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.

Automated Meta-Analysis of Transcranial Magnetic Stimulation Based on GraphRAG: A Methodology Study

Yu Zhong
Chinese Academy of Medical Sciences & Peking Union Medical College
Jingna Jin
Chinese Academy of Medical Sciences & Peking Union Medical College
Xin Wang
Chinese Academy of Medical Sciences & Peking Union Medical College
Wang He
Chinese Academy of Medical Sciences & Peking Union Medical College
Tao Yin
bme500@163.com
Chinese Academy of Medical Sciences & Peking Union Medical College

Research Article

Keywords: Transcranial magnetic stimulation, Large language model, Retrieval-augmented generation, Meta-analysis

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10200494/v1

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

Additional Declarations: No competing interests reported.

Automated Meta-Analysis of Transcranial Magnetic Stimulation Based on GraphRAG: A Methodology Study

Yu Zhong[12],Jingna Jin[12], Xin Wang[12], Wang He[12], Tao Yin[12]
Corresponding author: Tao Yin, bme500@163.com
1. State Key Laboratory of Advanced Medical Materials and Devices, Institute of Biomedical Engineering, Tianjin Institutes of Health Science, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin 300192, China
2. Tianjin Key Laboratory of Neuromodulation and Neurorepair, Tianjin 300192, China

Abstract

Background: The therapeutic efficacy of transcranial magnetic stimulation (TMS) is highly dependent on the selection of stimulation parameters. Different parameter combinations for the same disease may lead to heterogeneous clinical outcomes, necessitating quantitative evidence synthesis via meta-analysis to optimize stimulation protocols. However, manual meta-analysis is time-consuming and prone to human error, making automated meta-analysis a worthwhile area of investigation. Large language models (LLMs) are capable of extracting information from unstructured text, offering potential for automated data extraction. Nevertheless, conducting a full meta-analysis requires systematic retrieval and synthesis of all relevant literature, rather than processing individual papers or isolated sections. Conventional retrieval-augmented generation (RAG), which uses text chunks as retrieval units, cannot satisfy this requirement. Therefore, this paper proposes a literature extraction and retrieval method based on the GraphRAG paradigm.

Methods: The methodological workflow comprises three parts: parameter extraction, knowledge graph construction, and query-based retrieval. In the parameter extraction phase, we design a long-text processing strategy based on title-based chunking and sliding windows, along with prompt templates combining semantic blocks for basic information, outcome measures, TMS parameters, and risk of bias. In the knowledge graph construction phase, we use papers rather than tokens as entity nodes and perform community clustering to build the knowledge graph. In the query retrieval phase, we construct a hybrid retrieval mechanism integrating keywords, structured storage, and semantic vectors to meet the fine-grained requirements of queries.

Results: We downloaded 26 meta-analyses and their 250 supporting articles from platforms such as Web of Science and Google Scholar, and constructed a TMS-specific meta-analysis dataset, TMS-MAD, for method evaluation. Evaluation metrics included BLEU-4, Exact Match (EM), MAPE, Kappa, and coincidence rate. Across four models(GPT-5.4, DeepSeek-V4-flash, Gemini-3.1-flash-lite, and Claude-haiku-4.5), the EM values for fixed-field extraction all exceeded 90%, while semantic similarity for open-ended fields was above 0.8. The average time for extracting parameters from a single paper was 86.71 seconds, effectively reducing the time required for meta-analysis.

Conclusions: The proposed method effectively supports data extraction and retrieval tasks for meta-analysis in the TMS domain, providing a feasible solution for automated meta-analysis.

Keywords: Transcranial magnetic stimulation; Large language model; Retrieval-augmented generation; Meta-analysis

Introduction

Transcranial magnetic stimulation (TMS) is a non-invasive neuromodulation technique that induces local cortical electrical activity by applying time-varying magnetic fields over the scalp. It has been widely applied in research on the treatment of depression[1], post-stroke motor disorders[2], and other diseases. However, the therapeutic efficacy of TMS is highly dependent on stimulation protocols, including stimulation type, site, and other parameters[3]. For the same disease, different parameter combinations may lead to significantly divergent treatment outcomes; therefore, systematic synthesis of existing research evidence is necessary to guide the selection of optimal parameter regimens. Meta-analysis, as a method for quantitative synthesis of research evidence, can combine effect sizes to evaluate intervention efficacy and plays a pivotal role in evidence-based medicine[4]. Nevertheless, traditional meta-analysis relies heavily on manual operations, requiring 3 to 5 experts to collaborate for several months to manually extract information such as stimulation frequency, sample size, and effect sizes. This process is not only resource-intensive but also difficult to update as new literature emerges[5]. Hence, exploring automated approaches for meta-analysis is of substantial practical significance.

Large language models (LLMs) are deep neural network-based natural language processing systems built on the Transformer architecture. They are capable of capturing complex linguistic statistical regularities and contextual semantic dependencies, and have demonstrated strong performance in user question answering and text summarization. LLMs have been widely adopted in the medical domain[6], for tasks such as retrieval and extraction of medical information[7], answering patient queries[8], and optimizing clinical decision-making[9]. Moreover, with the advent of concepts such as multimodality[10], agents[11], and harness engineering[12], LLMs can further perform part of the text and data analysis work traditionally done by specialized physicians[13,14]. For instance, the BioResearcher model automates the entire biomedical research workflow from literature review to executable code generation through a modular architecture[15]; a multi-agent medical question-answering framework enhances complex reasoning capabilities through case generation and voting mechanisms[16].

Consequently, researchers have attempted to apply LLMs to certain steps of meta-analysis[17], such as using LLMs to screen relevant supporting articles based on titles and abstracts[18], and employing LLMs for data extraction[19]. However, conducting a complete meta-analysis requires systematic retrieval and integration of all relevant literature, not just processing individual papers or isolated paragraphs. LLMs themselves lack dynamic retrieval capabilities, and their knowledge is limited to the cutoff of model training[20,21], making them unable to cover real-time updated literature collections. Retrieval-augmented generation (RAG) technology can construct external documents into retrievable knowledge bases without fine-tuning the model[22], injecting relevant information at the model input stage to ensure source reliability. This paradigm aligns well with the evidence-based medicine principle of centering on research findings. For example, the MedTutor model generates teaching materials through hybrid retrieval from medical textbooks and literature[23]; a RAG framework designed for PubMed can automatically generate Boolean queries to improve retrieval efficiency[24]; the RefAI model, combining literature retrieval and ranking algorithms, effectively mitigates LLM hallucination issues[25]. In the TMS domain, one study constructed a vector database of patient information and repetitive TMS treatment protocols, enabling LLMs to generate personalized treatment

recommendations via similar patient matching[26]; another study built semantic mappings using pre- and post-TMS reports and generated personalized rTMS regimens through a voting mechanism[27]. These works all employ RAG but focus on depression, without validating on diseases with greater heterogeneity in treatment protocols. More critically, when treatment regimens for the same disease in the knowledge base vary substantially, these systems cannot assess the efficacy of the generated stimulation plans.

Integrating LLMs and RAG into the meta-analysis workflow holds promise for further enhancing LLM capabilities in the medical domain[28]. Existing studies have followed core meta-analysis principles, screening high-quality, low-heterogeneity literature via extrapolation and heterogeneity analyses as LLM inputs, improving question-answering accuracy on the MedQA dataset[29]. Other work has proposed a multi-agent framework integrating RAG-retrieved literature to improve systematic review and meta-analysis efficiency while ensuring text traceability; the authors suggest that LLMs can undertake part of the automation tasks, but experts should retain final decision-making authority[30]. The conventional approach to combining RAG and meta-analysis involves RAG handling literature retrieval, with LLMs performing partial meta-analysis operations based on retrieved content. Since RAG uses text chunks as retrieval units, it can only return relevant passages and lacks structured associations among medical entities (e.g., treatment methods, outcome measures). The retrieval granularity is thus ill-suited for the structured data aggregation required across multiple papers in meta-analysis[31]. Consequently, current solutions mainly focus on review report or abstract generation rather than true effect-size calculation for meta-analysis.

In summary, this study focuses on the specific therapeutic modality of TMS. We employ LLMs to extract detailed structured parameters from papers (including stimulation type, frequency, intensity, and inter-train intervals), and leverage the GraphRAG paradigm to construct a knowledge graph with papers as nodes, using hybrid storage and community clustering. Based on different queries, we filter literature and calculate effect sizes, thereby validating the feasibility and effectiveness of using LLMs for complete meta-analysis.

Methods

This study designs a knowledge graph storage structure based on the GraphRAG paradigm, using papers rather than text chunks as entities, to automate meta-analysis operations. The method follows the GraphRAG workflow and is divided into three parts: literature data extraction, knowledge graph construction, and multi-level querying.The methodological workflow is illustrated in Figure 1.

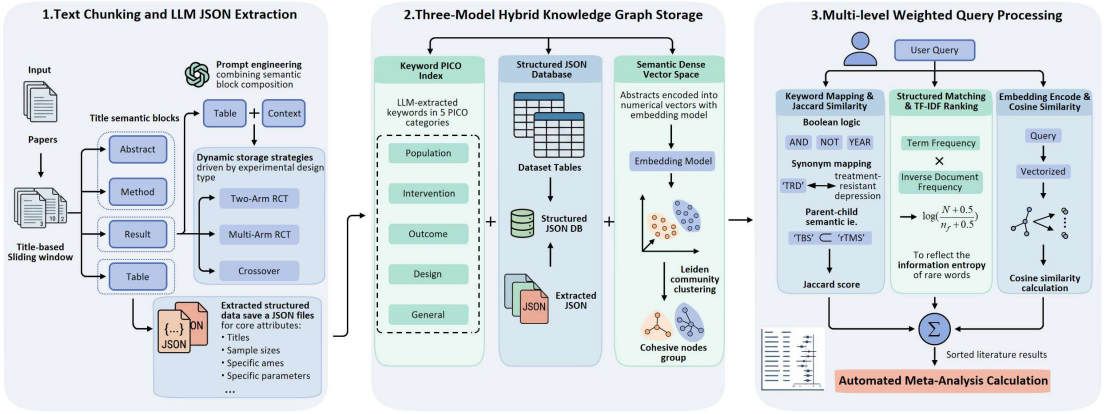


Figure 1 Methodological flowchart

For the extraction part, traditional regular expressions struggle to match the diverse expressions in papers and cannot interpret the meaning of data within tables. Therefore, LLMs are employed for automated parameter extraction. However, using LLMs also necessitates addressing two key issues: 1) The context window limitation of LLMs for long-text extraction. When the provided paper material exceeds the model's processing length, truncation may occur, leading to data loss. Moreover, outcome data are typically located in the middle of papers, where model attention is often weakest, resulting in significantly degraded retrieval and reasoning performance, a phenomenon known as the "lost-in-the-middle" problem[32]. 2) The diversity of experimental designs and complex text structures. Clinical trial papers adopt various design types, such as multi-arm RCTs and crossover designs. Even when targeting TMS meta-analysis, a single data storage format cannot accommodate the variability in experimental designs.

Chunking

To alleviate the information extraction difficulties caused by context window limitations when processing long texts with LLMs, this study adopts a title-based text chunking strategy. For the full text D of a paper D_i from the original paper collection, the title-based chunking algorithm H_{title} first divides D_i into semantic blocks of varying lengths. For text blocks exceeding the model's context window length, a sliding window algorithm T_{slide} further splits them into multiple partially overlapping sub-blocks, with table protection processing T_{table} applied to blocks containing tables to ensure table integrity. Finally, based on the actual content of each text block, semantic classification and reorganization C are performed to categorize them into abstract blocks, method blocks, result blocks, and table blocks, yielding the comprehensive information set S_i of the paper for LLM extraction. The above process can be formally expressed as:

$$S_i = C(T_{\text{slide}}(T_{\text{table}}(H_{\text{title}}(D_i))))$$

Data Extraction

After chunking and categorization, this study combines these four types of semantic blocks and designs different prompt templates to obtain corresponding information. For instance, combining abstract and method blocks yields basic experimental information such as study design, disease type, and intervention measures; combining table blocks with their surrounding context provides statistical indices including effect sizes and confidence intervals. Extraction results are stored in JSON format as structured data, serving as the foundation for subsequent entity-relationship extraction and knowledge graph construction. Given the regular correspondence between sections and information in clinical trial papers (e.g., stimulation protocols appear only in method blocks, effect sizes only in result blocks or tables), pre-classifying text blocks significantly compresses the LLM's retrieval window, thereby reducing information omission and cross-section interference.

Table 1 Prompt templates for chunking strategies

Text Block Combination	Extraction Target	Prompt Core Instruction Example	Output Field Examp
Abstract + Method	Basic experimental information	"Extract meta-analysis basic information from the following paper excerpts. Key focus: 1. Sample size refers specifically to the number randomized at group allocation (not study start or end); 2. Return null if keywords are not explicitly listed; 3. Return only JSON format."	{"title":"repetitive transcranial magnetic", "author": "E. M. Khedr", "year": 2003, "population": "Parkinson's disease", "sample_size": 36, "control_sample_size": 19, "intervention_sample_size": 17, "num_follow_up_people": 31, "intervention_intervention":"real-rTMS", "study_design_type": "two-arm rct"}
Method	Outcome measure types	"Extract outcome measure definitions from the paper. Key focus: 1. Clearly distinguish primary from secondary; 2.	{"primary_outcomes": [{"name": "Motor functions improvement", "scale": "Unified Parkinson's Disease Rating Scale (UPDRS)", "description": "Assessment of...",

		Extract scale names (HAMD-17/MADRS, etc.); 3. Record measurement time points (baseline/4weeks, etc.); 4. If no explicit classification is provided, let the LLM judge based on general conventions; 5. Output JSON, leaving arrays empty if not mentioned."	"measurement_time": "post-treatment"}], "secondary_outcomes": [{"name": "Walking speed", "scale": "Walking speed test", "description": "Time (seconds) to...", "measurement_time": "post-treatment"}]}
Method + Table	TMS experimental parameters	"Extract TMS study parameters from the paper. Key rules: 1. Frequency: combine burst stimulation expressions (e.g., '50Hz bursts at 5Hz'); 2. Intensity: distinguish RMT from AMT, annotated as 'xx% of RMT' or 'xx% of AMT'; 3. Inter-session interval: record only for aTMS with interval $\leqslant$ 1 day, otherwise null; 4. Sham angle and method must be explicitly extracted; 5. Set unmentioned parameters to null; return only JSON."	{"tms_type": "rTMS", "stimulation_frequency": "5Hz", "stimulation_intensity": "120% of RMT", "session_number": 10, "brain_target": "optimal position for lower limbs (EDB) and hand", "hemisphere": "right then left", "total_number_pulses_per_session": 2000, "train_duration": "2s", "train_time_interval": "8s", "inter_session_interval": null, "coil_tilt_angle_for_sham": "angled away from the head", "sham_type": "coil elevated and angled away from the head"}
Table + Table context	Meta-analysis data	"Extract meta-analytic numerical data from tables and surrounding context. Key rules: 1. Prioritize extracting complete statistics (n, mean, SD) for primary outcomes; 2. For multiple time points, take the longest follow-up; 3. If SD is unavailable, extract CI/p-value/t-value as alternatives; 4. Prefix 'change in' for change values; 5. Do not extract baseline values; 6. Set unmentioned to null; return only JSON."	"primary_outcomes": [{ "outcome_name": "Unified Parkinson's Disease Rating Scale (UPDRS)", "direction": "lower_is_better", "time_point": "1 month", "intervention_group": {"n": 16,"mean": 15.6,"sd": 6.5, ...}, "control_group": {"n": 15,"mean": 23.7,"sd": 7.6, ...}, "effect_size": {...}], "secondary_outcomes": [...],
Abstract + Method + Result	Literature quality assessment	"Assess RCT bias risk using the RoB2 tool. Key rules: 1. Evaluate five domains separately (Low/Some concerns/High); 2. Provide rationale and supporting evidence for each domain; 3. Overall rating follows the highest risk level among the five domains (weakest-link principle); 4. Base assessment on actual information in the paper, not speculation; 5. Return only JSON."	{"domain1_randomization": { "risk_level": "Some concerns", "rationale": "The description of the randomization method is insufficient. Only the closed envelope method is mentioned", "supporting_evidence": "Patients were assigned in a randomized pattern to one of the two groups... by using closed envelops."}, "domain2_deviations": {...}, ... "domain5_selective_reporting": {...}, "overall_bias_risk": {...}}

The JSON storage format for data extraction must be differentiated according to the experimental design type of the original literature. Since multi-arm and crossover trials introduce additional data structures compared to standard two-arm parallel designs. The former requiring pairwise comparisons among multiple intervention groups, the latter requiring consideration of

carryover effects and washout periods. A unified storage format cannot accommodate all design types. Therefore, this study adopts a design-type-driven dynamic storage strategy. The LLM first extracts the experimental design type from the paper, including open-label, two-arm RCT, multi-arm RCT, and crossover designs, and then selects the corresponding JSON storage structure based on predefined template mapping rules. Simultaneously, to ensure cross-design data integrability and code generalizability for subsequent meta-analytic computations, all prompt templates maintain consistent core attribute naming conventions and data types, adding corresponding fields only for design-specific structures. This approach preserves data completeness while reducing analytical complexity.

Knowledge Graph Construction

The storage architecture and query process constitute a pair of mutually constraining core design elements. The organization of query paths must strictly adapt to the underlying storage structure, while different storage formats directly affect retrieval accuracy and efficiency. Under the conventional RAG vector database or GraphRAG paradigm using tokens as basic entities, the system typically returns discrete text fragments lacking the encapsulation capability for complete clinical trial parameter units (e.g., triples of outcome measures, effect sizes, and time points), making it difficult to directly satisfy the quantitative synthesis requirements of meta-analysis.

To address the above issues, this study proposes a hybrid storage strategy combining three information representations: keyword storage, structured storage, and semantic storage. Keyword indexes provide precise matching and rapid filtering; structured storage, based on the JSON files from the data extraction phase, preserves complete parameter information; semantic storage enables generalized retrieval of near-synonymous expressions. On this basis, similarity scores between nodes are computed via weighted aggregation to establish entity associations, and the Leiden community detection algorithm is applied to partition the node graph into semantically cohesive communities. This design balances retrieval granularity and information completeness, providing structured support for subsequent relevant literature querying and meta-analytic computations.

For keyword storage, guided by the PICO principle in evidence-based medicine, the LLM extracts five categories of keywords from the abstract and full text: population characteristics, intervention methods, outcome measures, study design, and general keywords, with 3-8 keywords per category to ensure concise yet comprehensive summarization. For structured storage, we directly reuse the JSON files generated in the data extraction phase to obtain more specific experimental design parameters, supporting fine-grained user queries. For semantic storage, the BAAI/bge-base-en-v1.5 embedding model is employed to convert paper abstracts into numerical vectors, mapping similar TMS experimental designs into neighboring positions in the vector space.

Multi-Level Querying

The query logic is symmetrically designed with the storage architecture, controlling retrieval granularity by weighting keyword matching scores, structured matching scores, and semantic vector embedding scores.

$$\text{Score}_{\text{total}} = \alpha \cdot \text{Score}_{\text{keyword}} + \beta \cdot \text{Score}_{\text{structured}} + \gamma \cdot \text{Score}_{\text{embedding}}$$

where $\alpha + \beta + \gamma = 1$ are adjustable weights according to query type.

Keyword Query: Boolean retrieval logic is employed, with mandatory keywords (AND), exclusion keywords (NOT), and year restrictions for preliminary filtering. Under this method, keyword matching scores are calculated. Given the diversity of natural language expressions and the limitations of regular expression matching, this study constructs two types of semantic mapping tables: 1) Synonym mapping table. Normalizing different expressions of the same concept, e.g., mapping "treatment-resistant depression" to its abbreviation "TRD"; 2) Parent-child mapping table. Handling semantic inclusion relationships, e.g., "theta burst stimulation" (TBS) as a subtype of "repetitive transcranial magnetic stimulation" (rTMS), establishing hierarchical mapping from child to parent.

Through the above mapping relationships, semantically related literature can be retrieved to the greatest extent possible. In the scoring strategy, direct exact matches are assigned higher weights, while partial matches that cannot be precisely matched via mapping tables are computed using Jaccard similarity. The keyword score $\text{Score}_{\text{keyword}}$ is expressed as:

$$\text{Score}_{\text{keyword}} = \frac{\sum_{m=1}^{M} \text{Score}_{k_m}}{M}$$

$$\text{Score}_{k_m} = \begin{cases} 1, & \text{if } k_m \text{ exactly matches } K_n \\ J(k_m, K_n), & \text{if partial match via mapping tables} \end{cases}$$

where the user query keyword set is k, set k contains M elements, the paper keyword set is K, set K contains N elements, K_n denotes an element in set K, the Jaccard similarity function is J, and J is defined as:

$$J(k_m, K_n) = \frac{|\text{Seg}(k_m) \cap \text{Seg}(K_n)|}{|\text{Seg}(k_m) \cup \text{Seg}(K_n)|}$$

where the tokenization function $\text{Seg}(\cdot)$ splits strings into element sets based on spaces.

Structured Query: This study references the core ideas of the BM25 ranking algorithm, integrating term frequency (TF) and inverse document frequency (IDF) weighting mechanisms to compute matching scores for structured fields. Specifically, IDF quantifies the informational distinctiveness of feature terms across the document collection. In the TMS-specific knowledge base constructed in this study, "TMS" is a high-frequency general term appearing across most paper nodes, thus having a low IDF value. In contrast, terms like "deep TMS" appear only in specific literature and possess higher IDF values. Therefore, IDF weighting effectively suppresses interference from common high-frequency terms and assigns higher retrieval weights to more discriminative feature terms, thereby improving structured query precision. The IDF formula is:

$$\text{IDF}(f) = \log\left(\frac{N + 0.5}{n_f + 0.5}\right)$$

where f is the feature term, N is the number of paper nodes in the knowledge base, and n_f is the number of paper nodes containing that feature term.

Semantic Vector Embedding Query: The BAAI/bge-base-en-v1.5 embedding model encodes both the user query and paper abstracts into dense vectors. Cosine similarity between the two is then computed as the semantic matching score during querying, with the formula:

$$\text{Score}_{\text{structured}} = \frac{\mathbf{v}_Q \cdot \mathbf{v}_D}{\|\mathbf{v}_Q\| \|\mathbf{v}_D\|}$$

where $\mathbf{v}_Q$ and $\mathbf{v}_D$ are the embedding vectors of the query and paper abstract, respectively. The result serves as one of the inputs to the multi-level query.

Dataset Construction

To evaluate the proposed method, we constructed a dedicated dataset, designated as TMS-MAD (TMS Meta-Analysis Dataset), following the structured format defined in the data extraction section. The dataset was compiled by searching Web of Science and PubMed using the query: ("transcranial magnetic stimulation" OR "TMS") AND ("meta-analysis" OR "systematic review"). After screening, we included 26 TMS-related meta-analyses and further retrieved their 250 supporting original studies, yielding a total of 277 documents. The inclusion criteria were: (1) studies on TMS applications in neurological disorders; (2) document type as clinical original research or systematic review/meta-analysis; (3) meta-analyses published in English; and (4) availability of full text. The dataset composition is summarized in Table 2.

Table 2 Basic information of the TMS-MAD dataset

Dimension	Category	Count
Total literature	Meta-analysis original	26
	Supporting papers	250
Disease types covered by meta-analyses	Depression	8
	Alzheimer's disease	4

	Stroke	6
	Parkinson's disease	2
	ADHD	3
	OCD	2
	Headache	1
	Cerebellar ataxia	1
	Two-arm RCT	162
Supporting	Multi-arm RCT	50
paper study	Open-label	16
types	Crossover	15
	Case report	7

In terms of disease distribution, the included literature covers major TMS indications such as depression, stroke, Alzheimer's disease, and ADHD. Regarding study types, original studies are predominantly randomized controlled trials, with some non-randomized controlled studies. All literature was manually annotated according to the format defined in the data extraction section, forming a structured information repository for subsequent extraction performance evaluation and retrieval testing.

In summary, this study designed a title-based chunking and sliding window preprocessing strategy, multi-type parameter-oriented prompt engineering templates, and a three-modality hybrid retrieval mechanism integrating keywords, structured data, and semantic vectors for structured information extraction and retrieval from clinical trial literature. To evaluate the feasibility and effectiveness of the proposed method, the next section presents quantitative analyses in terms of extraction accuracy, retrieval recall, and system response efficiency.

Evaluation and Results

Evaluation Metrics

To validate the accuracy of the parameter extraction and query modules in the proposed method, we adopted appropriate evaluation metrics tailored to different data types.

For data extraction, based on the field attributes of the storage structure, the evaluated fields were categorized into four types: numeric, short-text, long-text, and semantic embedding vector, each with corresponding evaluation criteria.

1. Numeric fields: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were used to measure the deviation between extracted and benchmark values.

2. Long-text fields: BLEU-4 (Bilingual Evaluation Understudy-4) was employed, which computes similarity between generated and reference texts based on the weighted geometric mean of 1-gram to 4-gram matching degrees.

3. Short-text fields: For fields with few characters (e.g., first author, patient disease), where BLEU-4 is inappropriate due to insufficient 4-gram formation, Exact Match (EM) was used, counting a response as correct only if it exactly matches the reference.

4. Semantic embedding vector fields: Cosine similarity was used to assess semantic equivalence for non-numeric attributes, accommodating cases where expressions differ but meanings are similar.

Additionally, classification tasks were included in parameter extraction evaluation, such as the three-class classification (Low/Some concerns/High) for the five domains in literature quality assessment. For classification labels, accuracy, Cohen's Kappa coefficient, and semantic similarity were used to evaluate LLM classification capability based on paper text.

For the query phase, paper coincidence rate was adopted as the evaluation metric, comparing the retrieved literature set with the original supporting paper set. Coincidence rate is defined as the ratio of the intersection to the union of the two sets.

Parameter Extraction Performance Evaluation and Results

To evaluate the performance of different LLMs under the proposed parameter extraction method, we called the APIs of current mainstream LLMs, including GPT-5.4, DeepSeek-V4-flash, Gemini-3.1-flash-lite, and Claude-haiku-4.5. All models used identical prompt templates and chunking strategies, and were compared against the constructed dataset in terms of the above metrics, RoB2 literature quality assessment accuracy, and time/token consumption.

For fixed fields, as shown in BLEU-4, EM, and Kappa results, all models demonstrated high accuracy in extracting titles, first authors, and study design types. As shown in Figure 2(a), BLEU-4 values for title extraction across the four models ranged from 92.6% to 94.4%, while Kappa values for study design type ranged from 84.8% to 88.2%. For open fields, model performance under exact-match metrics was poor. ChatGPT-5.4 achieved a BLEU-4 of only 12.4% for intervention measures, DeepSeek-V4-Flash 22.0%, and Gemini-3.1-Flash-Lite, the best among them, only 32.5%.。

Therefore, to compensate for the limitations of exact-match metrics, we further used Sentence-BERT to compute cosine similarity between generated text and manual annotations to test semantic similarity. As shown in Figure 2(b), for open fields such as intervention measures and study populations, all models achieved significantly higher semantic similarity scores than exact-match metrics. For example, for ChatGPT-5.4, the BLEU-4 for intervention group interventions was only 12.4%, while cosine similarity was 83.3%; for the study population field, EM was 23.7% while cosine similarity was 86.6%. For TMS parameter fields (e.g., TMS type, stimulation intensity, stimulation target), cosine similarity scores exceeded 0.89 for all models. These results indicate that while LLM extraction results may not exactly match manual extraction, they are semantically similar.

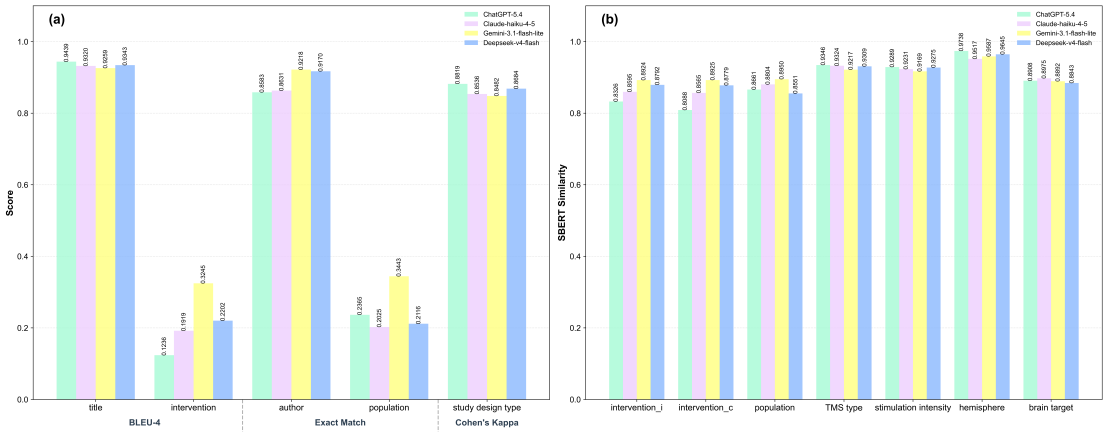


Figure 2 Extraction scores of four LLMs. (a) BLEU-4, EM, and Kappa scores for fixed and some open fields; (b) Similarity scores for open fields

For numeric parameters such as sample size, age, sex ratio, and pulses per session, MAE and MAPE were used for evaluation. As shown in Figures 3 and 4, all models performed stably on the year field, with MAPE between 0.2% – 0.3%; sample-size-related fields had MAPE in the 7 – 12% range; male percentage MAPE varied more widely, with DeepSeek-V4-Flash reaching up to 22.01%.

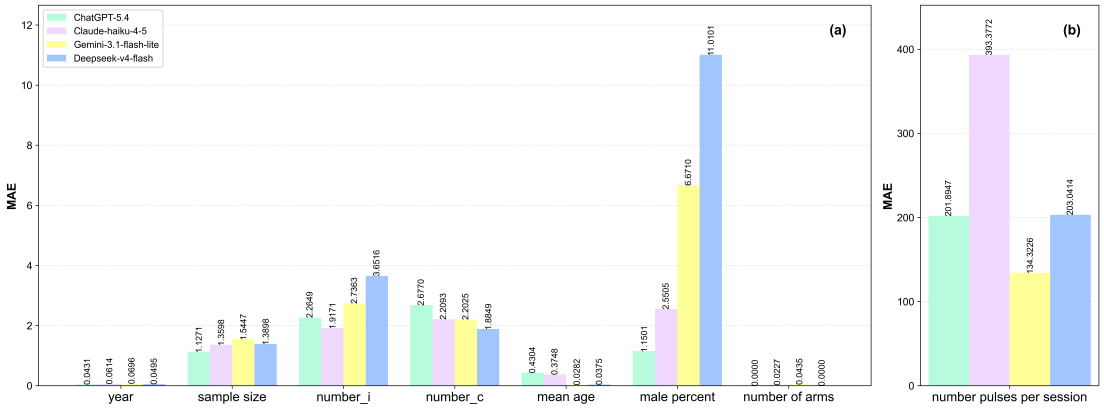


Figure 3MAE values of four LLMs on numeric fields. (a) Errors on conventional numeric fields; (b) Errors on pulse count fields

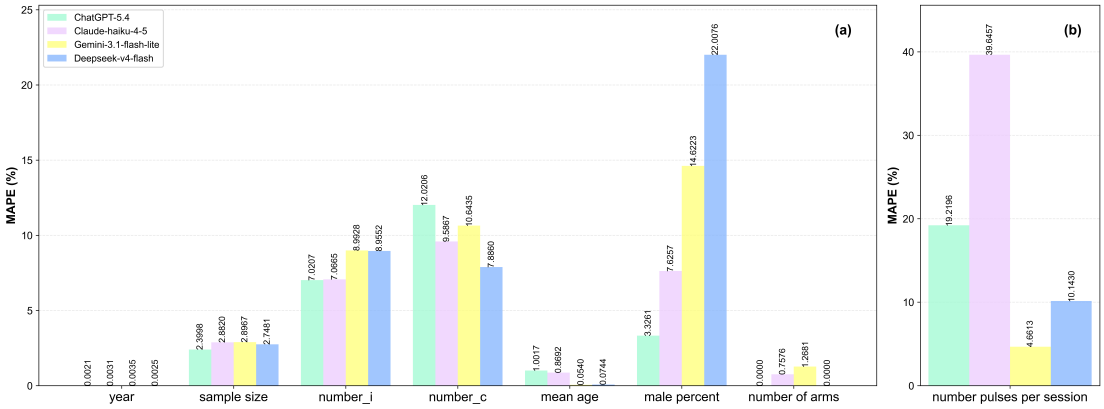


Figure 4 MAPE values of four LLMs on numeric fields. (a) Errors on conventional numeric fields; (b) Errors on pulse count fields

Literature Quality Assessment Evaluation and Results

The above evaluation focused on textual and numerical accuracy in parameter extraction. For meta-analysis, the quality assessment of included literature is also a critical factor affecting the reliability of pooled effect sizes. To validate the applicability of LLMs in evidence quality appraisal, we further evaluated model performance on the RoB2 risk-of-bias assessment tool. RoB2 covers five core domains and an overall bias judgment. For each domain, the model must output a risk level (Low/Some concerns/High) along with a rationale and supporting evidence. Table 3 presents the classification consistency and semantic similarity of rationales and evidence for each model.

Table 3 RoB2 assessment performance of four LLMs

Models	Evaluation index	randomization	deviation	missing data	outcome measurement	elective reporting	overall bias risk
ChatGPT-5.4	cohens kappa	0.48525	0.15354	0.16534	0.51350	0.10784	0.33182
	accuracy	0.70954	0.47718	0.51867	0.77593	0.49793	0.72614
	rationale similarity	0.83400	0.79591	0.79135	0.83326	0.76479	
	evidence similarity	0.79541	0.73674	0.71400	0.75407	0.71217	
Claude-haik	cohens	0.70152	0.52256	0.22985	0.59882	0.42342	0.42147

u-4-5	kappa						
	accuracy	0. 83058	0. 71488	0. 56198	0. 83058	0. 68182	0. 79339
	rationale similarity	0. 84448	0. 80075	0. 77268	0. 79521	0. 77197	
	evidence similarity	0. 79686	0. 75500	0. 70960	0. 74524	0. 71410	
Gemini-3.1-flash-lite	cohens kappa	0. 66706	0. 41577	0. 25559	0. 37992	0. 52695	0. 39733
	accuracy	0. 80328	0. 63525	0. 59426	0. 70492	0. 75	0. 72541
	rationale similarity	0. 83137	0. 78154	0. 78138	0. 82903	0. 76929	
	evidence similarity	0. 79022	0. 70336	0. 69902	0. 69928	0. 66876	
Deepseek-v4-flash	cohens kappa	0. 57909	0. 33727	0. 20596	0. 39341	0. 41873	0. 38171
	accuracy	0. 75620	0. 59917	0. 57025	0. 73140	0. 69008	0. 75207
	rationale similarity	0. 84409	0. 80446	0. 78761	0. 84247	0. 79242	
	evidence similarity	0. 80639	0. 75208	0. 71871	0. 73509	0. 68771	

Model performance on the RoB2 task varied. Claude-haiku-4-5 achieved the highest Kappa and accuracy scores in most domains, with overall bias risk accuracy reaching 79.34%, outperforming other tested models. Additionally, semantic similarity scores for rationales and evidence were all above 0.6 across models, with no significant differences. In specific domains, performance on the randomization process was relatively strong, with accuracy generally above 70%, while accuracy for the missing outcome data domain was poor, with all Kappa coefficients below 0.26.

Resource Consumption Evaluation and Results

To evaluate the reduction in human effort and time costs enabled by the proposed method for automated meta-analysis, we also examined token and time consumption for knowledge graph construction across models. Results are shown in Table 4.

Table 4 Token and time consumption across model APIs

	ChatGPT-5.4	Claude-haiku-4-5	Gemini-3.1-flash-lite	Deepseek-v4-flash
Total time(s)	21529.81	14868.33	8687.61	6,490,386
Total token consumption	6,510,105	7,739,532	6,490,386	7,142,761
Average time(s)	86.12	59.47	34.75	166.50
Average token consumption	26040.42	30958.13	25961.54	28571.04

Although API call variations (rate limits, request timeouts, etc.) caused notable differences in total time, across the four tested APIs, the average processing time per paper was 86.71 seconds, with an average token consumption of 27,882.78, substantially reducing the time and resources required for meta-analysis parameter extraction.

Hybrid Retrieval Strategy Evaluation and Results

After evaluating LLM performance on parameter extraction and literature quality assessment, we evaluated the effectiveness of the hybrid retrieval method. The evaluation used an ablation

design, comparing the hybrid method against three unimodal approaches (keyword-only, structure-only, and semantic-only) on 26 query tasks. Each query was designed to mimic the core question of the corresponding meta-analysis, and the evaluation metric was the coincidence rate between the retrieved literature set and the original supporting papers. Boxplots of the coincidence rates for the four query methods are shown in Figure 5.

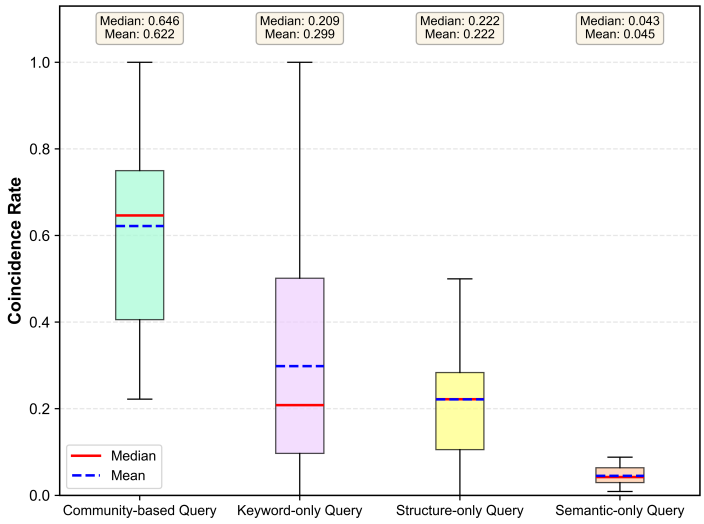


Figure 5 Coincidence rates of four query methods

The number of retrieved documents varied across methods: hybrid retrieval averaged 16.04 papers, keyword-only 62 papers, structure-only 22.58 papers, and semantic-only 231.54 papers. For semantic retrieval, most papers scored between 0.6 and 0.75, making threshold setting difficult; retrieving too many papers lowers the coincidence rate and fails to reflect semantic retrieval effectiveness. Therefore, for semantic-only queries, we additionally used Top-25 accuracy (Recall@25): the proportion of original supporting papers among the top 25 retrieved, to assess ranking quality. Recall@25 bar chart results are shown in Figure 6.

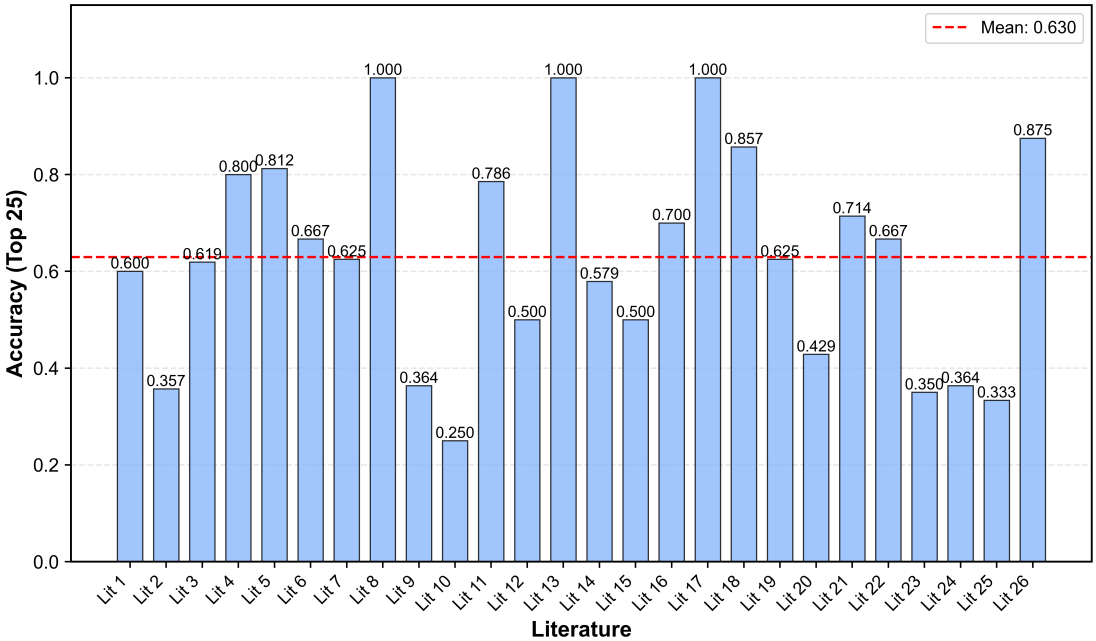


Figure 6 Recall@25 scores for semantic queries across 26 meta-analyses

The mean Recall@25 score was 0.6297, indicating that although semantic queries returned a

large volume of literature, the most relevant papers were prioritized at the top of the result list, enabling weighted filtering of relevant papers in the hybrid retrieval.

Forest Plot Replication Evaluation and Results

The above evaluations tested the performance of parameter extraction, quality assessment, and hybrid retrieval modules separately. In practical meta-analysis applications, these modules must work in concert to derive effect sizes for stimulation protocols. Therefore, we selected two meta-analysis papers with high coincidence rates from the hybrid retrieval results, used their core clinical questions as queries, and compared the forest plots reproduced by our method with those reported in the original meta-analyses. The query for Figure 7 was "To evaluate the cognitive function of rTMS in the treatment of patients with PSCI"; the query for Figure 8 was "I would like to know the efficacy and safety of repetitive transcranial magnetic stimulation in cerebellar ataxia."

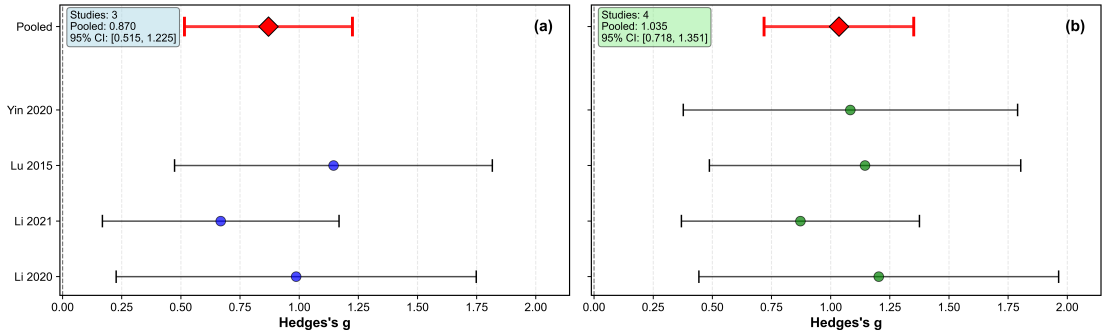


Figure 7 Forest plot comparison for the MoCA scale in Paper 10. (a) Code-generated results; (b) Original paper results

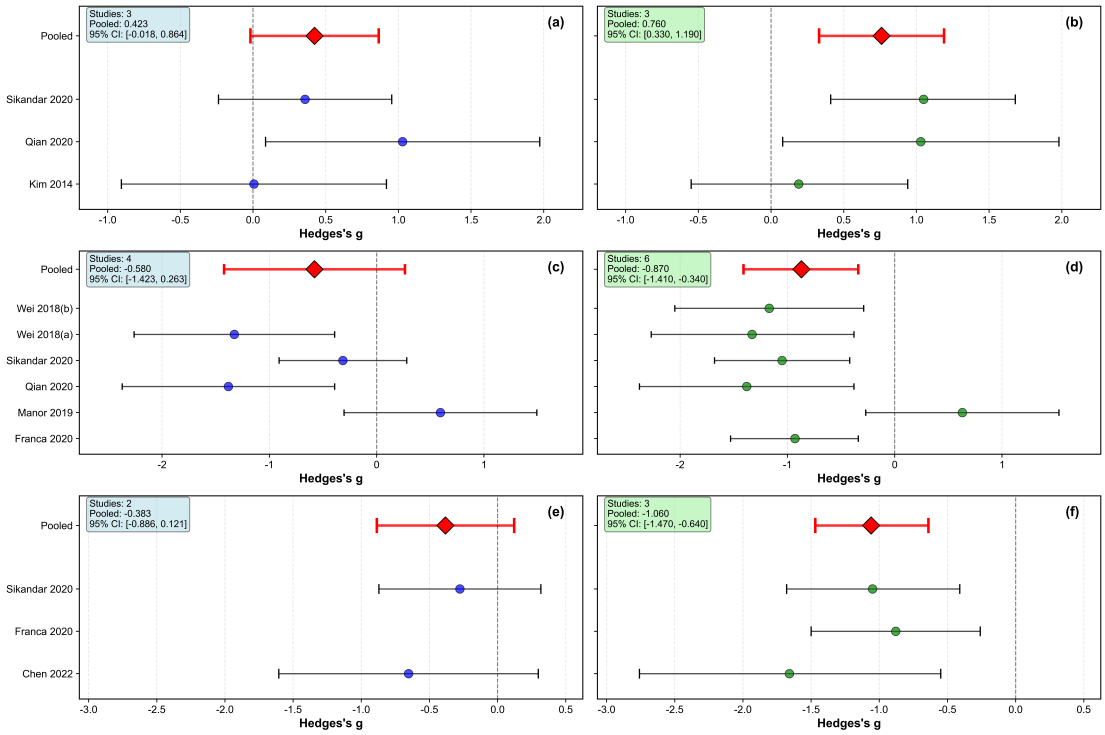


Figure 8Forest plot comparison for three outcome measures in Paper 17. (a,b) BBS; (c,d) SARA; (e,f) ICARS. Left: code-generated results; Right: original results

Missing data in the figures are due to the fact that some clinical study outcome measures

were presented as images, with specific data not described in the surrounding text. Since the proposed method does not involve image processing, data extraction for such cases was incomplete.

Discussion

This paper addresses the challenges of structured data extraction and cross-document information integration in LLM-based TMS meta-analysis by designing a GraphRAG-based literature extraction and retrieval method. We propose a title-structure- and sliding-window-based long-text chunking strategy to mitigate the impact of LLM context window limitations on extraction completeness. We design a set of semantic-block-combined prompt templates for multi-type parameter extraction, achieving automated conversion from unstructured medical literature to structured JSON data. We construct a paper-entity-based hybrid storage and retrieval mechanism that effectively balances retrieval granularity and information integrity. Using our self-constructed moderately sized TMS meta-analysis dataset covering depression, Alzheimer's disease, and other conditions (TMS-MAD), we validated the method on four models: GPT-5.4, DeepSeek-V4-flash, Gemini-3.1-flash-lite, and Claude-haiku-4.5, finding that EM values for fixed-field extraction all exceeded 90%, semantic similarity for open fields was above 0.8, and the average time for extracting parameters from a single paper was 86.71 seconds, effectively reducing meta-analysis time.

Following the evaluation of parameter extraction, literature retrieval, and forest plot replication, we analyze the observed results. First, in the parameter extraction phase, as shown in Figure 2, current mainstream LLMs can effectively understand and execute structured output instructions, achieving robust extraction for well-defined fields, consistent with findings by Gartlehner et al.[4]. Further analysis of extraction failures reveals that the primary cause is not insufficient model capability but limitations in the preprocessing stage. Some PDF documents, when converted to Markdown, suffer from text recognition errors, layout parsing errors, or loss of special characters, such as Greek letters or mathematical symbols being misrecognized as garbled text, or hyphens in author names being omitted. Future research should focus on optimizing the PDF parsing step or employing vision-capable LLMs to directly process raw PDF images, alleviating information loss during text conversion.

For open fields, all models performed poorly. Two main factors contribute: first, different models express the same medical concept differently; despite prompts constraining the output, LLMs cannot fully satisfy all constraints as requirements expand. Second, model-generated content tends to be verbose, including additional modifiers or explanatory text, preventing exact matching.

Additionally, in numeric parameter extraction, sample-size-related errors primarily originate from confusion between the sample size at randomization and the sample size used in analysis. When both values appear in the same paper, the model, despite prompt guidance, does not consistently select the former. As shown in Figure 4, MAPE for pulses per session varied widely (4.66% – 39.65%). Error analysis indicates that models sometimes confuse total pulses with pulses per session, leading to order-of-magnitude deviations. DeepSeek-V4-Flash exhibited the highest error in male percentage (MAPE=22.01%) while performing robustly in other fields, suggesting that different models have varying capabilities across numeric data types; model selection should consider task-specific characteristics.

From Table 3, model performance varied across RoB2 domains, with best performance on outcome measurement and worst on missing outcome data. This disparity may arise because models reason more effectively on judgments that can be made from a single chunk. For example, randomization methods are often explicitly stated in the methods section, allowing direct evidence retrieval and higher accuracy. In contrast, assessing missing outcome data requires consulting CONSORT flow diagrams, statistical analysis plans, and discussion sections; models are limited in integrating information across multiple chunks to answer whether authors transparently reported and properly handled missing data. Furthermore, semantic similarity scores for rationales and evidence were similar across models, indicating that models can locate original evidence at correct positions, but their internal evaluation criteria differ, leading to varying accuracy in risk level classification.

The ablation study for the retrieval mechanism shows that the three-modality weighted hybrid retrieval strategy significantly outperforms any single modality in coincidence rate. Although semantic retrieval performed poorly in coincidence rate due to excessive document returns, its Top-25 accuracy reached 0.63, validating the ranking quality advantage of dense retrieval. It can be seen that the proposed hybrid retrieval strategy uses keyword retrieval as the core, supplemented by structured data for detailed filtering, and employs semantic matching for similarity ranking.

Finally, compared to the original forest plots, some effect sizes differed. This is because some supporting papers provided only baseline and follow-up values rather than change values, requiring estimation of SD using imputed correlation coefficients. Data missingness also highlights a limitation of the current method: when key data are presented as images, text-based LLMs cannot extract them, underscoring the necessity of incorporating multimodal LLMs.

List of abbreviations

LLM	Large language model
RAG	Retrieval-Augmented Generation
GraphRAG	Graph Retrieval-Augmented Generation
TMS	Transcranial magnetic stimulation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
BLEU-4	Bilingual Evaluation Understudy-4
EM	Exact Match

Conclusion

This paper proposes a GraphRAG-inspired automated extraction and retrieval method for clinical trial literature, using papers as entity nodes. Through LLM-driven semantic chunking and prompt engineering, multi-type parameter extraction is achieved, and a hybrid storage and retrieval mechanism integrating keywords, structured data, and semantic vectors is constructed. The method effectively supports data extraction and retrieval tasks for meta-analysis, providing a feasible pathway toward automated meta-analysis.

Supplementary Information

Acknowledgements

The authors would like to thank Huichao Xie from Northwestern Polytechnical University for his valuable suggestions and insightful feedback during the implementation of this project.

Authors' contributions

YZ conceived and designed the study, collected the data, performed the analyses and led the manuscript drafting and writing. JJ and TY provided critical revisions for intellectual content. All the authors contributed to the interpretation of the findings, approved the final manuscript version, and agreed to be accountable for the accuracy and integrity of the work.

Funding

This research was funded by the National Natural Science Foundation of China under Grant 32330059 and the Tianjin Science and Technology Plan Project under Grant 24ZYCGYS00920.

Data availability

The source code and dataset used in this study are available at https://github.com/Zongyv/TMS-GraphRAG.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

References

[1] Morriss R, Briley P M, Webster L, et al. Connectivity-guided intermittent theta burst versus repetitive transcranial magnetic stimulation for treatment-resistant depression: a randomized controlled trial. Nature Medicine, 2024, 30(2): 403-413.

[2] Cheng S, Xin R, Zhao Y, et al. Evaluation of fMRI activation in post-stroke patients with movement disorders after repetitive transcranial magnetic stimulation: a scoping review. Frontiers in Neurology, 2023, 14: 1192545.

[3] Gogulski J, Ross J M, Talbot A, et al. Personalized repetitive transcranial magnetic stimulation for depression. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2023, 8(4): 351-360.

[4] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials, 1986, 7(3): 177-188.

[5] Riaz I B, Naqvi S A A, Hasan B, et al. Future of evidence synthesis: automated, living, and interactive systematic reviews and meta-analyses. Mayo Clinic Proceedings: Digital Health, 2024, 2(3): 361-365.

[6] Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 2024, 25(1): bbad493.

[7] Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof‐of‐concept study. Research synthesis methods, 2024, 15(4): 576-589.
[8] Ayers J W, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine, 2023, 183(6): 589-596.
[9] Li J, Zhou Z, Lyu H, et al. Large language models-powered clinical decision support: enhancing or replacing human expertise?. Intelligent Medicine, 2025, 5(1): 1-4.
[10] Qiu J, Yuan W, Lam K. The application of multimodal large language models in medicine. The Lancet Regional Health–Western Pacific, 2024, 45.
[11] Yang T, Xiao Y, Bao Z, et al. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. Briefings in Bioinformatics, 2025, 26(6): bbaf601.
[12] Zhou C, Chai H, Chen W, et al. Externalization in llm agents: A unified review of memory, skills, protocols and harness engineering. arXiv preprint arXiv:2604.08224, 2026.
[13] Tan Y, Zhang Z, Li M, et al. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. Computers in biology and medicine, 2024, 172: 108290.
[14] Yuan D, Rastogi E, Naik G, et al. A continued pretrained llm approach for automatic medical note generation. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). 2024: 565-571.
[15] Luo Y, Shi L, Li Y, et al. From intention to implementation: automating biomedical research via LLMs. Science China Information Sciences, 2025, 68(7): 170105.
[16] Yang H, Chen H, Guo H, et al. Llm-medqa: Enhancing medical question answering through case studies in large language models. /2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025: 1-8.
[17] 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, 1-48.
[18] Cai X, Geng Y, Du Y, et al. Utilizing large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. BMC Medical Research Methodology, 2025, 25(1): 116.
[19] Reason T, Benbow E, Langham J, et al. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. PharmacoEconomics-Open, 2024, 8(2): 205-220.
[20] Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. International Journal of Medical Sciences, 2025, 22(11): 2792.
[21] Anisuzzaman D M, Malins J G, Friedman P A, et al. Fine-tuning large language models for specialized use cases. Mayo Clinic Proceedings: Digital Health, 2025, 3(1): 100184.
[22] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 2020, 33: 9459-9474.
[23] Jang D, Shangguan Z, Tegtmeyer K, et al. MedTutor: A Retrieval-Augmented LLM System for Case-Based Medical Education. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2025: 319-353.
[24] Chen Z, Liu T, Mo Y, et al. MedRAGent: An Automatic Literature Retrieval and Screening System Utilizing Large Language Models with Retrieval-Augmented Generation. medRxiv, 2025:

2025.09. 18.25335860.

[25] Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. Journal of the American Medical Informatics Association, 2024, 31(9): 2030-2039.

[26] Xu L, Xie H, Tao X, et al. Personalized rTMS treatment recommendation with retrieval-augmented LLM reasoning. Brain Informatics, 2025, 12(1): 29.

[27] Xu L, Xie H, Qin S J, et al. AI-Powered Repetitive Transcranial Magnetic Stimulation Personalized Treatment Recommendation for Depression Based on Patient Profiles. IEEE Transactions on Computational Social Systems, 2026.

[28] Liu S, McCoy A B, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. Journal of the American Medical Informatics Association, 2025, 32(4): 605-615.

[29] Sun M, Zhao S, Chen J, et al. META-RAG: Meta-Analysis-Inspired Evidence-Re-Ranking Method for Retrieval-Augmented Generation in Evidence-Based Medicine. arXiv preprint arXiv:2510.24003, 2025.

[30] Shen J, Luo Z, Jia D, et al. Large language model enhanced framework for systematic reviews and meta-analyses. BMJ Digital Health & AI, 2025, 1(1).

[31] Tan Z, D'souza J. Diagnosing structural failures in LLM-based evidence extraction for meta-analysis. arXiv preprint arXiv:2602.10881, 2026.

[32] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts. Transactions of the association for computational linguistics, 2024, 12: 157-173.