AID-SLR: A Generative Artificial Intelligence-Driven Automated System for Systematic Literature Review

Kyeryoung Lee, PhD[1*], Surabhi Datta, PhD[1*], Hunki Paek, PhD[1], Majid Rastegar-Mojarad PhD[1], Liang-Chin Huang, PhD[1], Long He, MS[1], Siwei Wang, PhD[1], Jingqi Wang, PhD[1], Xiaoyan Wang, PhD[1]

[1]IMO Health, Rosemont, IL, USA

[*]Equal Contribution

Corresponding Author:

Xiaoyan Wang, PhD

IMO Health

Address: 9600 West Bryn Mawr Avenue Rosemont, IL 60018

United States

Phone: 2012828098

Email: xw108@caa.columbia.edu

Abstract

Systematic literature reviews (SLRs) are crucial for generating research evidence, supporting clinical decisions, advancing scientific knowledge, and informing policymaking. Despite their importance, manual SLRs are time-consuming, costly, and prone to errors. The increasing volume of published data and the complexity of clinical trials necessitate more efficient approaches. We present an automated SLR system using large language models (LLMs), designed to streamline the entire SLR process from initial query to data extraction, and customizable for various study fields. We developed an LLM-assisted SLR system, AID-SLR, accompanied by a user interface (UI) comprising 6 modules 1) Query, 2) Inclusion/Exclusion (I/E) criteria, 3) Abstract screening, 4) Full-text screening, 5) Data extraction, and 6) Data summary. The LLM model was utilized for abstract screening, full-text screening, and data extraction and its performance was evaluated using precision, recall, and F1 scores. We selected a non-small cell lung cancer use case to evaluate the system's performance. We additionally compared the performance of GPT-4 and GPT-4o models, focusing on data extraction across different categories. A qualitative evaluation was conducted to assess common error types and the reliability of extracted information. AID-SLR is user-centric, allowing users to specify study criteria and provide additional information and feedback. The LLM prompts are generalizable and automatically incorporate the user-entered details and instructions on the UI, such as domain-specific guidelines, thereby enabling easy adoption of the system to different study and disease areas. AID-SLR effectively screens relevant studies and extracts data elements. The system demonstrated high precision, recall, and F1 scores in screening both Irrelevant (1, 0.9286, and 0.9630, respectively) and Relevant (0.9737, 1, and 0.9867, respectively) articles, with an overall accuracy of 98.04%. Data extraction was granular with promising performance,

successfully identifying a wide range of treatment-related outcomes and statistical values. For data extraction, GPT-4o outperformed GPT-4, achieving higher precision (0.9984 vs. 0.9819), recall (0.9989 vs. 0.9519), and F1-score (0.9987 vs. 0.9651). GPT-4o also exhibited superior performance in cohort identification and value extraction, with fewer errors and more accurate capture of study design and demographic information. Our LLM system and UI provided a seamless end-to-end solution for automated SLRs. This automated SLR system can contribute to reducing the time, cost, and human errors associated with traditional manual SLRs. Integrating this model with other AI tools for comprehensive data analysis could further enhance its utility in SLRs.

Introduction

Systematic literature reviews (SLRs) are foundational in generating research evidence. They support clinical decisions, fill scientific knowledge gaps, and inform policymaking [1–3]. With the growing emphasis on evidence-based practice [4], the Food and Drug Administration (FDA)'s acceptance of real-world evidence (RWE) to support drug approval [5], and the health technology assessment (HTA) requirement for drug pricing and reimbursement guidance [6], SLRs have become indispensable methods for synthesizing high-quality, up-to-date evidence [7]. However, manual SLRs are labor-intensive, costly, and prone to errors. Conducting and publishing a single SLR typically takes between 12 to 24 months [8], with an average duration of approximately 17 months [9]. Major pharmaceutical companies spend over 5 million annually on these studies [10]. This substantial time and cost burden hinders the thorough conduct of SLR studies with the increasing volume of published data. Additionally, with around 20,000 new trials starting annually [11,12], the volume and complexity of ongoing clinical research create a pressing need for more efficient SLR approaches.

Automation has shown great potential to enhance SLRs [8], with various initiatives aiming to automate these processes [13–16]. Advances in artificial intelligence (AI), particularly in natural language processing (NLP), have notably accelerated SLR automation, especially in literature screening and data extraction [17–19]. The integration of large language models (LLMs) has further expedited each phase of the SLR process, though their implementation should be approached with caution and include human oversight [20,21]. LLMs have shown high accuracy in screening relevant titles, abstracts, and full-text [22,23], and effectively conducting quality assessment and risk-of-bias evaluation [24]. Moreover, the LLM system has been employed to automate the

extraction of Population, Intervention, Comparator, and Outcome (PICO) elements [25,26], the generation of evidence via data extraction [27,28], and the conduction of the network meta-analyses using generated R scripts and extracted data elements [29]. However, there has been limited exploration into the holistic integration of all steps [26], including querying relevant articles from databases like PubMed, screening abstracts and full texts based on the PICO eligibility criteria user-defined, and extracting user-specified data elements into a computable format for downstream analysis.

In this study, we detail the development of an automated SLR system, AID-SLR, that leverages LLMs and features a seamless end-to-end user interface (UI), incorporating human oversight. We apply this system to a non-small cell lung cancer (NSCLC) use case focusing on the first line of immunotherapy literature, demonstrating its potential to streamline the automated SLR process.

Materials and Methods

System Architecture and User Interface Overview

Our AID-SLR system consists of 6 modules: 1) Query Setup Module, 2) Inclusion/Exclusion (I/E) criteria Input Module, 3) Abstract Screening Module, 4) Full-text Screening Module, 5) Data Extraction Module, and 6) Data Summary Module. Users can select either abstract screening only or both abstract and full-text screening when creating a new project. The UI is designed to provide clear navigation through each module, ensuring an intuitive user experience. Figure 1 illustrates the system's architecture, and detailed functionality and implementation for each module are described below.

Module Functionality and Implementation Details

Each module is designed to ensure high accuracy and efficiency. We utilize the OpenAI GPT-4 model to develop the screening and data extraction modules. The LLM prompts are adaptable to various disease areas and user-defined study I/E criteria. The Supplementary Methods delineate a protocol for the UI application and LLM prompts, and screenshots of the UI design are available in Supplementary Figure 1A-F.

Query Setup Module

The system supports scalability and integration with two literature databases. It employs search APIs and techniques from literature databases such as PubMed and Embase to automatically expand queries. This includes mapping the search terms to medical subject headings terms, synonyms, and variants. For example, entering "lung cancer" will automatically expand to include "lung (adeno)carcinoma", "lung neoplasm", and other related terms.

Inclusion/Exclusion (I/E) Criteria Input Module

Users specify PICO (Population, Intervention, Comparator, Outcome) criteria for screening relevant articles in this module. Additionally, they specify any other criteria in free text under the "Other I/E Criteria" section. The system allows for the storage and retrieval of user-pre-defined PICO templates, ensuring that user inputs are seamlessly integrated into the screening workflow.

Abstract Screening Module

This module uses an LLM for the initial screening based on user-defined I/E criteria. The AI-recommended results can be adjusted through an interactive UI with human review. The LLM prompt for abstract screening considers user-specified information such as the I/E criteria and any domain interpretations along with generic screening instructions in its context to make the screening decision. Screening instructions are relaxed at this initial stage to encourage high recall. The model is guided to classify an article as "Irrelevant" if it matches at least one exclusion criterion and "Relevant" for all other scenarios, including unclear instances due to insufficient information in the abstract. As part of the screening output, the LLM provides an explanation that supports its decision and the specific exclusion reasons. Publication type and study design information are also considered in this screening prompt which are first identified using a separate prompt described in Supplementary Methods.

Full-Text Screening Module

This module enables the full-text screening of articles using stricter PICO I/E criteria if provided, and prompt instructions. The LLM first reviews the exclusion criteria and subsequently checks each inclusion criterion if no exclusion criteria are met. If all inclusion criteria are satisfied, the

article is classified as "Relevant." If the model is not fully confident, it labels the article as "Relevant - not confident," prioritizing it for further user review. This warrants more precise eligibility decisions and highlights the articles needing additional review. Explanations behind screening decisions and exclusion reasons are also generated in this module similar to abstract screening.

Data Extraction Module

We design a generalizable data extraction module with three prompts to identify: 1) Study details (e.g., sample size), 2) Study cohorts (different study arms), and 3) Study outcomes (e.g., overall survival). The user-provided data elements, along with any domain-specific descriptions and interpretations, are leveraged in constructing the context for the study details and outcome extraction prompts. The study cohorts returned by the second prompt are used in the Study outcomes (third) prompt to guide the association of each identified outcome with the relevant cohort. We highlight the corresponding text spans of the extracted data elements on the UI to facilitate easy manual review. Moreover, for data extraction from full-text articles, we extract the study outcome elements from both the tables and the main text of an article. For this, we separately process each table and the main text for LLM prompting.

Data Summary Module

This module compiles extracted data into a comprehensive summary. Users can review, filter, and download the summarized data. It provides a robust data management system, generating summary reports with key information such as PMID, citation title, authors, year of publication, and extracted data elements.

Evaluation

To evaluate the system's performance, we conduct both quantitative and qualitative assessments using the NSCLC use case. We utilized 10 articles to optimize the LLM prompts for screening and data extraction and assess the system's performance on a held-out set of 50 articles. For quantitative evaluation, we calculate the accuracy, precision, recall, and F1 scores for both screening and data extraction processes. Precision is calculated as the ratio of correctly predicted positive entities to the total predicted positive entities (Precision= TP/ (TP + FP)). Recall, also known as sensitivity, is calculated as the ratio of correctly predicted positive entities to all actual positive entities (Recall=TP/ (TP + FN)). The F1-score is the harmonic mean of precision and recall and is calculated using the formula: F1-score=2 × ((Precision × Recall) / (Precision + Recall)). In these equations, TP stands for true positives, FP stands for false positives, and FN stands for false negatives. Additionally, for abstract data extraction, we evaluate and compare the performance of two versions of GPT models, GPT-4 and GPT-4o, in two primary categories: Study Details and Study Outcomes. The qualitative evaluation involved analyzing common error types encountered during data extraction to gain insights into areas needing improvement.

Use Case: Non-Small Cell Lung Cancer

The inclusion and exclusion criteria, shown in Figure 2, were applied for the literature screening phases. During the abstract screening phase, broader criteria were used to capture a wide array of relevant studies. For instance, abstracts mentioning general categories like immunotherapy instead of specific drug names were included. The outcome criteria encompassed a wide range of treatment-related outcomes, such as efficacy, adverse events, treatment trends, and patient-

reported outcomes. In the full-text screening phase, more specific criteria were applied. For example, the inclusion intervention criteria were narrowed to include only studies specifically mentioning pembrolizumab.

Figure 1. The overview of the AID-SLR system

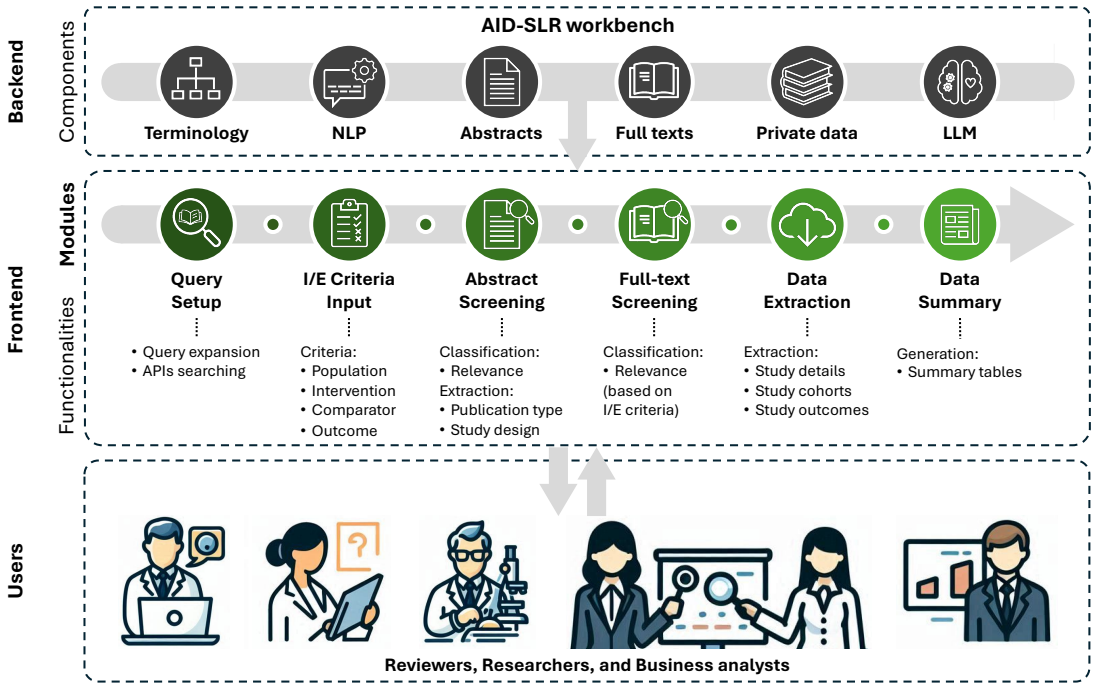
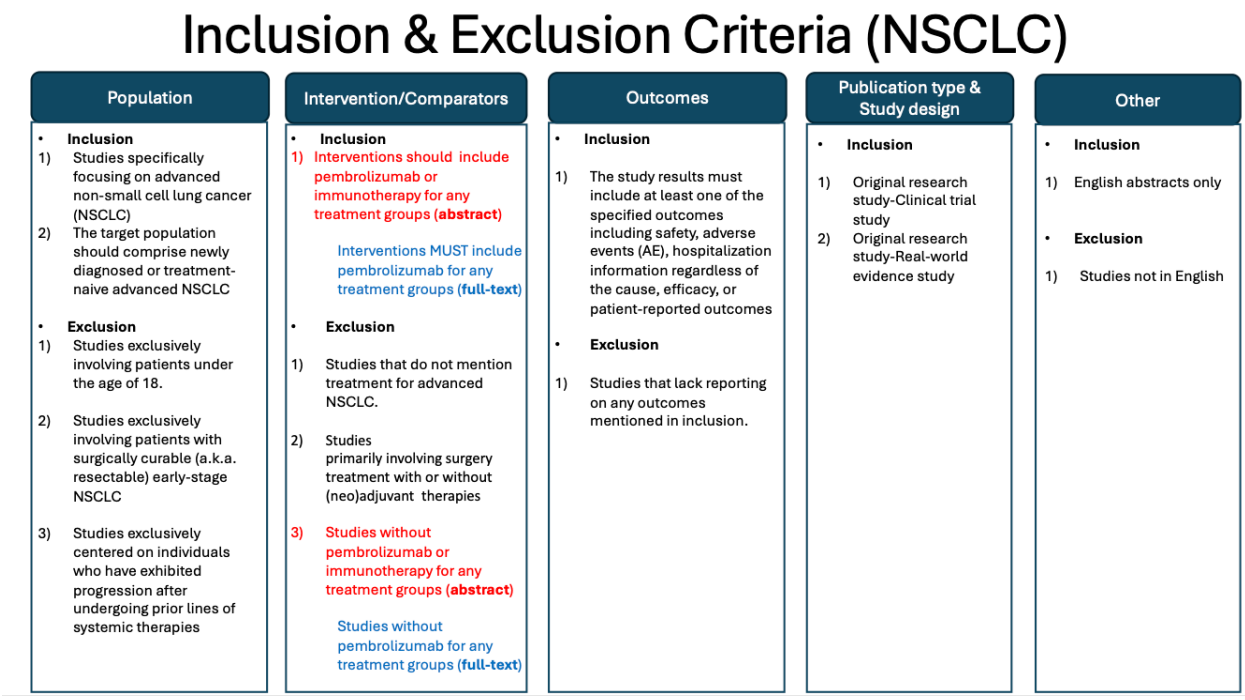


Figure 2. Eligible criteria for abstracts and full-text screening for non-small cell lung cancer. Black-colored criteria are for both, red-colored criteria are for abstract screening, and blue-colored criteria are for full-text screening.



Results

NSCLC Use Case Implementation on Our UI

The system found 2,135 articles when the user put the following query, limiting the publication year between 2023 and 2024.

- ◊ (((NSCLC) or (non-small cell lung cancer) AND ((advanced) or (metastatic) or (Stage III) or (stage IV)) AND ((immunotherapy) or (immune checkpoint inhibitor) or (pembrolizumab))) AND (English[Language])

Out of 2,135 articles, 898 and 1,172 abstracts were recommended as “Relevant” and “Irrelevant”, respectively (Supplementary Figure 1B), based on the PICO criteria we defined (Figure 2). 65 abstracts were not screened due to the absence of an abstract. We randomly selected relevant and irrelevant articles and reviewed the AI explanation. In “irrelevant” cases (Supplementary Figure 1C), the system shows the matched Exclusion criteria under the

"Exclusion details" section and provides a plain language explanation under the "AI explanation" section. For example, an article focusing on early-stage resectable NSCLC population, not advanced NSCLC, was marked as "Irrelevant". In contrast, in "relevant" cases (Supplementary Figure 1D), the "Exclusion details" section is empty, showing only the AI explanation for inclusion. Supplementary Figure 1E shows examples of user-selected progression-free survival (PFS) and overall survival (OS) outcomes values in each cohort with the corresponding text span where the information was extracted. For example,

- From the sentence, "Progression-free survival (PFS) was significantly shorter in patients with venous thrombotic events (VTE) compared to patients without VTE: 6.1 (95% CI 4.1-9.0) months vs. 8.3 (6.9-10.3) months (p=0.03)", PFS values were extracted for both Patients with VTE and without VTE cohorts.

Abstract and Full-Text Screening

The human reviewer annotated 36 out of the 50 articles as "Relevant" including 12 articles with a "Need to check full-text" tag and the remaining 14 as "Irrelevant". The precision, recall (sensitivity), and F1 scores for "Relevant" abstracts are 0.9737, 1, and 0.9867, respectively. For "Irrelevant" abstracts, the scores are 1, 0.9286, and 0.9630, respectively, with an overall accuracy of 98.04%. The macro and weighted averages for these performance metrics are shown in Table 1. The 12 "Relevant" articles with the "Need to check full-text" tag from the abstract screening stage were screened for full text, of which the human reviewer excluded 2 articles. The results of these 12 articles indicated perfect agreement between a human-annotated gold standard and AI recommendation, with both precision and recall (sensitivity) achieving scores of 1.00, resulting in an F1-score of 1.00. The system's overall accuracy was also 100%, reflecting the

efficient identification of studies that met the specific criteria for detailed clinical outcomes involving pembrolizumab.

Table 1. Performance metrics for abstract screening process on 50 articles

	Precision	Recall (sensitivity)	F1-score
Irrelevant	1.0000	0.9286	0.9630
Relevant	0.9737	1.0000	0.9867
Macro average	0.9868	0.9643	0.9748
Weighted average	0.9809	0.9804	0.9802
Specificity	0.9286	0.9286	0.9286

Data Extraction from Abstracts

Quantitative Evaluation

Table 2 presents the performance scores on 34 human-annotated final "Relevant" articles obtained after full-text screening. We observe high performance in both GPT-4 and GPT-4o models, with GPT-4o consistently outperforming GPT-4. Example output formats for extracted data elements are shown in Supplementary Tables 1 and 2. GPT4o achieved an overall precision score of 0.9984 compared to GPT-4's 0.9819. For recall, GPT-4o again showed an overall recall of 0.9989, while GPT-4 scored 0.9519. The F1-score for GPT-4o was 0.9987, compared to GPT-4' 0.9651(Table 2). These results demonstrate that GPT-4o is more precise and thorough in data extraction. Additionally, the accuracy metrics indicate that GPT-4o achieves near-perfect accuracy (0.9975) across all features, while GPT-4 scores were slightly lower (0.9379). We also

compare the performance of the models for individual features of the extracted elements such as value and text span. We found that GPT4o performs superiorly in study cohort identification and corresponding outcome value extraction. Detailed performance scores of GPT-4o and GPT 4 in identifying individual features are provided in Table 2 and illustrated in Figure 3.

Table 2. Performance metrics for data extraction from 34 abstracts

		Precision		Recall		F1-score		Accuracy	
		GPT-4	**GPT-4o**	**GPT-4**	**GPT-4o**	**GPT-4**	**GPT-4o**	**GPT-4**	**GPT-4o**
	Overall Average	0.9819	0.9984	0.9520	0.9989	0.9651	0.9987	0.9379	0.9975
Study Details	**Overall**	0.9938	1.0000	0.9956	1.0000	0.9946	1.0000	0.9895	1.0000
	Element	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Description	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Text span	0.9861	1.0000	0.9852	1.0000	0.9843	1.0000	0.9713	1.0000
	Value	0.9676	1.0000	0.9871	1.0000	0.9751	1.0000	0.9547	1.0000
Study Outcomes	**Overall**	0.9701	0.9968	0.9083	0.9979	0.9356	0.9973	0.8863	0.9950
	Element	0.9914	1.0000	0.9471	0.9977	0.9663	0.9988	0.9393	0.9977
	Description	0.9827	0.9923	0.9455	0.9968	0.9612	0.9943	0.9319	0.9907
	Text span	0.9672	1.0000	0.9213	0.9977	0.9341	0.9988	0.8963	0.9977
	Study cohort	0.9476	0.9880	0.7511	0.9968	0.8129	0.9920	0.7380	0.9864

	Value	0.8965	1.0000	0.8336	0.9977	0.8595	0.9988	0.8231	0.9977

Figure 3. Comparison of precision, recall, and F1 scores between GTP-4 and GTP-4o in various features of extracted data elements.

Error Analysis

The error analysis of data extraction between GPT-4 and GPT-4o models reveals several critical differences, as summarized in Table 3. In the domain of "Study Details," GPT-4 frequently misses essential information such as study design specifics (e.g., "randomized") and demographic details like sex ratios and prior treatments. Additionally, GPT-4 often misrepresents or omits details about the intervention. In contrast, GPT-4o rarely makes these

mistakes, demonstrating more consistent accuracy in capturing study design and demographic information, thereby providing a more reliable and complete extraction of study details.

In the "Study Outcome" category, GPT-4 performs well but has some limitations, particularly in accurately specifying study cohorts and subgroups. These errors were frequent in comparison studies involving multiple clinical trials. GPT-4o, however, performs much better in this regard, although it occasionally presents incorrect cohort names or confuses experimental groups with reference/comparator groups. Such errors are infrequent in GPT-4o, indicating higher precision and reliability in cohort-related data extraction.

In terms of data elements and statistical values, GPT-4 often misses important elements such as "Study duration" and "Data cut-off" dates and struggles to separate statistical values like hazard ratios (HR), odds ratios, 95% confidence intervals (CI), ranges, and p-values. This often results in marking values as "NA" or providing incorrect information. Conversely, GPT-4o exhibits higher accuracy in capturing these elements and distinguishing between various statistical metrics, though it occasionally fails to separate specific metrics clearly.

Table 3. Summary of qualitative evaluation in data extraction between GPT-4 and GPT-4o models.

Aspect		**GPT-4**	**GPT-4o**
Study Details	**Study design**	- Misses study design information (e.g., "randomized")	- Rarely makes such mistakes

	Demographic Information	- Misses demographic information (ratios of sex, prior treatments)	- More consistent accuracy in study design and demographic details
	Intervention	- Misrepresents/omits "Intervention" or its details	
Study Outcomes	**Study Cohort**	- Fails to present, misses, or incorrectly specifies study cohorts and subgroups	- Occasionally presents incorrect cohort names or confuses groups
		- Issues are frequent in comparison studies involving multiple trials	- Such instances are rare
	Data Element	- Misses important elements like "Study duration" and "Data cut-off"	- Higher accuracy in capturing study timelines
	Description	- Incorrectly presents comparator/reference information	- Improved accuracy in presenting descriptions
	Value	- Struggles to separate HR, odds ratios, 95% CI, ranges, and p-values	- Better at distinguishing and presenting values
		- Often marks values as "NA" or provides incorrect values	- Occasionally fails to separate specific metrics clearly
		- Does not reach necessary level of detail	- Accurately indicates when there is no significant difference

Error Analysis in Full-Text Data Extraction

We manually reviewed the extracted data from 5 full-text articles [30–34], particularly focusing on elements extracted from tables. These articles included a total of 16 tables, with 3 to 4 tables per article. Out of 16 tables, the LLM model achieved nearly complete data extraction from 10 tables, with only a few missing points. As shown in Supplementary Figure 2A-B, when a table had child components under one parent component, child components were sometimes captured on its own without the parent component mentioned. For example, “could not be evaluated” instead of “PD-L1 tumor proportion score-could not be evaluated” unlike other child components such as “PD-L1 tumor proportion score-1-49%”. Additionally, we noticed that some errors propagated from the “pdf-to-text" conversion step. For example, when the original PDF table contained the “greater than or equal to” symbol (“>=”), the converted text file only included the “greater than” symbol (“>”) without “equal to”, resulting in the final extracted outcome also displaying the “greater than” symbol (“>”) without “equal to” (Supplementary Figure 2A-B). Furthermore, as shown in Supplementary Figure 2C, in some cases where the table legend is included inside a table, it is recognized as the table column header. In 6 out of 16 tables which were long and contained dense information, each table content was further split, and the data elements were extracted in multiple output files under the same table column headers to tackle the model’s output token limitations. However, we noticed a few errors in correctly assigning the column headers against the extracted elements in these long tables.

.

Discussion

We present a comprehensive, easily customizable, end-to-end solution for an automated SLR system supported by a user interface that begins with an initial paper selection using a simple query via integrated literature databases such as PubMed and Embase, followed by abstract/full-text screening based on user-defined PICO I/E criteria and data elements, and subsequent extraction of these elements. Our system integrates LLMs into each separate module to increase the system's flexibility, particularly in the screening and data extraction steps. The LLM prompts in AID-SLR are designed to automatically capture important information (e.g., study criteria, interpretations of domain-specific terms, etc.) from the users critical to performing the SLR tasks, enhancing the customizability of the system. AID-SLR achieves high performance in both screening and extraction, emphasizing its utility in automating complex literature review processes. Furthermore, we compared the most recent versions of LLMs, GPT-4 and GPT-4o, to investigate their application in data extraction for SLRs, particularly given the complexity and volume of clinical data involved in studies on NSCLC and immunotherapy.

The literature screening process in our system is divided into two distinct phases: abstract screening and full-text screening. These phases address the inherent differences between abstracts and full texts in terms of scope and detail. Abstracts provide a concise overview, summarizing the research question, methods, key findings, and conclusions, while full texts offer a detailed understanding of the research methodology, findings, and their implications. Our system allows for different versions of I/E criteria for these two levels, enabling broader criteria

for abstract screening to capture a wide array of relevant studies and more specific criteria for full-text screening to ensure precision.

The evaluation of our system using the NSCLC use case revealed high levels of precision and recall in both screening phases, reflecting efficient and accurate identification of relevant studies. The system's ability to handle the extraction of detailed clinical study details and outcomes further enhances its value. It demonstrated granularity and depth in data extraction, successfully identifying a wide range of treatment-related outcomes and detailed statistical values. The extraction of hazard ratios (HRs), odds ratios (ORs) with confidence intervals (CIs), and P-values, along with intervention and comparator group information, underscores the system's comprehensiveness and accuracy.

Our comparison between GPT-4 and GPT-4o models highlighted the incremental improvements offered by GPT-4o. While GPT-4 already performs at a high level, GPT-4o consistently outperforms it in data extraction across all data elements. Higher precision scores in the GPT-4o model indicate its superior ability to correctly identify true positive instances, minimizing false positives. This enhanced precision is critical for ensuring the reliability of data extraction in SLRs, particularly when dealing with extensive and detailed clinical datasets. Higher recall in the GPT-4o model suggests it is more effective in capturing all relevant data elements, reducing the chances of missing critical information. Improved recall for specific features within "Study Outcomes," such as "Study Cohort" and "Value," highlights GPT-4o's enhanced capability in extracting complex cohort-related data, addressing shortcomings observed in GPT-4.

Moreover, our system's ability to extract data elements and associated values from tables, in addition to the main text of the articles, significantly enhances the comprehensiveness of the

SLR process. Tables often contain detailed information that is not fully discussed in the text due to word limits. By capturing this information, our system ensures that no critical data is overlooked, providing a more robust and complete synthesis of evidence

Limitations

We acknowledged some limitations in our study. Our whole system was tested using one use case, although a similar LLM-based data extraction module focused on abstracts was previously tested in other indications and contexts [28]. Evaluating the system's performance across different disease areas and various types of clinical trials or real-world studies is crucial for establishing its generalizability and robustness. Additionally, our evaluation for screening and data extraction was conducted with a relatively small sample size. Increasing the evaluation set will be the next immediate step. We also intend to extend our full-text data extraction module to consider extracting elements from article figures and further improve the extraction performance from long tables. Moreover, the I/E criteria are updated manually by users after reviewing the disagreements and the AI screening statistics. In the future, we plan to leverage the feedback provided by users while disagreeing with AI recommendations to automate the process of updating the I/E criteria, thus relieving the burden on the users to modify them manually.

In addition, our current UI can be improved in several ways. Firstly, the current UI does not provide a summary of all extracted data elements from selected relevant articles, even though these data elements and their corresponding values are extracted in the backend. Adding summary analytics of all extracted data elements from selected articles based on the user's PICO criteria could provide more valuable information. Future work could integrate these summary analytics modules, allowing users to choose which data elements to include in their own SLR

studies. Secondly, while our system allows human-AI-interaction at every step, the current UI is designed for one human reviewer and does not provide inter-rater agreement checking between multiple reviewers. Given that the involvement of more than two reviewers is essential in SLRs, further development to integrate this functionality would be required.

Conclusion

By establishing the human-in-the-loop, automated end-to-end AI solution for SLRs with high precision and recall (sensitivity) in both screening and data extraction, our system can reduce the time, cost, and human errors associated with traditional SLRs, ultimately contributing to more timely and comprehensive evidence generation. Additionally, our user-friendly UI allows users to conduct the entire SLR process seamlessly. Integrating our system with other AI tools for comprehensive data analysis could further enhance its utility in SLRs.

Reference

1. Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *J Family Med Prim Care*. 2013;2(1):9-14. doi:10.4103/2249-4863.109934

2. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med*. 2009;6(7):e1000100. doi:10.1371/journal.pmed.1000100

3. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71

4. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. *Lancet*. 2003;362(9391):1225-1230. doi:10.1016/S0140-6736(03)14546-1

5. Raphael MJ, Gyawali B, Booth CM. Real-world evidence and regulatory drug approval. *Nat Rev Clin Oncol*. 2020;17(5):271-272. doi:10.1038/s41571-020-0345-7

6. Ciani O, Jommi C. The role of health technology assessment bodies in shaping drug development. *Drug Des Devel Ther*. 2014;8:2273-2281. doi:10.2147/DDDT.S49935

7. Pournaghi Azar F, Ghojazadeh M. Embracing the future: The evolution of systematic reviews and meta-analyses in periodontology. *J Adv Periodontol Implant Dent*. 2023;15(2):65-66. doi:10.34172/japid.2023.025

8. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev*. 2014;3:74. doi:10.1186/2046-4053-3-74

9. 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. doi:10.1136/bmjopen-2016-012545

10. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials Commun*. 2019;16:100443. doi:10.1016/j.conctc.2019.100443

11. Gresham G, Meinert JL, Gresham AG, Piantadosi S, Meinert CL. Update on the clinical trial landscape: analysis of ClinicalTrials.gov registration data, 2000-2020. *Trials*. 2022;23(1):858. doi:10.1186/s13063-022-06569-2

12. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLoS Med*. 2010;7(9):e1000326. doi:10.1371/journal.pmed.1000326

13. Beller E, Clark J, Tsafnat G, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Syst Rev*. 2018;7(1):77. doi:10.1186/s13643-018-0740-7

14. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Wolfe MS. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). *Syst Rev*. 2018;7(1):3. doi:10.1186/s13643-017-0667-4

15. O'Connor AM, Tsafnat G, Gilbert SB, et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Syst Rev*. 2019;8(1):57. doi:10.1186/s13643-019-0975-y

16. O'Connor AM, Glasziou P, Taylor M, Thomas J, Spijker R, Wolfe MS. A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Syst Rev*. 2020;9(1):100. doi:10.1186/s13643-020-01351-4

17. Schopow N, Osterhoff G, Baur D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. *JMIR Med Inform*. 2023;11:e48933. doi:10.2196/48933

18. 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. doi:10.1002/jrsm.1553

19. Zhang Y, Liang S, Feng Y, et al. Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. *Syst Rev*. 2022;11(1):11. doi:10.1186/s13643-021-01881-5

20. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;12(1):72. doi:10.1186/s13643-023-02243-z

21. 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*. Published online March 14, 2024. doi:10.1002/jrsm.1715

22. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res*. 2024;26:e48996. doi:10.2196/48996

23. Robinson A, Thorne W, Wu BP, et al. Bio-SIEVE: Exploring Instruction Tuning Large Language Models for Systematic Review Automation. Published online 2023. doi:10.48550/ARXIV.2308.06610

24. Nashwan AJ, Jaradat JH. Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation. *Cureus*. 2023;15(8):e43023. doi:10.7759/cureus.43023

25. Ghosh M, Mukherjee S, Ganguly A, Basuchowdhuri P, Naskar SK, Ganguly D. AlpaPICO: Extraction of PICO Frames from Clinical Trial Documents Using LLMs. *Methods*. Published online April 19, 2024:S1046-2023(24)00089-6. doi:10.1016/j.ymeth.2024.04.005

26. Wang Z, Cao L, Danek B, et al. Accelerating Clinical Evidence Synthesis with Large Language Models. Published online 2024. doi:10.48550/ARXIV.2406.17755

27. 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*. Published online March 3, 2024. doi:10.1002/jrsm.1710

28. Lee K, Paek H, Huang LC, et al. SEETrials: Leveraging Large Language Models for Safety and Efficacy Extraction in Oncology Clinical Trials. *medRxiv*. Published online May 13, 2024:2024.01.18.24301502. doi:10.1101/2024.01.18.24301502

29. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *Pharmacoecon Open*. 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9

30. Yamamoto N, Satouchi M, Doi T, et al. KEYNOTE-434 part B: A phase 1 study evaluating the combination of epacadostat, pembrolizumab, and chemotherapy in Japanese patients with previously untreated advanced non-small-cell lung cancer. *Invest New Drugs*. 2024;42(3):261-271. doi:10.1007/s10637-024-01422-6

31. Maggie Liu SY, Huang J, Deng JY, et al. PD-L1 expression guidance on sintilimab versus pembrolizumab with or without platinum-doublet chemotherapy in untreated patients with advanced non-small cell lung cancer (CTONG1901): A phase 2, randomized, controlled trial. *Sci Bull (Beijing)*. 2024;69(4):535-543. doi:10.1016/j.scib.2023.12.046

32. Paz-Ares L, Vicente D, Tafreshi A, et al. A Randomized, Placebo-Controlled Trial of Pembrolizumab Plus Chemotherapy in Patients With Metastatic Squamous NSCLC: Protocol-Specified Final Analysis of KEYNOTE-407. *J Thorac Oncol*. 2020;15(10):1657-1669. doi:10.1016/j.jtho.2020.06.015

33. Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med*. 2018;378(22):2078-2092. doi:10.1056/NEJMoa1801005

34. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*. 2016;375(19):1823-1833. doi:10.1056/NEJMoa1606774

Supplementary Materials

Supplementary Methods.

UI operation protocol

Query

Users can choose to search in PubMed, Embase or both and can enter search terms, combining them using “AND” or “OR” logic. PubMed or Embase’s query expansion techniques will be applied when the user enters terms. For example, entering "Kahler disease" (a synonym for multiple myeloma) will automatically search for multiple myeloma other mapped terms, providing all articles that include any of the synonyms or mapped terms.

I/E criteria (Protocol)

Our I/E criteria follow the PICO (Population, Intervention, Comparator, Outcome) framework. Supplementary Figure 1A shows that users can add their inclusion and exclusion criteria under population, interventions/comparators, and outcomes. Users can then select “Both,” “Abstract,” or “Full-text” for each criterion. Abstracts often include broader terms rather than specific information. For example, a study might mention “immunotherapy” or “chemotherapy” in the abstract, rather than specifying pembrolizumab or platinum-based chemotherapy. To be more inclusive in the abstract screening step, users can provide different sets of PICO criteria that are broader for abstract screening and stricter for full-text screening. If a single criterion applies to both, users can select “Both.” Users can also assign publication type and study design by clicking the relevant options. Additional criteria can be added in the "Other I/E Criteria" section to specify any criterion that does not fall under the PICO categories.

Adding domain-specific knowledge in ambiguous or complex cases can improve model performance. For instance, if a user is screening for "Head and Neck Cancer," providing definitions and lists of head and neck anatomical sites can improve screening efficiency, as many studies might mention specific cancer types like "nasal cavity and sinus cancer" or "throat cancer" instead of "Head and Neck Cancer". Once this study protocol step is complete, the user can click the "Run AI Recommendation" button to start the abstract screening process.

Abstract screening

The abstract screening page displays 2 tables: "Screening counts" and "Confusion matrix" (Supplementary Figure 1B). The "Screening counts" table shows "AI recommended results for relevant and irrelevant articles. Initially, the Human screened column shows "0" for both relevant and irrelevant categories. As human reviewers label articles as relevant or irrelevant, the numbers in human screened column increase. If human reviewers disagree with AI recommendations, reviewers can click "relevant" or "irrelevant" for the individual article (See Supplementary Figure 1C, left panel). Supplementary Figures 1C and D show an example of AI-recommended "irrelevant" and "relevant" articles, respectively, including the AI-recommended exclusion details based on PICO criteria and the AI-generated explanation in plain language. The "Confusion matrix" table shows the agreement between AI recommendations and Human screening.

Full-text screening

Only articles labeled as "relevant" will proceed to the full-text screening step. Users can upload private PDF files on the full-text screening page if not available in public databases.

Data extraction

Users can specify their desired data elements and add them using the interface. The selected data elements will appear as shown in Supplementary Figure 1E.

Data Summary

The articles that are reviewed and marked as "approved citation" by users will appear on the Data Summary page. The summary table includes columns for PMID, citation title, Primary author's last name, Primary author's first name, and Year of publication. Users can download filtered or all data (Supplementary Figure 1F).

LLM prompts for screening and data extraction

We leverage the OpenAI GPT-4 model to develop the screening and data extraction modules. Besides specifying the criteria and data elements, users can optionally add domain-specific information such as disease knowledge and term interpretations which are incorporated into the LLM prompts to guide the model in screening and extraction tasks.

Abstract screening

We construct a generalizable prompt that produces a final screening decision ("Relevant" or "Irrelevant") based on an article's abstract and the given I/E criteria. The prompt contains four components – 1) main instruction to perform the screening task, 2) necessary context to take the decision (title, abstract, I/E criteria, publication type, study design, and domain knowledge), 3) general instructions describing the screening logic, and 4) output schema for response generation

which includes the screening decision, a supporting explanation, and the specific exclusion reasons for each PICO category.

Publication type and study design classification

We develop a prompt to classify an article's publication type and study design. The prompt takes in the article's title and abstract, followed by the step-by-step instructions for classification. The instructions contain two parts – 1) Publication type classification – Identify the type(s) given a list of publication type categories and their definitions, and 2) Study design classification – Identify the design(s) provided the guidelines defining the association between publication types and study designs (e.g., If the publication type is "Original research article", then the study design can be either "Clinical trial", "Real word evidence", or "Other/unspecified.")

Full-text screening

The prompt structure is like abstract screening except that we provide the full text of an article as context and the instructions to take the screening decision are stricter compared to abstract screening.

Data element extraction from abstract

We design a generalizable data extraction module that consists of three prompts to identify study details, study cohorts, and study outcomes.

Study Details

This prompt extracts key study characteristics, such as cohort size. The title, abstract, and any additional domain knowledge are provided as context. The output schema for each extracted study detail element includes four features: element name, description, value, and text span

Study Cohorts

This prompt aims to identify all the study cohorts or groups mentioned in the article's abstract. This starts with a query, "*Extract all detailed information of names or descriptions for all cohorts, sub-cohorts, sub-groups, and study arms mentioned in the following abstract.*", followed by the title and abstract as context, and finally instructs to list the names of all cohorts, sub-cohorts, sub-groups, and study arms separated by commas.

Study Outcomes

Similar to the study details prompt, this prompt instructs the model to identify outcome-related data elements using the abstract as context. The LLM output schema for outcomes includes five features: study cohort, element name, description, text span, and value. Additionally, this prompt instructs that every outcome value should be extracted with its associated cohort, selected from the ones returned by the "Study cohorts" prompt. To guide this association, we include a statement in the prompt, "Cohorts identified in the article include: -- <study cohorts returned by the above prompt>," to inform the model about the study cohorts. If an outcome applies to the entire population, the study cohort should be assigned "Entire cohort."

Data element extraction from full-text

The prompts are crafted similarly to abstract data extraction except that we provide the full text of an article as context to the model. Here, we extract the study outcome elements from both the tables and the main text by processing the tables and the article text separately. Further, for tables, each table is handled separately, that is, we provide each table content as context to the LLM to identify study outcomes from that specific table. In case the articles are available in pdf, we first convert their content into textual format using Amazon AWS Textract.