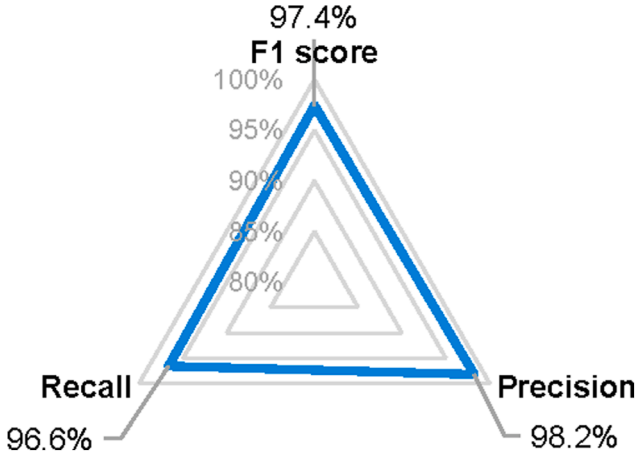
Among extraction errors, the variable with the highest rate was the Follow-Up Estimate Type in Study Characteristics, with 25% false positives, primarily due to the AI model extracting data where none was explicitly reported. Blinding had a 10% error rate because the AI model classified studies as "open label" when no information on blinding was provided in the publication. This assumption made by the AI model was incorrect because the appropriate classification should have been "not reported." For Intervention Characteristics, errors often stemmed from discrepancies between actual and planned durations or dosing schemes, such as for Treatment Duration Value and Unit. Maintenance Dose Value and Unit were frequently misidentified because of publications reporting only a single, general dose rather than distinguishing between a separate loading dose and maintenance dose.

In the Participant Characteristics domain, Estimate Value—representing the numerical value associated with a participant characteristic—had a 10% error rate. This error is primarily due to the model's difficulty distinguishing between treatment-arm level data and overall population-level data. In cases in which publications only reported participant characteristics for the overall population, the model incorrectly attempted to extract or infer treatment-arm level estimates, leading to inaccurate outputs based on assumptions and unsupported calculations. For the Outcomes domain, Standardized Outcome Name had a 10% error rate, often due to mislabeling (eg, classifying "grade 3 adverse events" as "adverse events" or standardizing "objective response rate" as "complete response").

Omission errors were largely driven by implicit or indirectly reported information. Among Study Characteristics, Safety Assessment Method had the highest omission rate (10%) because safety details were often implied rather than explicitly stated. Database Lock Date had a 10% omission rate, primarily due to alternative terminologies such as "cut-off date." Other study variables, (including National Clinical Trial) Identifier, Crossover Type, and Efficacy Assessment Method, each had a 4% omission rate due to terminology inconsistencies.

For Intervention Characteristics, Background Therapy (14%) and Line of Treatment (9%) were frequently missed because of inconsistent terminology across publications. Maintenance Dose Value and Unit had a 6% and 7% omission rate, respectively, often due to incomplete or complex dosing descriptions. Omission errors in Participant Characteristics were low (1%). In Outcomes, missing values were also low, with Number of Participants in Intervention Group at 3.9% and the remaining variables at 3.6%. A full overview of errors across variables is provided in Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.vhri.2025.101539.

Figure 4. Overall performance metrics of the AI model.

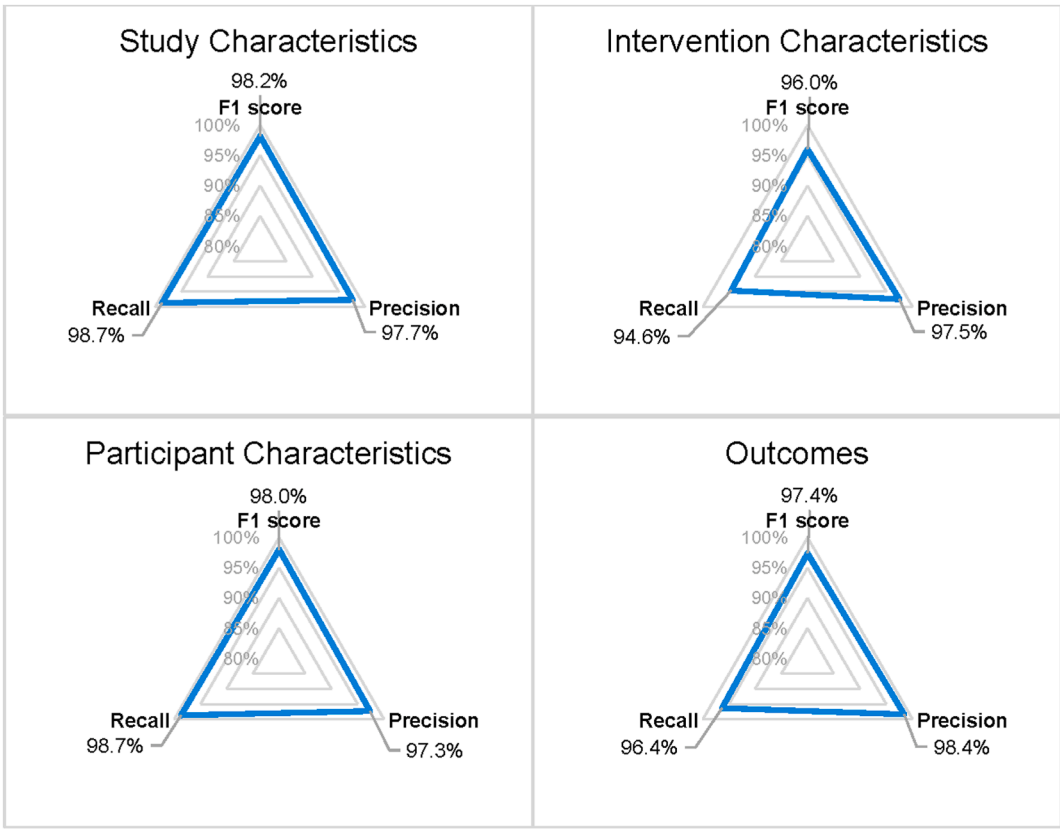


AI indicates artificial intelligence.

Efficiency Comparison

The articles that were extracted for this study averaged 8 pages in length, with a total range from 1 to 43 pages per study. The AI model required an average of 0.5 minutes for PDF parsing and approximately 4 minutes for full data extraction, resulting in a total processing time of 4.5 minutes per full-text article. In comparison, based on internal estimates, manual data extraction took approximately 2 hours (120 minutes) per full-text article when performed by a single reviewer, and 4 hours (240 minutes) when conducted in duplicate. These estimates did not account for time spent reconciling discrepancies between reviewers, which

Figure 5. Performance metrics of the AI model by data extraction type.



AI indicates artificial intelligence.

would have added additional time per study. Compared with duplicate human extraction, the AI model reduced the time required by approximately 235 minutes, a 98% decrease, per article.

Using a hybrid AI-driven extraction approach with human oversight, a manual review of the AI-extracted data required an average of 1 hour (60 minutes) per article. This estimate included time spent reconciling discrepancies between the AI-extracted output and human judgment, as reconciliation occured during the review process. Thus, the total time for the hybrid approach was approximately 64.5 minutes per study. Compared with full manual extraction, the hybrid approach reduced the time by approximately 175 minutes (73%) per article, still demonstrating a significant efficiency improvement.

In terms of cost, the average data extraction run in this study cost US dollars $2.8 per extraction, with the costs primarily comprising fees paid to Chunkr.ai for document parsing and the language model provider (Anthropic) for AI credits. When comparing the average cost of human labor for performing analogous tasks, there can be significant cost saving, even when using the hybrid AI-drive extraction approach described above.

Discussion

This study aimed to evaluate the performance and efficiency of a custom AI model developed to automate data extraction in SLRs, a foundational methodology in HEOR. To our knowledge, this is the first large-scale evaluation of an LLM-based model using a validated data set derived from previously conducted SLRs.

The model, built on Claude 3.7 Sonnet, was designed to streamline the extraction of structured data from clinical studies. Key technical advantages include the ability to directly process PDFs, handle entire articles in a single prompt, and integrate multimodal information (eg, text, figures, and document layout). The model was developed using structured extraction schemas informed by validated SLRs and refined through 10 iterative cycles of review and improvement. During this process, targeted extraction rules were implemented to reduce unnecessary inference, improve recognition of implicit or alternative terminology, and strengthen the model's ability to distinguish between similar data elements, contributing to enhanced performance.

A rigorous model evaluation of 20 full-text articles spanning 4 oncology-focused SLRs demonstrated high overall performance, with overall precision, recall, and F1-scores exceeding 96%. The model performed best in extracting Study Characteristics and Participant Characteristics. Although Outcomes data showed slightly lower recall, it remained above 95%, with consistently high precision. The Intervention Characteristics domain exhibited the lowest recall and F1-score, highlighting an area for targeted improvement.

In addition to its excellent performance, the AI model showed notable efficiency gains. This efficiency is particularly important given that data extraction is often the most time-consuming stage in SLRs. Prior studies report that SLRs can take more than a year to complete, with data extraction alone accounting for nearly a quarter of total effort.[1,2] In the current evaluation, a hybrid AI-human approach reduced total effort by nearly 4-fold compared with full manual extraction. This level of efficiency is

particularly valuable for HEOR deliverables, such as HTA submissions, global value dossiers, and cost-effectiveness analyses, which are often time sensitive.

The role of AI in supporting evidence synthesis and generation for health policy and HTA submissions is increasingly acknowledged by HTA bodies and professional societies. The United Kingdom's National Institute for Health and Care Excellence has issued a position statement emphasizing that AI-derived evidence must meet the same standards of transparency, reproducibility, and clear reporting as traditional methods.[21] Building on this, Canada's Drug Agency has also released a position statement recognizing the potential of AI while stressing the importance of transparency, appropriateness, and trustworthiness in its application.[22] In parallel, the ISPOR Working Group on Generative AI recently outlined a framework for evaluating and applying AI in HEOR contexts, highlighting both opportunities, such as automating aspects of SLRs, and challenges, including the need to maintain scientific rigor, ensure reliability, and mitigate bias.[23] All 3 organizations emphasize the critical role of human oversight in ensuring the responsible and effective use of AI tools in evidence synthesis. The hybrid AI-human approach evaluated in this study, featuring iterative model tuning, transparent error analysis, and human oversight, is closely aligned with these principles and demonstrates how AI can be responsibly integrated into HEOR workflows.

Importantly, the relevance of this approach is not limited to high-income countries. In LMICs, limited access to trained personnel and analytic infrastructure often constrain HEOR capacity. LLM-based automation offers a scalable solution enabling timelier and resource-efficient evidence synthesis. As LMICs expand HTA processes, such tools can support locally relevant, data-driven decision making and help promote more equitable access to care.

An error analysis identified that most extraction errors stemmed from implicit reporting, inconsistent terminology, or incorrect inferences. These included misinterpretations of population-level versus treatment-arm data and issues with variable definitions, particularly for outcomes and follow-up metrics. Although overall error rates were low, these findings underscore the importance of maintaining human oversight and continuing to refine AI extraction prompts.

An important consideration in applying AI to systematic reviews is the risk of bias. Models may reflect limitations of their training data, amplify reporting inconsistencies in biomedical literature, or infer details that are implied but not explicitly stated. Such biases can distort evidence synthesis if not carefully monitored, underscoring the need for human oversight to validate outputs and ensure methodological rigor.

Although the AI model was developed and evaluated through a rigorous process, the study has several limitations. First, we were unable to assess true negatives—instances in which irrelevant data were correctly not extracted—limiting our ability to calculate accuracy as a performance metric. Second, the model was trained and validated exclusively on oncology studies focused on treatment efficacy and safety. Consequently, its generalizability to other therapeutic areas or types of SLRs, such as those examining humanistic burden, unmet needs, or economic outcomes, currently remains untested. However, many of the training techniques and prompt design elements could be readily adapted to other therapeutic areas with sufficient time and domain-specific examples. The overall extraction pipeline and core prompts would remain consistent, with additional few-shot examples improving performance for new therapeutic areas or SLR types. Expanding this work to broader domains represents an important direction for future research.

Given these limitations, a hybrid AI-human extraction approach currently represents the most practical and effective solution. This strategy leverages the efficiency of AI-driven extraction while maintaining human oversight to ensure quality. By automating repetitive tasks, the model helps alleviate reviewer fatigue, particularly when processing large volumes of articles, while allowing human reviewers to focus on more nuanced or ambiguous cases.

Looking ahead, the AI model's performance can be further strengthened through additional refinements targeting key areas of improvement. First, enhancing terminology standardization will reduce inconsistencies in how data elements are labeled and interpreted, ensuring more reliable extraction across diverse studies. Second, expanding the training data set to include a wider range of therapeutic areas will help improve the model's generalizability and prevent overfitting to oncology-specific content, enabling broader applicability across different domains of medical research. Finally, incorporating more robust synonym recognition and inference-prevention rules will improve the model's ability to interpret varied reporting styles and reduce errors caused by assumptions or implicit information. Together, these refinements will support the continued advancement of AI-driven data extraction as a reliable, scalable, and high-quality solution for modern evidence synthesis.

Conclusions

The Claude 3.7 Sonnet-based AI model showed strong performance in automating data extraction for oncology-focused SLRs, with notable gains in efficiency. A hybrid approach, combining AI with human oversight, offers a practical balance between speed and quality, supporting more scalable and reliable evidence synthesis. By streamlining one of the most resource-intensive stages of the SLR process, this approach supports more timely, consistent, and transparent evidence synthesis for application in HEOR and other research domains.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.vhri.2025.101539.

Article and Author Information

Accepted for Publication: September 30, 2025

Published Online: xxxx

doi: https://doi.org/10.1016/j.vhri.2025.101539

Author Affiliation: Evidinno Outcomes Research Inc, Vancouver, BC, Canada (Kasireddy, Chow, Collet, Pourrahmat, Fazeli).

Correspondence: Mir Sohail Fazeli, PhD, 411-63 W 6th Ave, Vancouver, BC V5Y 1K2, Canada. Email: mfazeli@evidinno.com

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: All authors report employment with Evidinno Outcomes Research Inc. (Vancouver, BC, Canada). The authors received no financial support for this research.

Role of the Funder/Sponsor: All authors report employment with Evidinno Outcomes Research Inc (Vancouver, BC, Canada).

Ethical Approval: The study is solely aimed at advancing knowledge and understanding in the field of AI-based data extraction, without involving any human subjects or sensitive data. Therefore, ethical considerations and associated procedures do not apply to this research.

Data Availability: The data supporting the systematic reviews used in this study are derived from publicly available sources. All included articles were accessed through databases such as Embase, MEDLINE, and PubMed. Primary data collected as part of the validation process, prompts, and model responses are available upon reasonable request.

REFERENCES

1. 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. *J Clin Epidemiol*. 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. *Res Soc Admin Pharm*. 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. In: 2024. https://arxiv.org/abs/2405.14445. Accessed November 6, 2025.
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. *J Biomed Inform*. 2016;64:265–272.
7. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. *JCPP Adv*. 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 Support Publ*. 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. *J Clin Epidemiol*. 2022;144:22–42.
11. Santos ÁOD, 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. *J Biomed Inform*. 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.
13. Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus*. 2023;15(12):e50881.
14. Mondal H, Gupta G, Sarangi PK, et al. Assessing the capability of large language model chatbots in generating plain language summaries. *Cureus*. 2025;17(3):e80976.
15. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. 2024;15(4):576–589.
16. Stanford Center for Digital Health. *AI for health in low- and middle-income countries*. Round Table [discussion]. 2025; 2025. https://cdh.stanford.edu/our-research-portfolio/generative-ai-health-low-middle-income-countries/ai-health-lmics-roundtable. Accessed November 6, 2025.
17. Claude 3.7 Sonnet and Claude code. Anthropic. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed February 24, 2025.
18. Chiang WL, Zheng L, Sheng Y, et al. Chatbot arena: an open platform for evaluating LLMs by human preference. In: *Proceedings of the 41st International Conference on Machine Learning*; 2024. Vienna, Austria. proceedings.mlr.press/v235/chiang24b.html. Accessed November 6, 2025
19. White C, Dooley S, Roberts M, et al. LiveBench: a challenging, contamination-limited LLM benchmark. arXiv. Preprint. https://arxiv.org/abs/2406.19314. Accessed November 6, 2025.
20. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5; updated 2024. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current; 2024. Accessed November 6, 2025
21. Use of AI in evidence generation: NICE position statement. National Institute for Health and Care Excellence (NICE). https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation–nice-position-statement. Accessed June 11, 2025.
22. Canada's drug agency position statement on the use of artificial intelligence in the generation and reporting of evidence. Canada's Drug Agency. https://www.cda-amc.ca/sites/default/files/MG%20Methods/Position_Statement_AI_Renumbered.pdf. Accessed June 11, 2025.
23. Fleurence RL, Bian J, Wang X, et al. Generative artificial intelligence for health technology assessment: opportunities, challenges, and policy considerations: an ISPOR working group report. *Value Health*. 2025;28(2):175–183.