MSR28

Evaluating the Performance of GPT-4o and Retrieval-Augmented Generation in Extracting Data from Journal Articles: A Comparative Study

Wei-Hua Huang, Varadraj Poojary, Ellen Kasireddy, Mir Sohail Fazeli

Evidinno Outcomes Research Inc., Vancouver, British Columbia, Canada

Background

- Systematic reviews are essential for evidence-based research as synthesized published data can inform clinical practice and policy
- However, manual data extraction from scientific articles is time-consuming, labor-intensive, and prone to errors, potentially affecting review quality[1,2]
- Advancements in natural language processing (NLP) and artificial intelligence (AI), particularly large language models, offer a solution[3]

Objective

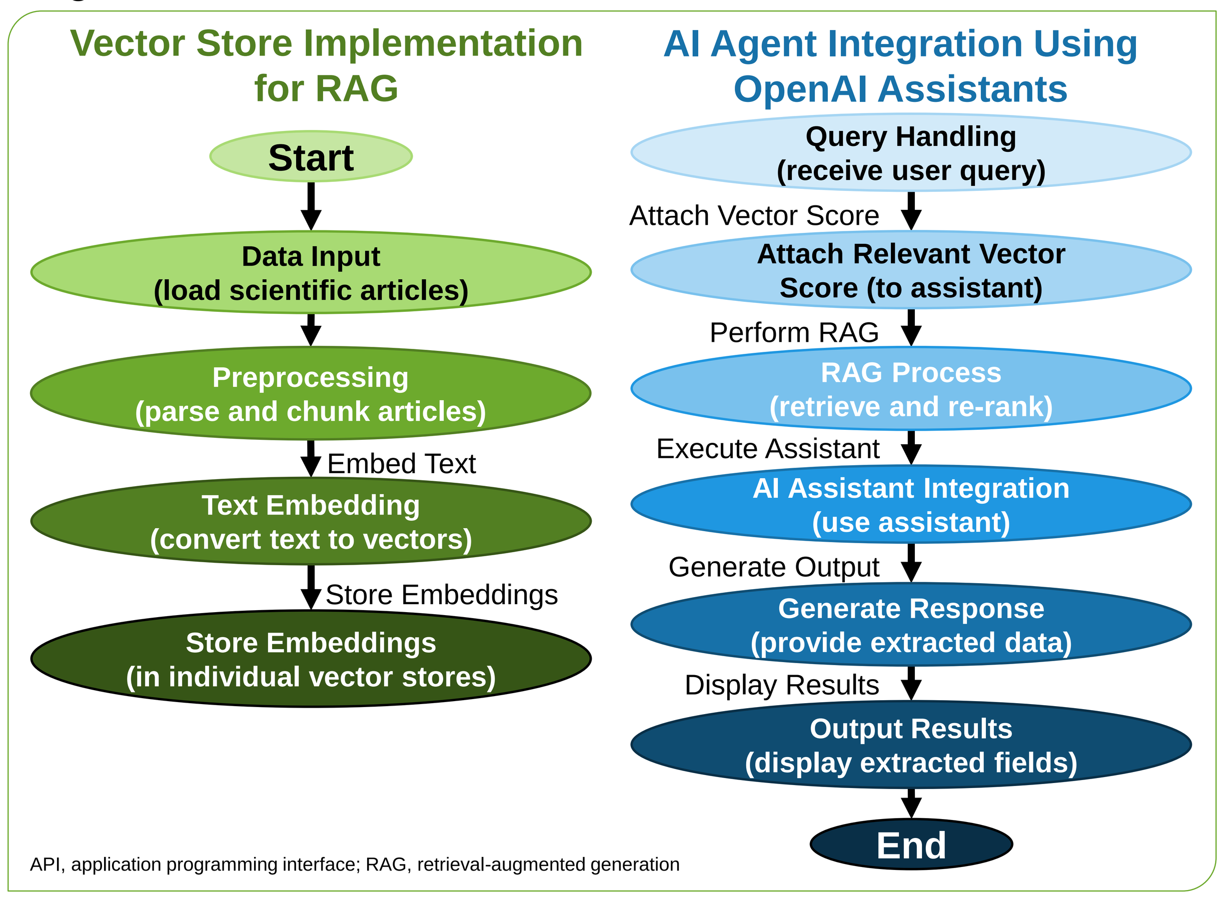
- To evaluate the performance of a custom-designed system utilizing GPT-4o and retrieval-augmented generation (RAG) for extracting specific fields from scientific journal articles, compared with domain expert extraction

Methods

SYSTEM DEVELOPMENT

- A system using OpenAI's GPT-4o model[4] integrated with RAG capabilities was developed to automate the extraction of key data fields for both straightforward and nuanced data with accuracy comparable to that of domain experts
- System architecture consisted of 2 primary components (**Figure 1**):
  1. **Vector Store Implementation for RAG**
     - Journal articles were parsed, chunked, and embedded into vector stores
  2. **AI Agent Integration Using OpenAI Assistants**
     - OpenAI assistant was tailored to perform data extraction tasks
     - System leveraged the File Search tool to retrieve and extract relevant data from Vector Stores, enabling multi-step, context-aware searches

Figure 1: Workflow for Automated Data Extraction



API, application programming interface; RAG, retrieval-augmented generation

EVALUATION PROCESS

Study Selection

- To evaluate system performance, 4 unpublished systematic reviews including 36 published clinical trials and observational studies across diverse medical fields were selected
  - **Systematic review 1:** 10 full-text studies on prognostic value of sentinel lymph node biopsy in melanoma (9 cohort studies, 1 cross-sectional study)
  - **Systematic review 2:** 10 full-text studies on humanistic burden of systemic lupus (8 cross-sectional studies, 1 cohort study, 1 case-control study)
  - **Systematic review 3:** 8 full-text studies on indicators of symptomatic progression in oncology (7 randomized controlled trials [RCTs], 1 post-hoc analysis of an RCT)
  - **Systematic review 4:** 8 full-text studies on humanistic burden of kidney transplant rejection (5 cross-sectional studies, 2 cohort studies, 1 RCT)

Data Extraction and Analysis

- System tasked with extracting 6 data fields from full-text articles
  - Study design, location, setting, sample size, trial phase, blinding
  - Fields were chosen to test system's ability to handle extractions that were considered straightforward (information typically explicitly reported in articles; e.g., "Location", "Sample Size"), and those that were complex (varied reporting styles and terminologies in articles; e.g., "Study Design")

Comparative Analysis

- AI-extracted data were compared with those of domain experts by a third reviewer to determine if the AI-extracted data were "consistent" with domain experts. Two elements were considered:
  - **Similarity**: How closely AI's extractions matched those of experts in terms of both content and format
  - **Completeness**: System's ability to accurately capture all relevant data points that domain experts captured
- If both metrics were satisfied, the data field would be considered as "consistent" against the expert's extraction. The overall consistency rate was then calculated by using the following formula:
  - Consistency Rate = ((Number of Correct Extractions by AI) / (Total Number of Extractions of the Same Field by Expert)) × 100
  - Consistency was categorized as high (consistency >90%), moderate (75–90%), or low (<75%)

Results

GPT-4o and RAG-based system achieved an average consistency rate of 84% across diverse data types compared with extractions from domain experts

Figure 2: Overall Distribution of Consistent Extractions

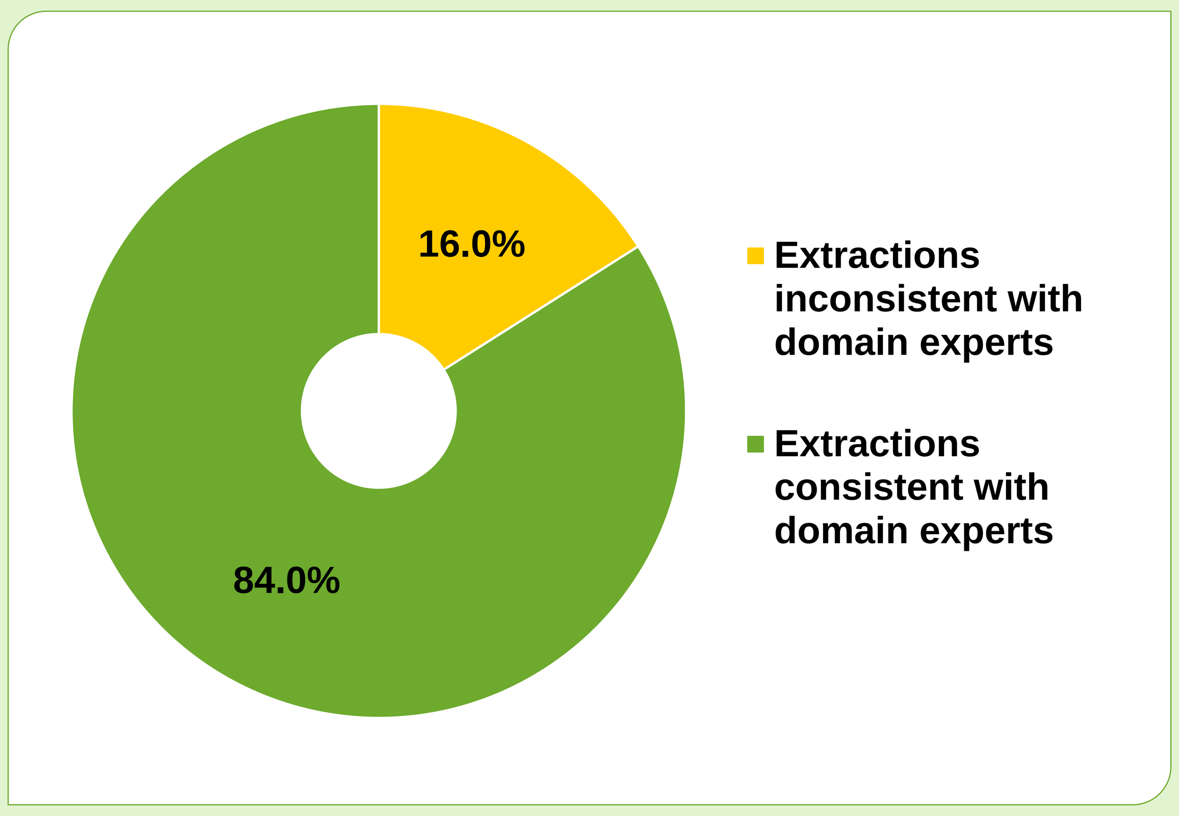
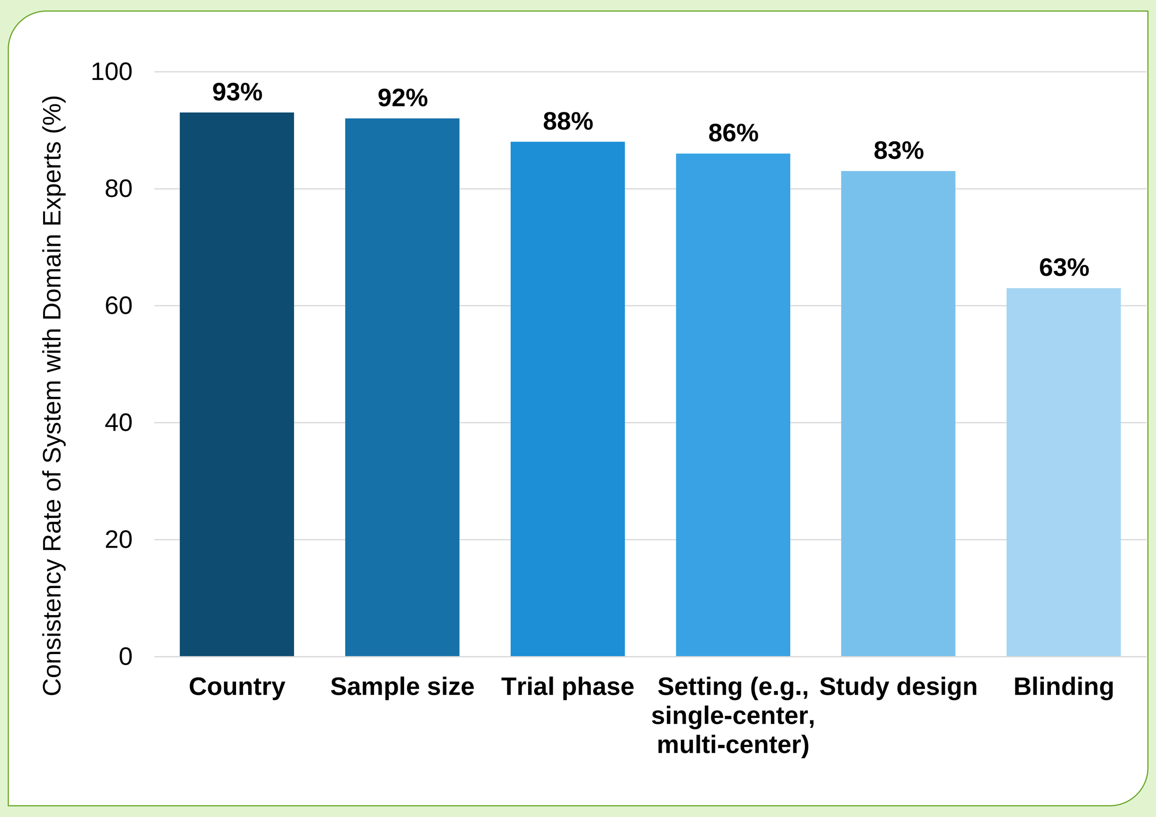


Figure 3: Consistency Rates by Data Field Type



OVERALL PERFORMANCE

Consistency

- System successfully extracted 168 data points from 36 studies, with 141 (84%) extractions considered consistent with those of domain experts (**Figure 2**)
- Consistency rate of the system varied across different data types, reflecting diversity and complexity of information reported in scientific literature (**Figures 3** and **4**)
- Performance of the system was categorized into three levels:

1. **High Consistency**
   - **Study Location:** Extracted 26/28 data points correctly (93% consistency)
     - High accuracy reflects the consistent way in which study location was reported across studies
   - **Sample Size:** Extracted 33/36 data points correctly (92% consistency)
     - This data type is often clearly stated, allowing more precise extraction
2. **Moderate Consistency**
   - **Trial Phase:** Extracted 7/8 data points correctly (88% consistency)
     - Occasional misidentifications occurred when there were subtle differences in the way phases were reported across studies
   - **Setting:** Extracted 31/36 data points correctly (86% consistency)
   - **Study design:** Extracted 30/36 data points correctly (83% consistency)
     - Most difficult field to extract due to the complexity and variability of study design descriptions
3. **Low Consistency**
   - **Blinding:** Extracted 5/8 data points correctly (63% consistency)
     - Inconsistencies in how blinding information was reported across studies led to lower extraction performance

ERROR ANALYSIS

Contextual Misinterpretations

- Errors typically occurred due to the system misinterpreting context, especially for complex fields (e.g., Study Design)
  - E.g., in studies that included multiple designs or exploratory sub-studies, the system sometimes incorrectly identified the primary design

Incomplete Extractions

- Some fields were partially extracted correctly, but had missing data
  - E.g., For "Location", the system sometimes only extracted one country when the study was conducted across multiple counties.

Discussion

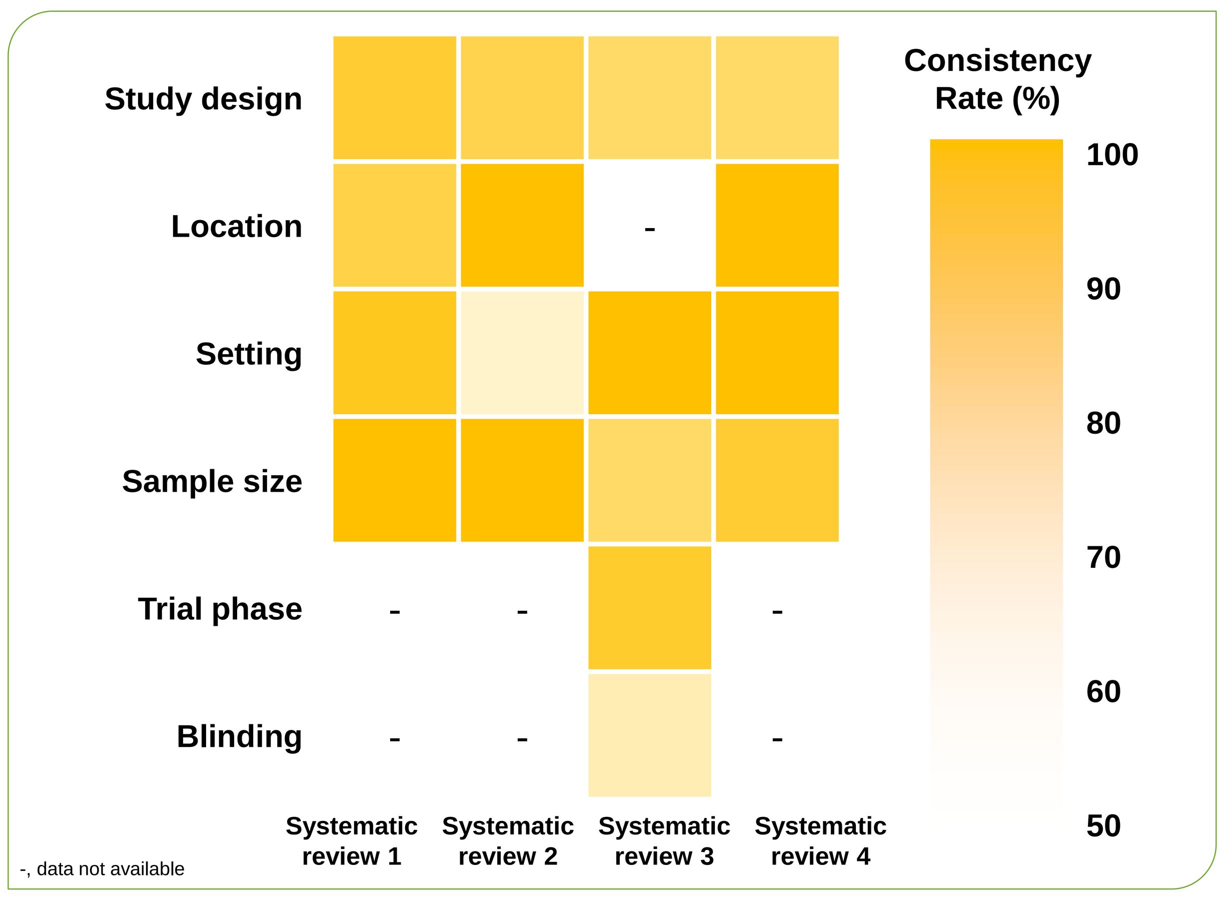
STRENGTHS

- System excelled at extracting simple, well-defined fields (e.g., study location, sample size) with consistency rates over 90%
  - This demonstrates the system's robustness when handling standardized data that is uniformly reported across scientific studies, suggesting strong potential for use in structured data environments

AREAS FOR IMPROVEMENT

- Contextual understanding
  - Fields such as "Blinding" and "Study Design" require the system to better understand and interpret complex, nuanced information
  - Enhancing model's contextual recognition could significantly improve accuracy in these more challenging fields
- Handling synonym variations
  - Performance could also be improved by refining the system's ability to handle varied phrasing and synonyms
    - Particularly in fields such as "Trial Phase", in which minor wording differences impact extraction
- Advanced NLP techniques
  - Incorporating more sophisticated NLP models for semantic understanding could help the system navigate the complexity of unstructured data
    - E.g., variable formats of study design reporting

Figure 4: Heatmap of Consistency Rates Across Different Systematic Reviews and Fields



Conclusions

- GPT-4o and RAG-based system shows high level of accuracy for certain measures of data extraction from published articles, although variability in performance across different fields indicates the need for further refinement
- Future development will focus on enhancing contextual understanding for complex fields, improving synonym recognition/semantic analysis, expanding/fine-tuning the system using broader datasets, and improving data extraction accuracy for additional fields, such as efficacy and safety measures
  - These improvements aim to create a more robust and comprehensive tool for data extraction and evidence synthesis

References

1. Kim JSM et al. *Syst Rev*. 2022;11:206.
2. Borah R et al. *BMJ Open*. 2017;7:e012545.
3. Ofori-Boateng R et al. *Artif Intell Rev*. 2024;57:200.
4. OpenAI GPT-4. 2023. https://openai.com/index/gpt-4-research/

Acknowledgements

Authors report employment with Evidinno Outcomes Research Inc. (Vancouver, BC, Canada)
Authors would like to thank Kimberly Hofer of Evidinno Outcomes Research Inc. for her assistance in poster development

EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

Presented at ISPOR Europe 2024; November 17-20; Barcelona, Spain

Email: mfazeli@evidinno.com