Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Configurable LLM System for Automated Data Extraction in Meta-Analysis Using a Minimal Atomic Unit Framework

Yijia Yin
Xi'an Jiaotong-Liverpool University
Haoxin Feng
Xi'an Jiaotong-Liverpool University
Mengqi Shao
The Second People's Hospital of Yangzhou
Yujia Pan
Yunnan Normal University
Chuyu Zhao
Xi'an Jiaotong-Liverpool University
Chenxin Zhu
Xi'an Jiaotong-Liverpool University
You Wan
Xi'an Jiaotong-Liverpool University
Tiejun Tong
Hong Kong Baptist University
Xiaoyu Tang
xiaoyu.tang@xjtlu.edu.cn

Xi'an Jiaotong-Liverpool University

Research Article

Keywords: data extraction, evidence synthesis, large language models, meta-analysis, minimal atomic unit, natural language processing, systematic review

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10662331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Configurable LLM System for Automated Data Extraction in Meta-Analysis Using a Minimal Atomic Unit Framework

Yijia Yin[1], Haoxin Feng[1], Mengqi Shao[2], Yujia Pan[3], Chuyu Zhao[1], Chenxin Zhu[4], You Wan[4], Tiejun Tong[5], Xiaoyu Tang[4,6*]

[1]School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China.
[2]Department of Rehabilitation Medicine, The Second People's Hospital of Yangzhou, Yangzhou, China.
[3]School of Information Science, Yunnan Normal University, Kunming, China.
[4]Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, 215123, Jiangsu, China.
[5]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
[6]Institute of Population Health, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, United Kingdom.

*Corresponding author(s). E-mail(s): xiaoyu.tang@xjtlu.edu.cn;

Abstract

Background: Extracting numerical data from randomized controlled trial reports is a major bottleneck in systematic reviews. Large language models (LLMs) can assist with this task, but complex combinations of study arms, time points, subgroups, and outcomes create risks of omission, misalignment, and unsupported values. We developed a configurable LLM extraction system based on a Minimal Atomic Unit (MAU) representation.

Methods: The system uses GPT-4.1 and a three-level pipeline that produces study-level, arm-level, and outcome-level records, with source text attached to every extracted field. During prompt development, an exploratory comparison used the same three continuous positive airway pressure trial reports for prompt refinement and for comparing MAU and non-MAU prompts. The two settings used the same model, extraction scope, output fields, and source-tracing requirements; only the MAU setting explicitly decomposed the task. The finalized system was independently evaluated on 88 full-text study reports from five clinical domains. Performance was assessed at the aligned-cell level using precision, recall, F1 score, and hallucination rate.
Results: In the exploratory prompt-development comparison (1,160 cells), precision, recall, and F1 were 1.0000, 0.9595, and 0.9793 with MAU and 0.9288, 0.2138, and 0.3476 without MAU; neither setting produced a hallucinated response. In the independent validation dataset (25,887 evaluated cells), precision was 0.997 (95% confidence interval [CI], 0.996–0.997), recall was 0.984 (95% CI, 0.983–0.985), F1 was 0.990 (95% CI, 0.989–0.991), and the hallucination rate was 0.2% (95% CI, 0.2%–0.3%).
Conclusions: The MAU framework produced structured, source-traceable extraction with high cell-level performance across five domains. The MAU versus non-MAU finding is exploratory because the comparison reused the prompt-development studies. External validation with other models and review teams remains necessary before unsupervised use.

Keywords: data extraction, evidence synthesis, large language models, meta-analysis, minimal atomic unit, natural language processing, systematic review

1 Background

Systematic reviews and meta-analyses are fundamental to evidence-based medicine because they synthesize results from independent studies and support clinical and policy decisions [1]. As the biomedical literature expands, evidence synthesis must process an increasing number of reports. The Cochrane Handbook requires reviewers to extract study design, participant characteristics, intervention details, outcome measures, and statistical parameters from each included study [2]. This work is labor intensive and prone to error [3, 4]. Duplicate extraction improves reliability but increases reviewer time [5], making data extraction a persistent bottleneck.

Natural language processing has created opportunities to automate parts of this workflow. Earlier approaches used rules or conventional machine-learning models [6–9], but these methods often require task-specific annotation and transfer poorly across domains [10–15]. Instruction-tuned large language models (LLMs) can instead perform complex information-extraction tasks through prompts without task-specific model training [16–20].

Direct application of LLMs to meta-analysis data extraction nevertheless remains difficult. Much existing work focuses on Population, Intervention, Comparator, and Outcome (PICO) elements, screening, or narrative summaries, whereas meta-analysis requires structured qualitative and numerical fields that can be aligned to statistical synthesis [21–24]. Trial reports may include multiple arms, time points, subgroups, outcomes, and analysis populations, increasing the risk that an extracted number is assigned to the wrong analytical context [25–28]. A recent BMC Medical Research Methodology study found moderate-to-high performance for some binary fields but substantially poorer extraction of continuous outcomes, reinforcing the need for more structured methods [29].

Hallucination is a further concern. When evidence is absent or ambiguous, an LLM may generate a plausible but unsupported value; if such a value enters a pooled analysis, it can introduce difficult-to-detect bias [30–32]. Source-linked outputs are therefore important for verification and auditability [33–35].

We developed a configurable extraction system based on GPT-4.1 and a Minimal Atomic Unit (MAU) data representation. Each outcome record is bound to a study, study arm, assessment time point, subgroup, and outcome measure. The system produces study-level, arm-level, and outcome-level tables and attaches source text to every field. This study aimed to develop the MAU workflow, explore whether explicit MAU decomposition was associated with better extraction during prompt development, and independently evaluate the finalized system across five clinical domains.

2 Methods

2.1 System design and extraction workflow

The workflow transforms full-text reports into standardized, source-traceable data for meta-analysis (Fig. 1). Development and experiments were conducted from December 2025 through March 2026. The pipeline comprises document preprocessing, configurable task definition, three-level LLM extraction, post-processing, and comparison with a manually adjudicated gold standard.

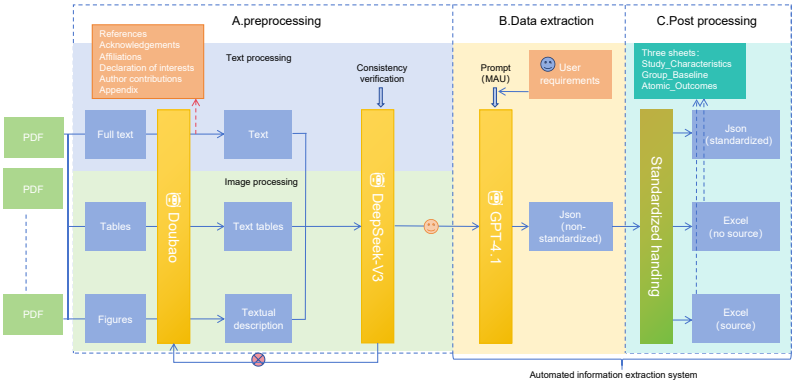


Fig. 1 Workflow of the MAU-based extraction system

The MAU is the smallest independently interpretable record required for a meta-analysis. It is defined by five dimensions: study identifier, study arm, assessment time point, subgroup, and outcome measure. Multi-arm trials are represented as separate arm-level records, and every outcome record remains linked to its arm, comparison type, and comparison target. This decomposition converts a broad document-level task into a series of context-specific extraction tasks. Every extracted field includes supporting source text to enable verification.

Source PDFs were converted to plain text using Doubao. References, acknowledgements, appendices, ORCID identifiers, author-contribution statements, affiliations, and declarations were removed because they were not inputs to outcome extraction. Tables were converted into structured text, and relevant graphical content was represented in textual descriptions. DeepSeek-V3 then checked whether the converted text was consistent with the source report. When it identified missing or inconsistent content, the relevant item and reason were returned for another conversion pass. Doubao and DeepSeek-V3 were preprocessing tools; neither identified MAUs nor produced the final structured extraction.

Before extraction, users can specify intervention and comparator definitions, outcomes and outcome types, primary time points, comparison types, and eligible study designs. Unspecified parameters remain unrestricted. The configured inputs first determine eligible study arms, time points, subgroups, and study designs; subsequent levels link arm-level evidence and extract outcome statistics within the same scope.

Final extraction used GPT-4.1 through the OpenAI application programming interface. No model fine-tuning or parameter updating was performed. Temperature was set to 0 to reduce sampling variability, with a 120-second per-request timeout and up to three retries for transient failures. Level 1 (`Study_Characteristics`) contains 17 study-level fields. Level 2 (`Group_Baseline`) contains 17 arm-level fields. Level 3 (`Atomic_Outcomes`) contains 31 fields describing outcomes, timing, comparison structure, descriptive and inferential statistics, denominators, uncertainty, and clinical direction. Representative outputs are shown in Additional file 1, Supplementary Tables S2–S4.

Model-generated JavaScript Object Notation (JSON) files underwent rule-based post-processing. Time-point expressions and synonymous outcome names were standardized, and derived fields such as assessment phase and follow-up duration in weeks were calculated. Outputs were exported as JSON and as workbooks containing the

three extraction tables and source-text columns. The retained JSON outputs record GPT-4.1 as the extraction model. Temperature was fixed at 0 in the executable code, but this parameter was not written to the legacy JSON metadata.

2.2 Exploratory comparison during prompt development

Three randomized controlled trial (RCT) reports from a meta-analysis of continuous positive airway pressure adherence and cardiovascular-event risk were used to develop the prompt sequence, MAU schema, input constraints, and JSON format. They were selected because they included both table-based statistics and narrative outcome reporting. The same three reports were subsequently used for an exploratory comparison of the finalized MAU and non-MAU prompts; consequently, this comparison was not an independent validation and was not designed to provide a causal estimate of the MAU effect. The prompts were finalized using these three development reports; none of the 88 reports in the independent validation set was used to revise the prompts.

Both settings used GPT-4.1 at temperature 0 and had the same extraction scope, target fields, output format, and requirement to attach supporting source text. The only intended structural difference was that the MAU setting explicitly decomposed extraction by study arm, time point, subgroup, and outcome, whereas the non-MAU setting requested the same fields without MAU task decomposition. Outputs were compared with manually annotated values at the cell level. A concise comparison of the prompt structures is reported in Additional file 1, Supplementary Table S8; de-identified historical prompt templates extracted directly from the surviving non-MAU implementation are reproduced in Additional file 2.

2.3 Independent multi-domain validation

The finalized MAU system was evaluated using studies from published systematic reviews or meta-analyses in breast cancer surgery, schizophrenia treatment, cardiovascular disease prevention, type 2 diabetes management, and postoperative orthopedic rehabilitation. These domains include multi-arm designs, multiple follow-up periods, subgroup analyses, dichotomous and continuous outcomes, time-to-event outcomes, complex scales, and heterogeneous terminology. The only user-configured input in these experiments was the outcome domain; intervention, comparator, outcome type, time point, comparison type, and study-design settings were left unrestricted. Domain-specific outcome inputs are listed in Additional file 1, Supplementary Table S1.

We initially considered 104 study reports involving more than 186,000 participants. Sixteen were excluded before extraction: 12 schizophrenia reports and two diabetes reports because full text was unavailable, and two cardiovascular reports because they did not contain the prespecified core cardiovascular outcomes for the simulated meta-analysis. The final validation set contained 88 full-text study reports: breast cancer surgery (5), schizophrenia treatment (44), cardiovascular disease prevention (19), type 2 diabetes management (11), and postoperative orthopedic rehabilitation (9). The dataset included 51 outcome measures. Exclusions and outcomes are detailed in Additional file 1, Supplementary Table S5.

2.4 Gold standard

The gold standard used the same MAU definitions and fields as the extraction system. HF and MS independently performed the initial extraction. XT and TT independently validated those records. Disagreements were resolved by returning to the source report and reaching consensus. Only information directly identifiable in article text, tables,

or figures was entered. If a field was unreported, the gold-standard cell was left empty and retained so that unsupported model values could be detected.

2.5 Evaluation and statistical analysis

Model output and gold-standard records were aligned by table name, row identifier, and field name. Reviewers manually inspected each aligned cell, classified it as correct, incorrect, hallucinated, or missing, and compiled the aggregate category counts (Table 1). A hallucination was a model value in a cell whose gold standard was empty. An omission was an empty model cell whose gold standard contained a value.

Table 1 Cell categories and aggregate evaluation counts

Count	Correct	Incorrect	Hallucinated	Missing	Definition
Correctly extracted	Yes	–	–	–	Model value matched the gold standard.
Hallucinated	–	–	Yes	–	Model value was unsupported by the gold standard.
Valid model responses	Yes	Yes	Yes	–	All nonempty model responses.
Gold-standard targets	Yes	Yes	–	Yes	All nonempty gold-standard cells.
Evaluated cells	Yes	Yes	Yes	Yes	All cells included in evaluation.

Metrics were calculated as follows:

$$\text{Precision} = \frac{C}{M}, \qquad \text{Recall} = \frac{C}{G}, \tag{1}$$

$$\text{Hallucination rate} = \frac{H}{M}, \qquad F_1 = \frac{2C}{M+G}, \tag{2}$$

where C is the number of correctly extracted cells, H the number of hallucinated cells, M the number of valid model responses, and G the number of gold-standard targets. The reported metrics were calculated from the manually compiled counts.

Precision, recall, and hallucination rate were treated as binomial proportions and assigned Wilson 95% confidence intervals (CIs). F1 CIs were estimated with 10,000 nonparametric bootstrap iterations, and figures were produced in Python. Metrics were first calculated for all validation cells and then by clinical domain. No domain-specific model tuning was performed.

3 Results

3.1 Exploratory MAU versus non-MAU comparison

The prompt-development comparison contained 1,160 evaluated cells. With the non-MAU prompt, 248 cells were correct and 267 contained a valid model response; with the MAU prompt, 1,113 cells were correct and 1,113 contained a valid response (Supplementary Table S6). No hallucinated response was observed in either setting; the Wilson 95% CI was 0–0.0033.

Precision was 0.9288 (95% CI, 0.8915–0.9540) without MAU and 1.0000 (95% CI, 0.9966–1.0000) with MAU. Recall was 0.2138 (95% CI, 0.1912–0.2383) without MAU and 0.9595 (95% CI, 0.9465–0.9694) with MAU. F1 was 0.3476 (95% CI, 0.3159–0.3791) without MAU and 0.9793 (95% CI, 0.9735–0.9851) with MAU (Fig. 2). In this exploratory dataset, the largest difference was in recall. Because these reports had been used to refine the prompts, these estimates describe performance on the development data and should not be interpreted as an unbiased estimate of MAU effectiveness.

3.2 Overall performance in the independent validation dataset

Across 88 full-text study reports, 51 outcomes, and 25,887 evaluated cells, precision was 0.997 (95% CI, 0.996–0.997), recall was 0.984 (95% CI, 0.983–0.985), F1 was 0.990 (95% CI, 0.989–0.991), and the hallucination rate was 0.2% (95% CI, 0.2%–0.3%) (Fig. 3; see Additional file 1, Supplementary Table S7). There were 55 hallucinated responses, 384 omissions, and 33 incorrect extractions.

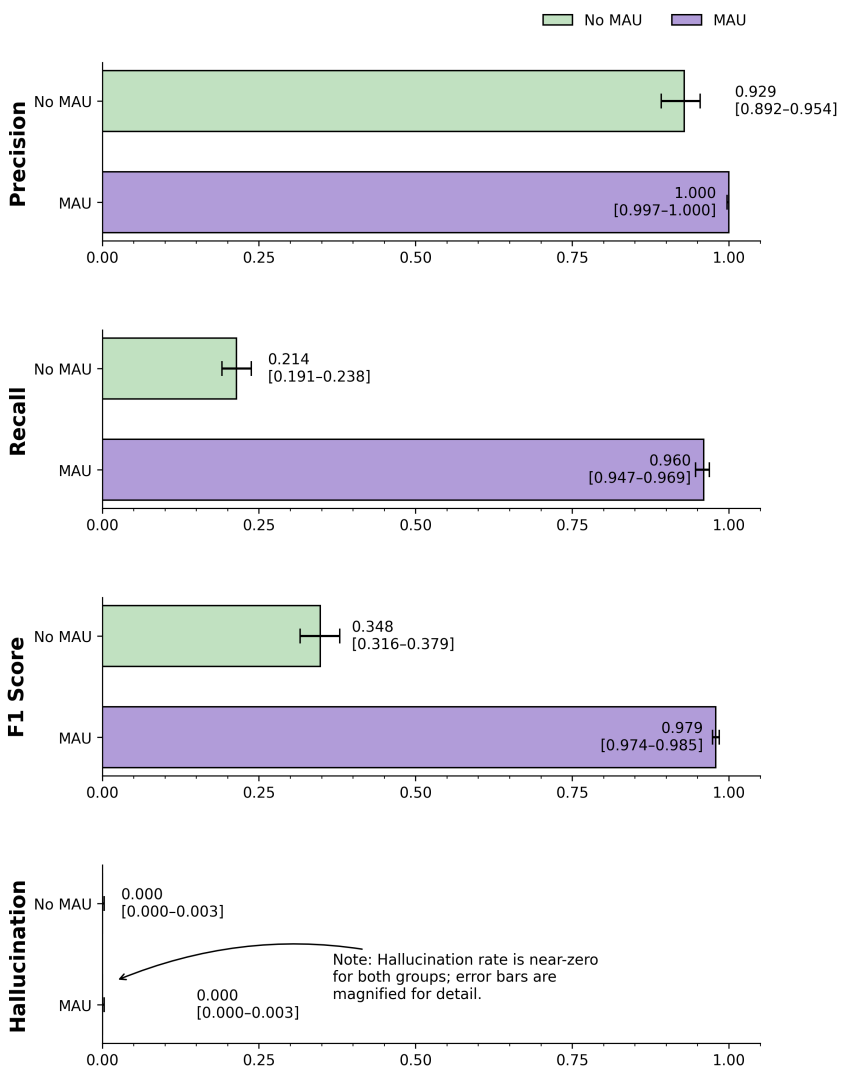


Fig. 2 Exploratory comparison of MAU and non-MAU prompt structures

Of the 55 hallucinated responses, 52 occurred in `Atomic_Outcomes`, especially in `total_count` and `event_count`. These outputs commonly reflected quantities inferred from surrounding context rather than directly reported values. Most omissions occurred when the same information, such as a denominator, P value, intervention description, or baseline descriptor, applied to multiple rows and was entered in only one. Long narrative fields, including multi-item eligibility criteria, were sometimes

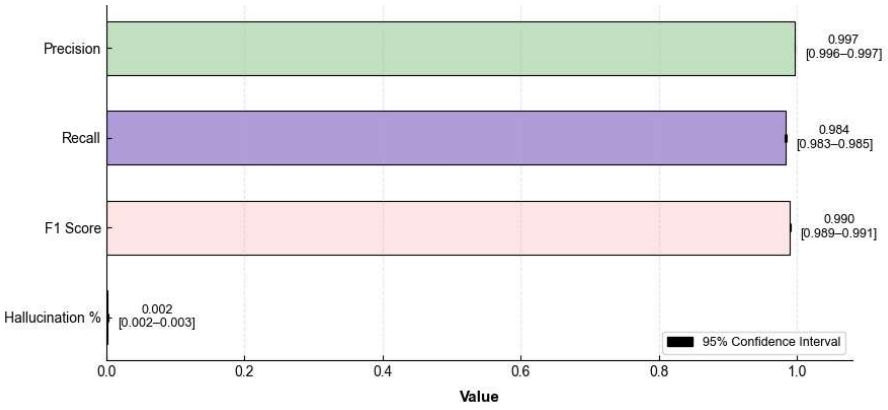


Fig. 3 Overall performance in the multi-domain validation dataset

incomplete. Incorrect extractions most often involved alignment among randomized, analyzed, follow-up, subgroup, and outcome denominators, or distinctions among related follow-up descriptions.

3.3 Performance across clinical domains

Precision ranged from 0.983 (95% CI, 0.976–0.988) to 1.000 (95% CI, 0.999–1.000), recall from 0.921 (95% CI, 0.907–0.932) to 0.999 (95% CI, 0.995–1.000), and F1 from 0.951 to 0.998 (Fig. 4). Hallucination rates were below 0.004 in every domain, with none observed in postoperative orthopedic rehabilitation.

Breast cancer surgery had precision 0.996, recall 0.999, and F1 0.998. Schizophrenia treatment had precision 0.997, recall 0.992, and F1 0.995. Cardiovascular disease prevention had precision 0.997, recall 0.986, and F1 0.991. Type 2 diabetes had the lowest recall (0.921) and F1 (0.951), primarily because repeated denominators and analysis populations were omitted across multi-endpoint rows. Postoperative orthopedic rehabilitation had precision 1.000, recall 0.979, and F1 0.989; remaining omissions commonly involved terminology mapping among expressions for knee range of motion.

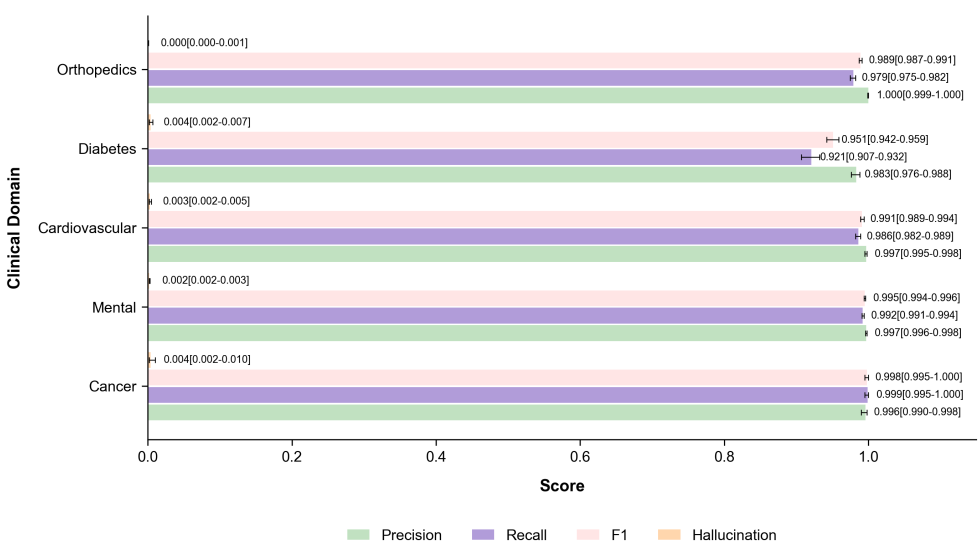


Fig. 4 Performance across five clinical domains

4 Discussion

4.1 Principal findings

The finalized MAU system produced high precision, recall, and F1 with a low hallucination rate across 88 full-text study reports from five domains. The error analysis nevertheless shows that high aggregate performance does not eliminate the need for verification: repeated-field omissions, long narrative fields, and denominator alignment remained important failure modes.

Performance was also higher with MAU than without MAU on the three prompt-development reports. The magnitude of the recall difference suggests that explicit decomposition may help the model preserve arm, time-point, subgroup, and outcome context. However, the same reports were used for prompt refinement, so the comparison is exploratory and susceptible to development-set optimism. It supports the decision to evaluate the finalized architecture but does not isolate the causal contribution of MAU or replace an independently designed ablation study.

Recent work in BMC Medical Research Methodology provides useful context. Yisha and colleagues reported that general-purpose LLMs extracted some binary RCT fields well but performed poorly on continuous means and standard deviations [29]. A schema-constrained pipeline published after completion of our experiments emphasized typed schemas, evidence gating, and sentence-level provenance at corpus scale [36]. The present system shares the goal of auditable structured extraction, while its MAU representation explicitly binds each meta-analysis value to an arm, time point, subgroup, outcome, and comparison structure and was evaluated against an adjudicated cell-level gold standard across five clinical domains.

4.2 Implications for evidence synthesis

The system converts dispersed trial information into standardized records that more closely resemble meta-analysis input tables. Source links can make checking and correction more targeted, and the stable three-level schema may support reuse when reviews are updated. These are workflow capabilities rather than demonstrated savings: the present study did not measure reviewer time, monetary cost, or downstream meta-analysis agreement. Human review remains necessary before extracted values are used in evidence synthesis.

4.3 Limitations

First, the MAU versus non-MAU comparison reused the three prompt-development reports and therefore was not held out. Although a surviving historical non-MAU implementation was located and its de-identified prompt templates are reproduced in Additional file 2, the exact prompt-freeze timestamp was not preserved. Second, although validation covered five domains, the reports were selected through existing reviews and may not represent all publication formats or clinical specialties. Third, all final extraction experiments used GPT-4.1; performance may differ with other models or with later versions of a hosted API. Fourth, the confidence intervals are cell-level

estimates and do not account for correlation among cells within the same report; they may therefore be narrower than study-clustered intervals. Fifth, the study did not measure time, cost, inter-reviewer effort, or the effect of residual errors on pooled estimates. Sixth, preprocessing depended on hosted services, and the exact historical Doubao and DeepSeek-V3 API identifiers were not retained in the repository. Finally, terminology normalization and repeated-field completion require further improvement.

Future evaluation should use a preregistered, independently held-out comparison of MAU and non-MAU prompts, report study-clustered uncertainty, include additional model families and review teams, and measure time, cost, correction burden, and the effect of extraction errors on meta-analysis results.

5 Conclusions

The configurable MAU system produced structured, source-traceable data with high cell-level performance in an independent set of 88 full-text study reports across five clinical domains. The development-set comparison suggests that explicit MAU decomposition may improve completeness, but independent ablation is needed to quantify that contribution. The framework is best considered a verification-supported extraction method rather than a replacement for expert review.

List of abbreviations

API: application programming interface; CI: confidence interval; CPAP: continuous positive airway pressure; JSON: JavaScript Object Notation; LLM: large language model; MAU: Minimal Atomic Unit; NLP: natural language processing; ORCID: Open Researcher and Contributor ID; PDF: portable document format; PICO: Population, Intervention, Comparator, and Outcome; RCT: randomized controlled trial.

Supplementary information. Additional file 1 (.pdf). Title: Supplementary Tables S1–S9. Description: Validation inputs, representative extraction records,

validation-set composition, exploratory prompt comparison, evaluation counts, prompt structures, and input-token lengths.

Additional file 2 (.txt). Title: De-identified historical non-MAU prompt templates. Description: The shared study-level prompt and direct arm/outcome extraction prompt were extracted directly from the surviving historical non-MAU implementation. Runtime values and report text are represented by placeholders; credentials and local file paths are not included.

Declarations

Ethics approval and consent to participate. Not applicable. This methodological study analyzed published scientific reports and did not recruit participants, access identifiable participant-level data, or conduct new research involving humans or human tissue.

Consent for publication. Not applicable.

Availability of data and materials. The code, configuration examples, processed text inputs, derived datasets, gold-standard records, and evaluation outputs are available in the MAU-RCT-Extraction repository at the archived Git commit [37]. Publisher-formatted source PDFs are not redistributed because of copyright restrictions; users must obtain source articles through lawful access.

Code availability. Project name: MAU-RCT-Extraction. Project home page and archived version: see Ref. [37]. Operating system: tested with Windows; the Python code is otherwise platform independent. Programming language: Python 3. Requirements: OpenAI Python package, pandas, openpyxl, requests, access credentials for the configured model API, and lawful access to source reports. License: MIT. Additional restrictions: none for the authors' code and derived outputs; third-party publication rights continue to apply to source material.

Competing interests. The authors declare that they have no competing interests.

Funding. XT was supported by the Research Development Fund of Xi'an Jiaotong-Liverpool University (grant RDF2401022). The funder had no role in the design of the study; collection, analysis, or interpretation of the data; preparation of the manuscript; or decision to submit the work for publication.

Authors' contributions. YY conceived the study, developed the methodology and software, conducted the analyses and visualization, and drafted the manuscript. HF and MS curated the datasets and performed independent initial gold-standard extraction. XT and TT independently validated the gold-standard records, supervised the methodology, and contributed to interpretation. YP, CZ, CXZ, and YW contributed to interpretation of the results and critical revision of the manuscript. All authors reviewed and approved the final manuscript and accept accountability for the work.

Acknowledgements. An AI-assisted tool (OpenAI Codex) was used for language editing and LaTeX formatting during preparation of the submission files. The authors are responsible for reviewing and approving all scientific content.

Authors' information. Not applicable.

References

[1] Cook D, Mulrow C, Haynes R. Systematic reviews: synthesis of best evidence for clinical decisions. Annals of Internal Medicine. 1997;126(5):376–380. https://doi.org/10.7326/0003-4819-126-5-199703010-00006.

[2] Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. John Wiley & Sons; 2023.

[3] Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Medical Research Methodology. 2017;17(1):152. https://doi.org/10.1186/s12874-017-0431-4.

[4] Jonnalagadda S, Goyal P, Huffman M. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews. 2015;4:78. https://doi.org/10.1186/s13643-015-0066-7.

[5] Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen T. Single data extraction generated more errors than double data extraction in systematic reviews. Journal of Clinical Epidemiology. 2006;59(7):697–703. https://doi.org/10.1016/j.jclinepi.2005.11.010.

[6] Basu T, Goldsworthy S, Gkoutos G. A Sentence Classification Framework to Identify Geometric Errors in Radiation Therapy from Relevant Literature. Information. 2021;12(4):139. https://doi.org/10.3390/info12040139.

[7] Sim J, Huang X, Horan M, Stewart C, Robison L, Hudson M, et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. Artificial Intelligence in Medicine. 2023;146:102701. https://doi.org/10.1016/j.artmed.2023.102701.

[8] Imaichi O, Yanase T, Niwa Y. A Comparison of Rule-Based and Machine Learning Methods for Medical Information Extraction. In: Proceedings of the First Workshop on Natural Language Processing for Medical and Healthcare Fields. Asian Federation of Natural Language Processing; 2013. p. 38–42.

[9] Han H, Giles C, Manavoglu E, Zha H, Zhang Z, Fox E. Automatic document metadata extraction using support vector machines. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2003). IEEE Computer Society; 2003. p. 37–48.

[10] Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson E, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Medical Informatics and Decision Making. 2019;19:1. https://doi.org/10.1186/s12911-018-0723-6.

[11] Marshall I, Wallace B. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews. 2019;8(1):163. https://doi.org/10.1186/s13643-019-1074-9.

[12] Sundaram G, Berleant D. Automating systematic literature reviews with natural language processing and text mining: A systematic literature review. In: Proceedings of the Eighth International Congress on Information and Communication Technology (ICICT 2023). vol. 693 of Lecture Notes in Networks and Systems. Springer; 2023. p. 73–92.

[13] Yang Y, Wu Z, Yang Y, Lian S, Guo F, Wang Z. A Survey of Information Extraction Based on Deep Learning. Applied Sciences. 2022;12(19):9691. https://doi.org/10.3390/app12199691.

[14] Finn A, Kushmerick N. SVM based learning system for information extraction. In: Deterministic and Statistical Methods in Machine Learning. vol. 3635 of Lecture Notes in Computer Science. Springer; 2004. p. 319–339.

[15] Adnan K, Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data. 2019;6:91. https:

//doi.org/10.1186/s40537-019-0254-8.

[16] Sahoo S, Plasek J, Xu H, Uzuner Ö, Cohen T, Yetisgen M, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. Journal of the American Medical Informatics Association. 2024;31(9):2114–2124. https://doi.org/10.1093/jamia/ocae074.

[17] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877–1901.

[18] Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Medical Informatics. 2024;12:e55318. https://doi.org/10.2196/55318.

[19] Xu X, Zhu Y, Wang X, Zhang N. How to unleash the power of large language models for few-shot relation extraction? In: Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP). Association for Computational Linguistics; 2023. p. 190–200.

[20] Xu D, Chen W, Peng W, et al. Large language models for generative information extraction: a survey. Frontiers of Computer Science. 2024;18:186357. https://doi.org/10.1007/s11704-024-40555-y.

[21] Witte C, Schmidt D, Cimiano P. Comparing generative and extractive approaches to information extraction from abstracts describing randomized clinical trials. Journal of Biomedical Semantics. 2024;15(1):3. https://doi.org/10.1186/s13326-024-00305-2.

[22] Chen F, Zhang G, Fang Y, Peng Y, Weng C. Semi-supervised learning from small annotated data and large unlabeled data for fine-grained Participants, Intervention, Comparison, and Outcomes entity recognition. Journal of the American Medical Informatics Association. 2025;32(3):555–565. https://doi.org/10.1093/jamia/ocae326.

[23] Ghosh M, Mukherjee S, Ganguly A, Basuchowdhuri P, Naskar S, Ganguly D. AlpaPICO: Extraction of PICO frames from clinical trial documents using LLMs. Methods. 2024;226:78–88. https://doi.org/10.1016/j.ymeth.2024.04.005.

[24] Cumpston M, McKenzie J, Thomas J, Brennan S. The use of 'PICO for synthesis' and methods for synthesis without meta-analysis: protocol for a survey of current practice in systematic reviews of health interventions. F1000Research. 2020;9:678. https://doi.org/10.12688/f1000research.24469.2.

[25] Schmidt L, Finnerty Mutlu A, Elmore R, Olorisade B, Thomas J, Higgins J. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Research. 2021;10:401. https://doi.org/10.12688/f1000research.51117.3.

[26] Li L, Mathrani A, Susnjak T. Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of AI. Research Synthesis Methods. 2026;p. 1–48. https://doi.org/10.1017/rsm.2025.10065.

[27] Tsafnat G, Glasziou P, Choong M, et al. Systematic review automation technologies. Systematic Reviews. 2014;3:74. https://doi.org/10.1186/2046-4053-3-74.

[28] Schmidt L, Hair K, Graziosi S, Campbell F, Kapp C, Khanteymoori A, et al. Exploring the use of a large language model for data extraction in systematic reviews: A rapid feasibility study. In: Proceedings of the 3rd Workshop on Augmented Intelligence for Technology-Assisted Reviews Systems (ALTARS 2024). vol. 3832 of CEUR Workshop Proceedings. CEUR-WS.org; 2024. p. 1–30. Available from: https://ceur-ws.org/Vol-3832/paper2.pdf.

[29] Yisha Z, Zou P, Li S, Zhang L, Guo L, Gu A, et al. Assessing data extraction in randomized clinical trials with large language models. BMC Medical Research Methodology. 2026;26:33. https://doi.org/10.1186/s12874-025-02729-5.

[30] Das A, Ahmed S, Sakib S. Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models. arXiv preprint. 2025;https://doi.org/10.48550/arXiv.2504.19061. arXiv:2504.19061. [cs.CL].

[31] Li J, Chen J, Ren R, Cheng X, Zhao X, Nie J, et al. The dawn after the dark: An empirical study on factuality hallucination in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2024. p. 10879–10899.

[32] Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models. arXiv preprint. 2024;https://doi.org/10.48550/arXiv.2401.11817. arXiv:2401.11817. [cs.CL].

[33] Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023). Singapore: Association for Computational Linguistics; 2023. p. 6465–6488.

[34] Haan S. SemanticCite: Citation verification with AI-powered full-text analysis and evidence-based reasoning. arXiv preprint. 2025;https://doi.org/10.48550/arXiv.2511.16198. arXiv:2511.16198. [cs.CL].

[35] García F, Shi Q, Feng Z. Enhancing factual accuracy and citation generation in LLMs via multi-stage self-verification. arXiv preprint. 2025;https://doi.org/10.48550/arXiv.2509.05741. arXiv:2509.05741. [cs.CL].

[36] Mortezaagha P, Shaw J, Sun B, Rahgozar A. From chaos to clarity: schema-constrained AI for auditable biomedical evidence extraction from full-text PDFs. BMC Medical Research Methodology. 2026;26:119. https://doi.org/10.1186/s12874-026-02847-8.

[37] Yin Y. MAU-RCT-Extraction: Minimal Atomic Unit framework for extracting meta-analysis data from full-text study reports. GitHub repository. 2026;Commit f62c788c4279d7851d184e5dc907f6ae42b41632. Available from: https://github.com/yijiajiajiajiajia/MAU-RCT-Extraction/tree/f62c788c4279d7851d184e5dc907f6ae42b41632. Accessed 3 August 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2nonMAUprompt.txt