META-RAG: Meta-Analysis Inspired Re-Ranking for RAG in Evidence-Based Medicine

Anonymous ACL submission

Abstract

Evidence-based medicine (EBM) holds a crucial role in clinical application. Doctors can reduce diagnostic errors by integrating high-quality evidence. Moreover, large language models (LLMs) based methods like RAG can make EBM tasks more efficient. However, RAG applications retrieve irrelevant or conflicting evidence and struggle to validate. This will increase the risk of incorrect clinical decisions. Therefore, inspired by the meta-analysis, we provide a new method to re-rank and filter the medical evidence. We employ a hybrid re-ranking pipeline called Meta-RAG, which includes reliability analysis, heterogeneity analysis, and extrapolation analysis, inspired by the meta-analysis. Meta-RAG can filter and re-rank medical evidence in a training-free manner to meet clinical needs. For evaluation, We test META-RAG with three baselines on multiple datasets in open-domain and multiple-choice clinical QA tasks. The experimental results show there is a stable improvement on quality of evidence and answer accuracy across models of different types and sizes. Meta-RAG also effectively enables RAG to extract more consistent and more patient-specific.

1 Introduction

Evidence-Based Medicine (EBM) is gradually being embraced by doctors as an essential discipline in the medical field (Subbiah, 2023). Using EBM can significantly reduce the risk of misdiagnosis by referring to the retrieved medical articles. As the volume of medical evidence grows, doctors start to rely on artificial intelligence (AI) technology to assist in the practice of EBM (Djulbegovic and Guyatt, 2017). The key requirement from AI is to leverage all available resources, extracting and synthesizing all relevant evidence to arrive at a comprehensive conclusion (Clusmann et al., 2023). However, due to the limitation of memory capacity, small-scale models often struggle to deal with a large amount of evidence (Friedman et al., 2013; Nadkarni et al., 2011). Recently, Large Language Models (LLMs) have been presented, which are equipped with a long input restriction and exceptional comprehension ability. There have been breakthroughs in using LLMs to assist EBM.

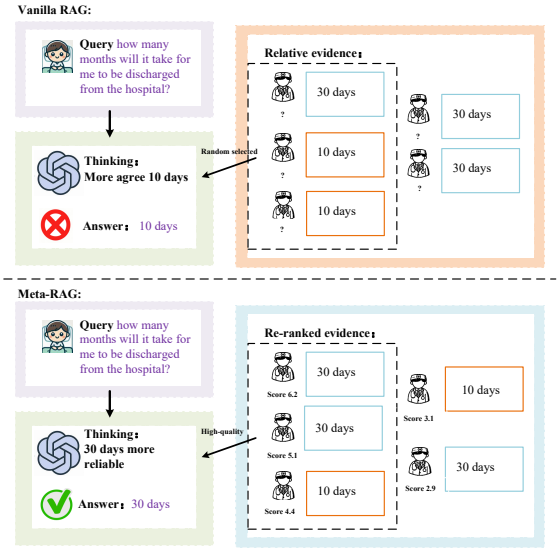

Figure 1: When traditional RAG processes a query, it probably retrieves evidence including conditional results and outdated conclusions. This will mislead the generator to mistakes.

With the iterative advancements in LLM technology, innovative methods like Retrieval-augmented Generation (RAG) and knowledge fine-tuning have emerged (Alam et al., 2023). They can minimize the knowledge errors made by LLMs (Zhang et al., 2023; Huang et al., 2023). The core process of RAG, which involves retrieving evidence and generating diagnoses, closely aligns with the fundamental principles of EBM. As a result, RAG has the most potential to enhance the efficiency of EBM. However, RAG faces several limitations when applied to clinical medicine. EBM requires a highly rigorous process for selecting and filtering the re-

trieved evidence (Sackett et al., 2008). Traditional RAG fails to adequately address this process because of the complexity of medical articles. This oversight often leads to the retrieval of conflicting and redundant evidence. For instance, as illustrated in Figure 1, the vanilla RAG probably retrieves a large volume of unhelpful and non-reliable evidence. This evidence may include conditional results and outdated conclusions. Consequently, RAG selects this evidence to mislead the response, which will significantly restrict the accuracy.

To address the above issues, we develop META-RAG for evidence re-ranking and filtering in RAG for EBM. By acquiring more reliable and valid evidence, this method enables RAG to retrieve evidence that is both more trustworthy and consistent, thereby reducing erroneous judgments. We emulate the principles of meta-analysis, which focuses on three key aspects: (1) reliability, (2) heterogeneity, and (3) extrapolation (Lipsey, 2001; Egger et al., 1997; Hansen et al., 2022). META-RAG filters out inconsistent evidence and presents reliable and rigorous evidence to the response model. As shown in Figure 2, first, we gather the related medical articles and assign a base score to each article based on its publication type. Then, we analyze the evidence and compute the reliability score. We filter the heterogeneous articles and evaluate the extrapolation parameters. Finally, the reliable and high-quality articles are selected and passed to the generator. We present the experiments and results to prove our method effectively resolves the issues of low-quality and conflicting evidence.

Our contributions can be summarized in three aspects:

- We propose a meta-analysis inspired re-ranking pipeline for RAG, which evaluate the evidence in reliability, heterogeneity and extrapolation.
- We propose LLM-based heterogeneity analysis and extrapolation analysis for selecting the optimal medical evidence set using.
- We conduct an evaluation method for the quality of evidence in multiple-choice question task. This method can clearly highlight the improvement in evidence quality brought by re-ranking.

2 Related Works

2.1 EBM and Meta-Analysis

EBM aims to make the best clinical decisions by integrating the best research evidence, clinical expertise, and patient preferences (Subbiah, 2023; McMurray and Packer, 2021). However, doctors can not trust AI models because of hallucinations. They would like to choose the time-consuming and subjective manual approach unless the LLMs (Li et al., 2024). To eliminate biases arising from subjective choices, researchers propose the method known as meta-analysis. Meta-analysis is a quantitative research technique designed to systematically integrate the results of multiple independent studies to provide more rigorous conclusions. It is widely used in fields like medicine, social science, and education, especially in studies derived from experiments (Borenstein et al., 2021). In meta-analysis, researchers aggregate data from multiple independent studies and conduct uniform statistical analyzes to determine overall effect sizes or other relevant statistical metrics (Hansen et al., 2022). However, each meta-analysis requires manually compiling more relevant literature, which is highly complex. Therefore, we hope to utilize the core comparative elements of meta-analysis and employ LLMs to assist users in evaluating evidence.

2.2 RAG in EBM

LLMs have recently made significant progress in natural language processing. High-performance models like GPT-4o (Achiam et al., 2023) have achieved substantial breakthroughs in fields such as medicine, military, and law. Google MED-PALM (Singhal et al., 2023) suggests that LLMs can be applied in many tasks within clinical. With RAG method, the LLMs can deal with these complex tasks with few hallucinations (Lewis et al., 2020). The principle of EBM, which relies on extensive medical evidence for decision-making, aligns well with this approach. RAG generative method is particularly well-suited for EBM and serves as an effective tool for assisting doctors in resolving clinical issues.

However, medicine constantly evolves at a rapid pace, leading to inconsistencies in viewpoints among publications like the articles in PubMed (White, 2020). RAG may retrieve outdated, incorrect, and restricted theories. They may have once been accepted but no longer correct because of the proposal of a new theory. This phe-

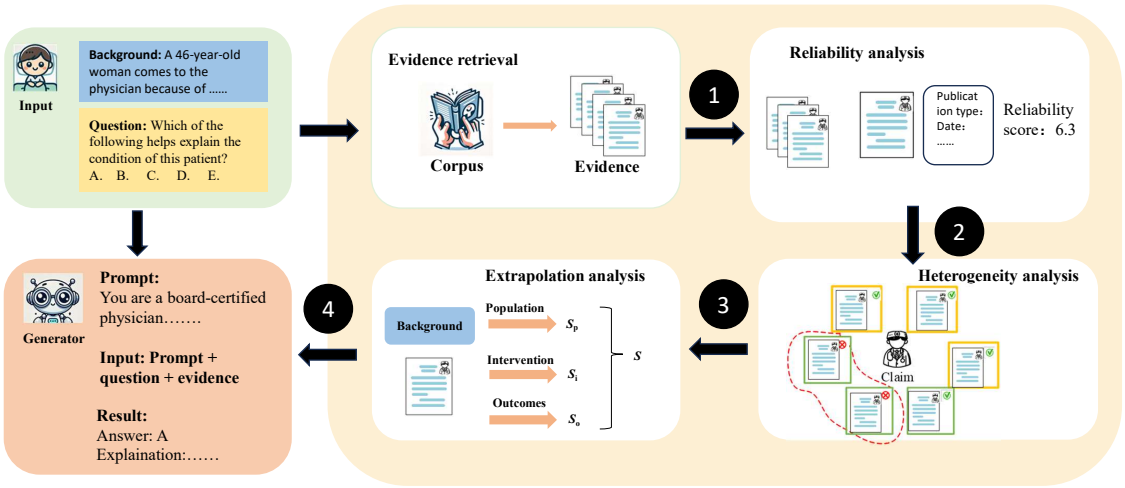

Figure 2: The pipeline of META-RAG includes (1) reliability analysis, (2) heterogeneity analysis, and (3) extrapolation analysis. Our method incorporates these three stages to re-rank and filter evidence, providing as high-quality evidence as possible to (4) generator LLM.

nomenon will result in some conclusions being inapplicable to the actual situation.

2.3 Evidence Re-Ranking

Currently, there are three main methods for optimizing the evidence retrieved during the RAG process: scoring based on similarity, training models, and LLMs that have re-ranking capabilities (Wang et al., 2025; Movin and Hauff, 2025; Gao et al., 2023). However, in medical settings, evidence ranking is too complex for the first two approaches. Clear supervision on evidence order is also unavailable for training. In addition, these two approaches lack interpretability. Their ranking results are therefore hard to justify to users. For the third method, using inter-agent deliberation to perform evidence self-ranking or self-reordering is an efficient zero-shot approach (Miao et al., 2023; Zhai et al., 2025; Ma et al., 2023). Inter-agent collaboration can indeed prioritize important evidence. However, a simple reordering scheme does not clearly indicate how the model should evaluate evidence. We therefore propose META-RAG.

3 Method

3.1 Task Definition

To align with the principles of EBM, we aim not only to deliver convincing answers but also to present high-quality evidence. We define medical queries Q from users as system inputs and then respond A and retrieved evidence E as the output. As shown in Figure 2, our main pipeline focuses on the re-ranking and filtering steps of the evidence in RAG. At the end of the re-ranking and filtering section, we pass high-quality articles with their orders to the generator. In this task, we evaluate the evidence across three distinct dimensions: reliability analysis, heterogeneity analysis, and extrapolation analysis. These analyses enable us to assess the reliability of the evidence, exclude untrustworthy findings, and determine whether the results can be applied to the patient. After re-ranking evidence, the most effective pieces of evidence and their order are passed to the response model to generate recommendations for the queries.

3.2 Evidence Retrieval

In the first step, we construct evidence set by dense retrieval based on semantic embeddings. However, there are some article types lacking key fields such as abstracts in PubMed (White, 2020). To address these problems, we employ a hybrid retrieval approach. We simultaneously search the article titles, abstracts, and MeSH (Medical Subject Headings) keys in the articles. By calculating and aggregating the similarity scores across these three different tags and ranking them, we ultimately select the evidence set E with the each query.

3.3 Reliability Analysis

After obtaining highly relevant evidence, we first grade the articles by their background information. As shown in Figure 3, we mainly score the evidence E with the rules of the publication type, publication date, and LLM judgments.

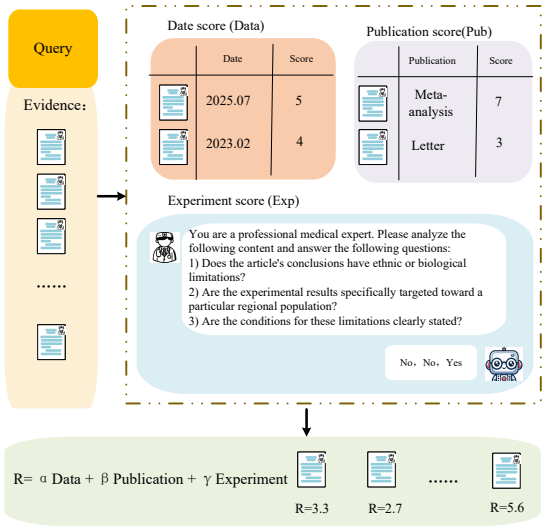

Figure 3: The pipeline of the reliability analysis. We synthesize the information and the judgments of LLM to show the reliability of each evidence.

Initially, we assign base scores based on the publication type and publication date (Polit and Beck, 2004). We categorize all evidence types into seven levels based on the evidence hierarchy in evidence-based medicine, and we assign scores according to these levels. Also, we set additional scores for the articles by their publication dates.

We also employ an LLM for a more fine-grained reliability analysis. Meta-analyses typically analyze the randomization of literature, data integrity, presence of bias, and choices regarding blinding. These principles can reflect the validity of the experimental conclusions in the article. We implement this method, evaluating the evidence by three questions as detailed in Figure 3. The detailed architecture of prompts and questions is provided in the appendix. Ultimately, we sum three types of scores to provides the reliability score r_i with E_i. A larger r_i signify a more rigorous methodology.

3.4 Heterogeneity Analysis

After we score the evidence on reliability, we filter the evidence and enforce stance consistency. We apply an evidence filtering method inspired by heterogeneity analysis (Higgins and Thompson, 2002). This analysis can remove studies with low quality and high heterogeneity. This step guarantees coherent evidence fed to the generative model.

In this step, we apply the definition of heterogeneity in the DerSimonian-Laird method (DerSimonian and Laird, 2015). The inputs are article-claim pairs. Based on the characteristics of datasets, we approximate part of the model parameters and define the measurement metric to represent the stance of each article.

First, we create claims by combining the query with each option. Each option defines a separate claim. We ask LLMs to determine the stance of each piece of evidence on each claim. We define the label of each evidence as y_i, and mark these pairs as support, oppose, or irrelevant.

$$y_i = \begin{cases} 1, & \text{if } i \text{ labeled “Support”}, \\ 0, & \text{if } i \text{ labeled “Oppose”}, \\ \text{NaN}, & \text{if } i \text{ labeled “Irrelevant”}. \end{cases} \quad (1)$$

Second we need to compute the heterogeneity of the evidence set associated with each query. We define k as the total number of studies retrieved for a single query and v_i as the variance estimate of the i^{th} study. We set the random-effects variance τ^2_{DL} as the stance divergence of evidence group. To calculate this, we should gather the pooled effect θ_{RE}, and the study weights w_i at this step. However, most original studies do not report standard errors in the abstract. Therefore, it is not feasible to compute these two quantities using the standard meta-analysis pipeline. We assign an article-level weight using the reliability score r_i obtained in the previous step. More reliable articles receive higher weights. Then we can get the fixed-effect $\hat{\theta}_{\text{FE}}$. Formally,

$$\hat{\theta}_{\text{FE}} = \frac{\sum_{i=1}^{k} r_i\, y_i}{\sum_{i=1}^{k} r_i} \quad (2)$$

Then we calculate the heterogeneity statistic Q. This variable represents the total standard deviation of the entire set of articles. It serves as a preliminary indicator of the consistency of stances within the article cluster.

$$Q = \sum_{i=1}^{k} r_i \left(y_i - \hat{\theta}_{\text{FE}}\right)^2 \quad (3)$$

Therefore, we get the heterogeneity τ^2_{DL} as follows. We compare the observed between-study dispersion Q with the dispersion expected under sampling error alone $(k-1)$. If Q is larger, the excess is converted into the between-study variance after weight adjustment; otherwise, we assume no heterogeneity ($\tau^{2\,(i)}_{\text{DL}}$=0). Formally,

$$\tau^2_{\text{DL}} = \max\left\{ \frac{Q-(k-1)}{\sum_{i=1}^{k} r_i - \frac{\sum_{i=1}^{k} r_i^2}{\sum_{i=1}^{k} r_i}},\ 0 \right\} \quad (4)$$

Finally, we define S as the set of all k studies and $S^{(-i)}$ as the set obtained by removing study i from S. As for the formula 4, we compute the $\tau^{2\,(-i)}_{\mathrm{DL}}$ for the $S^{(-i)}$ and calculate the decrease caused by this evidence. We define an acceptable maximum heterogeneity contribution M and a minimum reliability score R_c. Based on the final outcomes, we determine whether each article should be excluded. Formally,

$$\Delta_i = \frac{\tau^2_{\mathrm{DL}} - \tau^{2\,(-i)}_{\mathrm{DL}}}{\tau^2_{\mathrm{DL}}} \tag{5}$$

Algorithm 1 summarizes this process.

Algorithm 1 Heterogeneity Analysis

Input: Query q, Evidence $(E, R) = \{(E_1, r_1), \dots (E_k, r_k)\}$, Hyperparameter M, R_c
Output: filtered evidence $E_f = \{E_1, \dots E_m\}$

$v_i \leftarrow \sigma$ ▷ Initialize
$c_1, c_2, \dots \leftarrow \mathcal{C}(q)$ ▷ Combine claims
$E_f \leftarrow \{\}$
for $c_i \in \mathcal{C}(q)$ **do**
  $y \leftarrow \mathcal{G}(\mathcal{P}_0(c, E_i))$ ▷ Generate evidence labels
  $\tau^2_{\mathrm{DL}} \leftarrow \max D(q, y, v_i, k)$ ▷ Calculate DL variance
  for $e \in \{E_i \mid i = 0 \dots k\}$ **do**
    Compute $Q_i = \mathcal{M}(y_i, v_i)$
    Compute Δ_i by Eq. (5)
    if $\Delta_i < M \ \wedge \ e \notin E_f$ **then**
      Add e to E_f
    end if
    if $\Delta_i \geq M \ \wedge \ e \notin E_f \ \wedge \ r_i > R_c$ **then**
      Add e to E_f
    end if
  end for
end for
return E_f ▷ Return the filtered evidence

3.5 Extrapolation Analysis

To analysis the gaps between the users' statics and the experimental conditions in the evidence, we design an extrapolation parameter for each evidence. This parameter is calculated based on three steps. First, we split the query into background and question.

Then, we use LLM with a carefully designed prompt to compare the background information and the evidence across the population, intervention, and outcomes (Methley et al., 2014). Each piece of evidence is assigned a fine-grained score along each of these dimensions, and the detailed architecture of this process is provided in the appendix. Formally,

$$T_j \leftarrow \alpha T_p + \beta T_i + \gamma T_o \tag{6}$$

Finally, we compute an overall extrapolation score for each evidence relative to the user background. We calculate the final ranking score S by both the extrapolation score and the reliability score. Formally,

$$S = \sum_{j=1}^{k} r_j^2 T_j \tag{7}$$

Algorithm 2 summarizes this process.

Algorithm 2 Extrapolation Analysis

Input: Query q, Evidence $E_f, R_f = \{(E_1, r_1), \dots (E_f, r_f)\}$, Hyperparameter α, β, γ
Output: Scored evidence $(E_f, S) = \{(E_1, S_1), \dots (E_m, S_m)\}$

$S \leftarrow \{\}$ ▷ Initialize
$Back, Que \leftarrow \mathcal{C}(q)$ ▷ split the background
for $e \in \{E_j \mid j = 0 \dots m\}$ **do**
  $T_p \leftarrow \mathcal{G}(\mathcal{P}_0(Back, E_j))$ ▷ Generate Population score
  $T_i \leftarrow \mathcal{G}(\mathcal{P}_1(Back, E_j))$ ▷ Generate Intervention score
  $T_o \leftarrow \mathcal{G}(\mathcal{P}_2(Back, E_j))$ ▷ Generate outcome score
  $T_j \leftarrow \alpha T_p + \beta T_i + \gamma T_o$ ▷ Calculate Extrapolation score
  $S_j \leftarrow r_j^2 T_j$ ▷ Calculate total ranking score
end for
return (E_f, S) ▷ Return the filtered evidence

4 Experiments and Results

4.1 Experimental Setup

Datasets In our experiments, we select and extract three datasets: **Asclepius** (Kweon et al., 2023): Asclepius is a publicly available clinical instruction dataset built from synthetic clinical notes and QA-style instructions for training and evaluating medical LLMs. The dataset provides comprehensive patient information, and both the questions and gold answers are generated by GPT-3.5.(n=2000) **MedQA** (Jin et al., 2020): MedQA is a medical multiple-choice question(MCQ) benchmark

	Method	D = 1				D = 2			
		Soft-Recall	Soft-F1	SemSim	LLM-E	Soft-Recall	Soft-F1	SemSim	LLM-E
Llama-3.0-8B	Meta	0.765	0.654	0.780	0.463	0.760	0.652	0.771	0.473
	w/o Evi	0.701	0.649	0.758	0.428	–	–	–	–
	Sim-Evi	0.697	0.652	0.756	0.375	0.694	0.650	0.755	0.385
	Self-Evi	0.677	0.625	0.736	0.342	0.677	0.633	0.738	0.340
Qwen2.5-7B	Meta	0.782	0.569	0.785	0.524	0.760	0.652	0.771	0.324
	w/o Evi	0.701	0.649	0.758	0.375	–	–	–	–
	Sim-Evi	0.697	0.652	0.756	0.36	0.694	0.650	0.755	0.380
	Self-Evi	0.748	0.637	0.790	0.449	0.748	0.637	0.787	0.466
Qwen2.5-14B	Meta	0.803	0.562	0.805	0.790	0.760	0.652	0.771	0.604
	w/o Evi	0.787	0.623	0.809	0.832	–	–	–	–
	Sim-Evi	0.786	0.617	0.804	0.880	0.786	0.621	0.804	0.861
	Self-Evi	0.684	0.632	0.747	0.614	0.674	0.630	0.733	0.633

Table 1: Performance of Meta-RAG and baselines on 2000 Asclepius queries. D denotes the number of evidence articles provided during generation. For the w/o Evi setting, the number of evidence items is independent of D. The model generates responses without any evidence. Detailed evaluation setup is discussed in Section 4.2.

derived from USMLE examinations, designed to test clinical knowledge and reasoning.(n=5000) **MMLU** (He et al., 2019): From MMLU, we use only the MMLU-clinical subset. This is also a MCQ dataset. Since the questions are largely factual, it serves as a good test of whether the evidence can locate the right piece of knowledge rather than rely on complex reasoning.(n=300)

Evidence We take the PubMed (White, 2020) as the evidence database. This dataset provides a thorough organization of information from the literature. For each query, 15 articles are initially retrieved. In the step of reliability analysis, we also divide these articles into different levels by other rules and LLM shown in Figure 4. We then categorize and rank studies based on their PubMed publication types. This procedure yields seven distinct levels.

Baselines We select three different baselines to compare the performance of the META: **w/o Evi**: To test the base performance of each model, we give no evidence to the LLM as w/o Evi. **Sim-Evi**: We provide the similarity based order of evidence as the Sim-Evi. We use the output after the retrieval directly serving as the most straightforward control group for our method. **Self-Evi**: We provide evidence extracted based on the LLM as Self-Evi. This baseline is designed to demonstrate that our method offers a significant improvement over a straightforward LLM-based approach.

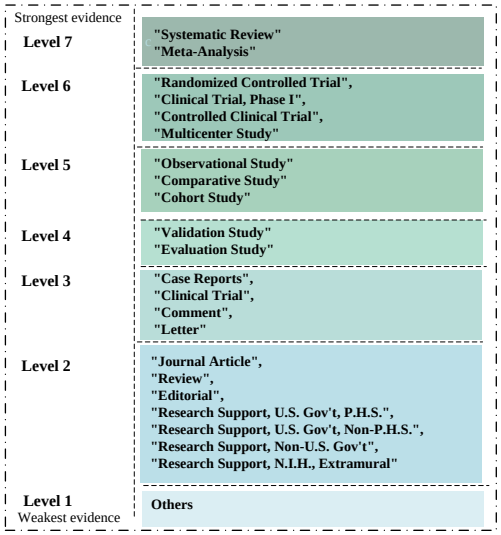

Figure 4: We divide the evidence type into 7 levels. In reliability analysis, we categorize evidence from different publication types and LLM judgments. The higher level of evidence means a better publication type score.

4.2 Results on Open-Domain QA Datasets

As shown in Table 1, we report the performance of our method on the Asclepius dataset. The queries are too long for many models to produce valid outputs. We therefore select the models reported in the table. Also, we utilize following four evaluation metrics to demonstrate the improvements of META-RAG. **Soft-Recall**: Sentence-level semantic coverage of the gold answer by the prediction, computed using SBERT cosine similarity. **Soft-F1**: Sentence-level semantic F1 score that summarizes how well the prediction semantically

	Method	MedQA					MMLU				
		D=1	D=2	D=3	D=4	Best	D=1	D=2	D=3	D=4	Best
Llama-3.0-8B	Meta	**44.0**	38.0	40.7	39.3	**44.0**	**42.7**	42.0	42.0	39.3	**42.7**
	w/o Evi	–	–	–	–	38.7	–	–	–	–	36.2
	Sim-Evi	25.3	29.3	31.9	32.6	32.6	25.3	29.3	31.9	32.6	32.6
	Self-Evi	38.0	30.0	33.3	28.3	38.0	36.0	40.0	42.7	37.2	42.7
Qwen2.5-7B	Meta	51.5	**52.0**	48.5	42.5	**52.0**	49.3	46.0	**50.7**	46.7	**50.7**
	w/o Evi	–	–	–	–	49.6	–	–	–	–	49.3
	Sim-Evi	44.5	43.5	42.5	43.5	44.5	43.3	43.7	48.4	44.3	48.4
	Self-Evi	42.5	39.5	43.5	41.5	41.5	48.0	48.7	48.0	48.4	48.7
Mistral-7B	Meta	**47.5**	45.0	46.5	46.5	**47.5**	45.0	47.3	**48.0**	47.7	**48.0**
	w/o Evi	–	–	–	–	43.5	–	–	–	–	44.0
	Sim-Evi	42.0	42.5	40.5	45.5	45.5	43.3	45.0	46.8	45.4	46.8
	Self-Evi	42.5	39.5	43.5	41.5	43.5	43.3	44.7	45.3	46.7	46.7
Gemma-1.1-7B	Meta	41.0	41.5	**43.0**	40.0	**43.0**	36.0	34.7	35.3	**40.0**	**40.0**
	w/o Evi	–	–	–	–	40.5	–	–	–	–	34.7
	Sim-Evi	34.0	31.5	30.0	31.0	34.0	35.3	36.6	35.5	37.1	37.1
	Self-Evi	31.0	29.5	30.0	31.0	31.0	34.7	29.3	33.3	34.7	34.7

Table 2: Accuracy (%) of Meta-RAG and baselines on MedQA and MMLU. D denotes the number of evidence articles provided during generation. Best reports the highest score among D=1–4 for each setting.

matches and covers the gold answer. **SemSim**: Global semantic similarity between the full prediction and the full gold answer. SemSim captures global semantic closeness, whereas Soft-F1 measures sentence-level semantic coverage. **LLM-E**: LLM-Evaluation label in {correct, partial, incorrect} given the question, gold answer, and prediction. The results in the table are strict accuracies computed using only the correct option.

4.3 Results on MCQ Datasets

Table 2 shows the performance of different types of LLMs, while Table 3 shows the performance of different sizes. We require the model to output the selected option letter first. If the response does not follow this format, the model is allowed up to five regeneration attempts. We then compute accuracy based on the final selected answers.

4.4 Overall Analysis

Our method consistently improves upon all the baselines of almost all LLMs. Across LLMs of different sizes and types, our Meta-RAG achieves substantial improvements over traditional evidence-ranking methods. In open-domain QA datasets, we find that the exact magnitude of improvement is hard to quantify. However, META-RAG markedly increases the recall of the model's outputs. This pattern indicates that our method provides more comprehensive information, and the model incorporates it into its responses. In MCQ datasets, we find that META-RAG yields a stable improvement in answer accuracy across settings. The best-performing model, Llama-3-8B, achieves up to an 11.4% gain on the MedQA dataset.

4.5 Analysis on Evidence Numbers

The performance floats by different evidence budgets have multiple causes. Some models show substantial prompt and question forgetting when given long evidence. Large evidence set becomes a heavy burden for these models. In addition, when the task is not difficult for a given model, excessive evidence can constrain its reasoning and reduce its headroom. These effects explain why accuracy can drop in some settings even when more evidence is provided.

4.6 Analysis on LLM Sizes

As shown in Table 3, our method delivers steadily increasing performance on stronger models. Smaller models suffer significant drops when exposed to too many input tokens. However, for Qwen-14B and Qwen-32B, adding more evidence consistently improves accuracy, indicating that our evidence has low heterogeneity.

4.7 Ablation Study

To validate the efficiency of each step, we set multiple ablation experiments as shown in Table 4.

w/o Reliability: We set all the reliability scores same. When calculating the highest-scoring ev-

	Method	MedQA				
		D=1	D=2	D=3	D=4	Best
0.5B	Meta	**25.00**	24.70	23.92	23.88	**25.00**
	w/o	–	–	–	–	24.64
	Sim	23.80	23.50	23.62	23.80	23.80
	Self	23.50	23.78	24.00	23.34	24.00
1.5B	Meta	28.42	30.26	30.56	**30.82**	**30.82**
	w/o	–	–	–	–	**35.08**
	Sim	26.52	28.84	28.12	27.40	28.84
	Self	25.98	28.48	27.94	27.08	28.48
7B	Meta	50.76	**51.04**	50.74	50.56	**51.04**
	w/o	–	–	–	–	50.58
	Sim	47.04	47.08	46.66	47.18	47.18
	Self	49.70	49.12	49.52	49.18	49.70
14B	Meta	59.20	60.00	60.72	**60.90**	**60.90**
	w/o	–	–	–	–	58.36
	Sim	56.58	56.76	56.32	56.94	56.94
	Self	55.48	55.90	55.88	55.66	55.90
32B	Meta	63.28	64.08	64.00	**64.32**	**64.32**
	w/o	–	–	–	–	62.06
	Sim	59.14	60.24	60.30	60.10	60.30
	Self	59.46	60.24	60.10	60.18	60.24

Table 3: Accuracy (%) for different sizes of Qwen-2.5 models on 5,000 MedQA queries. D denotes the number of evidence articles. Best reports the highest score among D=1–4.

idence for heterogeneity analysis, we randomly select the first piece of evidence from the list. Both the quality of the evidence and the accuracy of the responses decrease in this ablation.

Setting	MedQA				
	D=1	D=2	D=3	D=4	Best
Meta	**44.0**	38.0	40.7	39.3	**44.0**
w/o R	36.7	40.0	34.0	37.3	40.0
w/o H	34.0	38.7	37.3	35.3	38.7
w/o E	34.0	33.3	34.7	34.7	34.7

Table 4: Ablation study of Meta on MedQA. D denotes the number of evidence articles. *w/o R*: remove reliability checking by setting all reliability scores to 1. *w/o H*: remove heterogeneity analysis. *w/o E*: disable evidence weighting by setting $T_j = 1$.

w/o Heterogeneity: we remove the heterogeneity judgment process and directly re-rank the extracted evidence based on reliability and extrapolation. We observe a noticeable decline in the evidence contribution score and accuracy. This phenomenon suggests that some highly reliable but paradoxical evidence scores remain in the evidence and mislead the judgment of the generation model.

w/o Extrapolation: We set all the extrapolation parameters to 1. The accuracy decreases quickly. Most queries in MedQA have background restriction. Evidence with wrong population mislead the judge of LLM.

4.8 Is the Evidence Better?

As shown in Figure 5, we employ another evaluation metric to evaluate the evidence quality. We assess the similarity between re-ranked evidence and the gold clam. The higher similarity means better evidence is provided. We observe that the average quality of the evidence is effectively enhanced after Meta-RAG. Additionally, we can also analyze that as the model size grows, the Self-Evi group becomes more sensitive to good evidence. Our experiments show that some models can select higher-quality evidence but they are controversial. As a result, the generative model becomes confused. Therefore, most baselines can not surpass Meta-RAG.

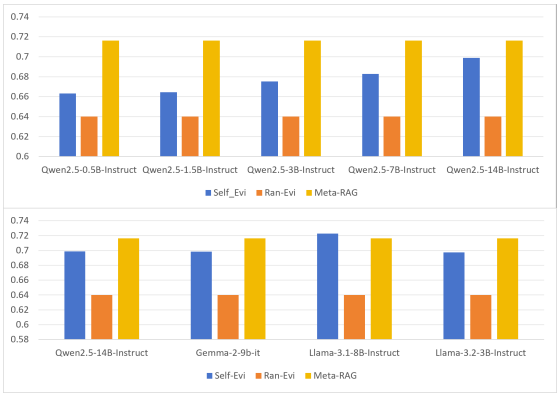

Figure 5: The similarity of each method between the provided evidence and the ground-truth answer. We use this metric to evaluate whether Meta-RAG can better guide the model to the correct answer.

5 Conclusion

EBM currently needs robust automated tools to assist in medical tasks. However, existing RAG for EBM cannot ensure the evidence meets the stringent requirements of medicine. Therefore, inspired by the principles of meta-analysis, we propose a META-RAG filtering and re-ranking method to ensure the evidence is effective and reliable. We conduct practical experiments on our method and verify its improvements in accuracy and evidence quality. We hope this work will assist researchers in the medical field, promoting safer and more effective deployment of LLMs in medical applications.

Limitations of the work

Traditional tasks can't effectively evaluate our evidence-ranking method. Our approach aims to guide the model to adhere to the principles of evidence screening and ranking in EBM, retrieving more effective and reliable evidence. Our method can enhance the acceptance of the screened evidence among the medical community. However, for traditional tasks, the improvement provided by our method is indeed not significant. This phenomenon compels us to add the level of evidence as an evaluation metric. In our subsequent work, we will seek evaluation methods that better reflect the unique characteristics of medical evidence.

Ethical Statement

This work aims to re-rank and filter the evidence to assist doctors in identifying diagnostic errors. We suggest that LLMs can never completely replace doctors. Additionally, all datasets used in this study are derived from publicly available datasets, and there is no leakage of personal privacy or confidential information.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fakhare Alam, Hamed Babaei Giglou, and Khalid Mahmood Malik. 2023. Automated clinical knowledge graph generation framework for evidence based medicine. *Expert Systems with Applications*, 233:120964.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2021. *Introduction to meta-analysis*. John Wiley & Sons.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Rebecca DerSimonian and Nan Laird. 2015. Meta-analysis in clinical trials revisited. *Contemporary clinical trials*, 45:139–145.

Benjamin Djulbegovic and Gordon H Guyatt. 2017. Progress in evidence-based medicine: a quarter century on. *The lancet*, 390(10092):415–423.

Matthias Egger, George Davey Smith, and Andrew N Phillips. 1997. Meta-analysis: principles and procedures. *Bmj*, 315(7121):1533–1537.

Carol Friedman, Thomas C Rindflesch, and Milton Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765–773.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Christopher Hansen, Holger Steinmetz, and Jörn Block. 2022. How to conduct a meta-analysis in eight steps: a practical guide.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making*, 19:91–100.

Julian PT Higgins and Simon G Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2023. Publicly shareable clinical large language model built on synthetic clinical notes. *Preprint*, arXiv:2309.00237.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jin Li, Yiyan Deng, Qi Sun, Junjie Zhu, Yu Tian, Jingsong Li, and Tingting Zhu. 2024. Benchmarking large language models in evidence-based medicine. *IEEE Journal of Biomedical and Health Informatics*.

Mark W Lipsey. 2001. Practical meta-analysis. *Thousand Oaks*.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

John JV McMurray and Milton Packer. 2021. How should we sequence the treatments for heart failure and a reduced ejection fraction? a redefinition of evidence-based medicine. *Circulation*, 143(9):875–877.

Abigail M Methley, Stephen Campbell, Carolyn Chew-Graham, Rosalind McNally, and Sudeh Cheraghi-Sohi. 2014. Pico, picos and spider: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research*, 14(1):1–10.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Maria Movin and Claudia Hauff. 2025. Zero-shot reranking with large language models and precomputed ranking features: Opportunities and limitations.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Denise F Polit and Cheryl Tatano Beck. 2004. *Nursing research: Principles and methods*. Lippincott Williams & Wilkins.

D Sackett, WS Richardson, and WMC Rosenberg. 2008. What is evidence-based medicine (ebm). *Patient care model*, 36:26–33.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Vivek Subbiah. 2023. The next generation of evidence-based medicine. *Nature medicine*, 29(1):49–58.

Pinhuan Wang, Zhiqiu Xia, Chunhua Liao, Feiyi Wang, and Hang Liu. 2025. Realm: Recursive relevance modeling for llm-based document re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23875–23889.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Zhouwei Zhai, Mengxiang Chen, Haoyun Xia, Jin Li, Renquan Zhou, and Min Yang. 2025. Beyond retrieval-ranking: A multi-agent cognitive decision framework for e-commerce search. *arXiv preprint arXiv:2510.20567*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

6 Appendix

6.1 Retrieval

In the first step, we conduct evidence retrieval based on query similarity with the datasets. We aim to retrieve as much relevant evidence as possible. Therefore, we utilize the PubMed database, a vast repository of biomedical and life sciences research articles managed by the U.S. National Library of Medicine. This database primarily consists of academic articles from plenty of publications with extensive information, like mesh heading and the article date.

During the retrieval, we observe that many PubMed entries lack an abstract due to their publication format, so relying solely on abstract-based retrieval is incomplete. Moreover, selecting articles using only title similarity fails to capture other facets of the claim. To address these limitations, we employed three distinct retrieval strategies and then aggregated and deduplicated their results. For each candidate article, we extracted its abstract, title, and keywords, and computed similarity scores between each of these elements and the claim. We then selected the top ten articles per dimension—thirty in total—and deduplicated this set. Finally, we randomly sampled ten articles from the deduplicated pool and added them to the query's document directory as our ten extracted pieces of evidence.

6.2 Reliability Analysis

After obtaining highly relevant evidence, we first grade the articles by their fundamental information. We combine rules and LLMs judgments to score their reliability. Initially, we access the Publication type from the information and evaluate the articles' quality level. We assign scores ranging from 1 to 7 based on the medical principles (Polit and Beck, 2004). Recognizing that the publication date of an article can significantly influence its conclusions, we then sort the articles by their publication dates. We award an extra point to the most recently published articles on their base score. And as the article becomes less recent, the score we reward gradually decreases in tiers. This process results in our base score derived from rule-based filtering.

Subsequently, we employ an LLM for a more fine-grained reliability analysis. Meta-analyses typically analyze the randomization of literature, data integrity, presence of bias, and choices regarding blinding. These four principles can reflect the validity of the experimental conclusions in the article. We implement this method with an LLM, evaluating the evidence from these four aspects as detailed in Figure 4. We give the LLM a prompt to let it judge the article and respond only four letters to answer the questions. For each imperfection identified, we apply a penalty parameter to the article score. Ultimately, this process provides us with a more reliable ranking order.

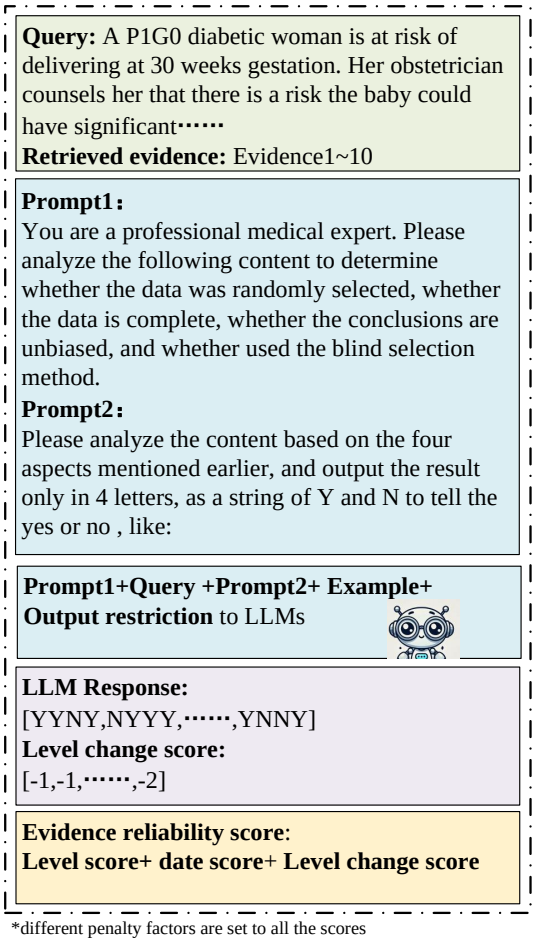

Figure 6: Details of the reliability score counting. This step ensures the articles are ranked by their quality.

6.3 Heterogeneity Analysis

After obtaining the reliability scores, we need a heterogeneity analysis to ensure the content provided to the generation model is closely aligned with the scientific consensus. The traditional meta-analysis need all experiment data to calculate the heterogeneity among all articles. However, we can only access the abstracts of each medical publication, and the raw data of the medical articles is always Closed-source data. Therefore, our approach is to filter out heterogeneous information only based on

statistical reliability scores.

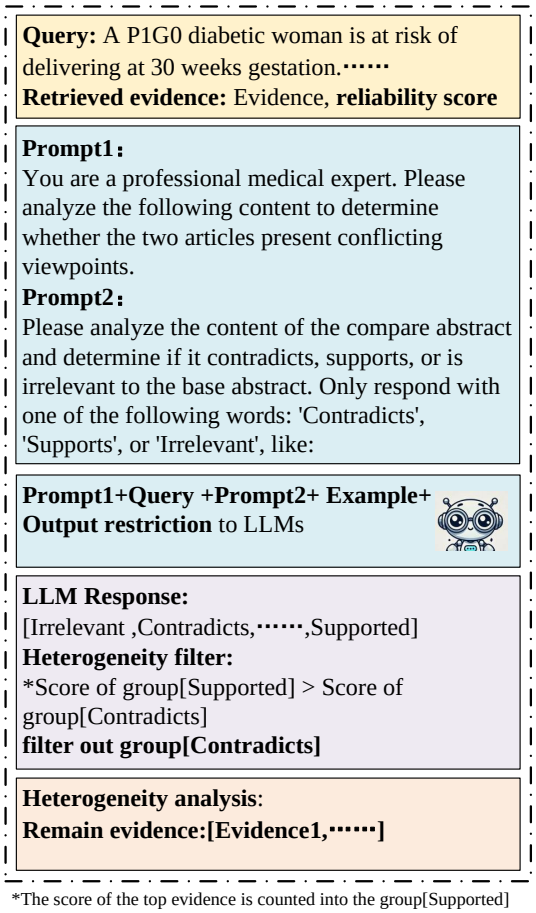

Figure 7: Details of the Heterogeneity analysis. This ensures the consistency of the evidence provided to the generation model.

Our filter method starts with selecting the highest-scoring documents from the reliability analysis as the base articles. We compare the remaining articles to this article to determine if there are supportive, irrelevant, or contradictory relationships. We employ an LLM to categorize all articles accordingly. Ultimately, we compare the total scores of articles that support the baseline against those that oppose it. The lower-scoring group has all its articles removed. This method helps retain a consistent and coherent set of evidence that enhances the quality and relevance of the information fed into the generation model.

6.4 Extrapolation Analysis

In addition to the generalizability analysis used in the main text, we also design an alternative variant that relies solely on LLM assistance for generalizability assessment. For the articles after the heterogeneity analysis, we prompt the LLM to discern whether the conclusions of these articles have any racial or biological limitations and whether the experimental results could cause significant side effects for specific populations as shown in Figure 8. We adjust the scores positively for articles without conditions, negatively for articles with applicable conditions, and neutrally for those with unclear conditions. This adjustment ensures that we obtain

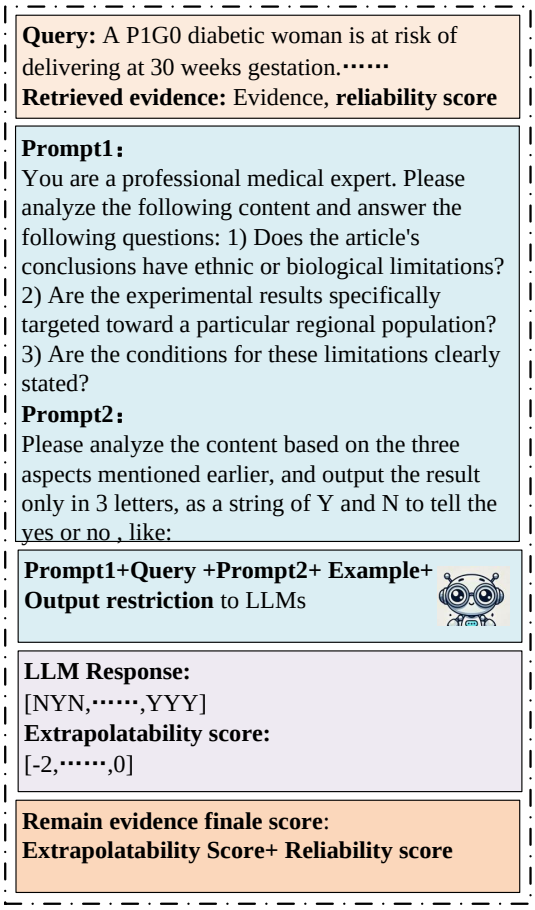

Figure 8: Details of the Extrapolation analysis. This adjustment ensures that we obtain a score that truly represents the quality of the help to the users.

a score that truly represents the quality of the help to the users. After this analysis, we can provide the top-scored evidence to the generation model, resulting in a safe and reliable high-quality response. This process not only enhances the applicability of the responses but also ensures they are tailored to the specific biological context of the user, increasing the accuracy and safety of the medical advice provided.

However, in our empirical experiments, we

find that the large model's extrapolability assessment was overly coarse-grained. As a result, it tend to judge the vast majority of evidence as non-extrapolable, rendering the evaluation meaningless. To address this, we adapt the prompt shown in Figure 9, which provides a more fine-grained criterion for directly assessing the extrapolability of each piece of evidence to the user's background.

6.5 Experiments

Parameters In our experiments, we set the input length of every model to its maximum to ensure effective responses for all baselines and META-RAG. We cap the output at 512 tokens for clarity and readability, and we truncate any excess. To maximize performance, we sweep the temperature over 0.3, 0.5, and 0.7 and report the best result. During retrieval, each computation retrieves at least five papers. We then deduplicate after merging evidence groups and ensure that at least ten papers remain.

Other Evaluation As shown in Table 5, We employ another evaluation metric to evaluate the evidence quality. We assess the contribution of our input evidence to the evidence options as a measure of evidence quality. We observe that after reranking and filtering, the average quality of the evidence is effectively enhanced. Additionally, We also analyze the proportion of articles that have positive evidence relevance. Our method successfully demonstrates that good evidence has been provided for the generation model. The evidence extracted by META-RAG is more closely aligned with the correct answers among the options.

Method	ECS	PPA (%)
Baseline1	N/A	N/A
Baseline2	-1.8924	12.33
Baseline3	-0.8355	12.33
Meta-RAG	**-0.5718**	**12.62**

Table 5: Evaluation of retrieved evidence quality. ECS (Evidence Contribution Score) measures the average similarity between retrieved articles and candidate options. PPA (Proportion of Positive Articles) denotes the ratio of evidence articles that support the correct option. ECS and PPA are not applicable to Baseline1 due to unrelated evidence retrieval.

Also, we test the extrapolatation analysis component of the model. The results mainly reflect whether the evidence can effectively serve the user. While this issue occurs less frequently in medical problems, it is critical in EBM. We include this module to ensure that users can access relevant knowledge more precisely. For LLMs, the influence of this part seems less than the other two. Therefore, as shown in Table 6, we implement a more detailed experiment to test the influence of extrapolatation penalty score. The experimental results show that the model's accuracy is highly sensitive to the penalty coefficient. With an incorrect penalty coefficient, the output accuracy significantly decreases, which indirectly demonstrates the necessity of this step in the process.

EP value	Accuracy (%)	ECS	PPA (%)
-0.2	**46.60**	**-0.5718**	12.62
-0.3	41.75	-0.7156	**14.56**
-0.4	37.86	-0.5800	12.62

Table 6: Ablation study of the extrapolatation penalty (EP) on MedQA. Higher accuracy and PPA indicate better performance, while ECS closer to zero indicates higher evidence contribution quality.

6.6 How about other LLMs

We also evaluated several of today's top-performing large language models. As illustrated in Figure 7, we report GPT-4o's performance on the two datasets used in our study. We observe substantial instability across different baselines. On the MEDQA dataset, GPT-4o achieves its best results when answering directly—without any supporting evidence—whereas on the MMLU dataset, supplying random evidence yields superior performance.

For MEDQA, we hypothesize that GPT-4o was trained on an excessive amount of the same or very similar data, leading it to rely far more heavily on memorized internal evidence than on external inputs. As a result, providing additional evidence actually misleads the model and degrades its performance. By contrast, for MMLU, GPT-4o appears not to have been exposed to equivalent training data; consequently, the direct-answering baseline produces the worst results. In this case, the injected evidence does not constrain the model's reasoning path but rather "awakens" its latent knowledge, enabling it to recall relevant facts more clearly.

Taken together, these findings suggest that for widely studied benchmarks like MEDQA and MMLU, large models already possess substantial internal reserves of knowledge. As a result, apply-

"""
You are a medical evidence applicability evaluator.

Your task is to assess how applicable a piece of clinical evidence is to a specific patient case. Focus your evaluation on three key dimensions:

1. Similarity between the patient and the studied population — including demographics (age, sex, ethnicity), comorbidities, disease stage or subtype, and previous treatment history.

Scoring guidance:
- 0.90–1.00: Nearly identical population; key clinical characteristics closely match.
- 0.70–0.89: Minor differences exist (e.g., slightly different age group, manageable comorbidities).
- 0.40–0.69: Moderate differences in critical variables (e.g., different subtype or significant comorbidities).
- 0.00–0.39: Major mismatch (e.g., different disease type or demographic group).

2. Relevance of the intervention — considering the treatment modality, dosage, route, duration, and feasibility in the patient's setting.

Scoring guidance:
- 0.90–1.00: Intervention is the same or virtually identical; fully feasible for the patient.
- 0.70–0.89: Intervention is similar with minor differences (e.g., dose adjustment).
- 0.40–0.69: Significant modifications are needed or partial incompatibility exists.
- 0.00–0.39: The intervention is not available, feasible, or differs substantially.

3. Alignment of the clinical outcome — focus on whether outcomes in the evidence are clinically meaningful to the patient, distinguishing surrogate vs. direct endpoints.

Scoring guidance:
- 0.90–1.00: Clinical outcomes fully align with patient goals (e.g., mortality, symptom relief).
- 0.70–0.89: Outcomes are partially aligned or secondary but still meaningful.
- 0.40–0.69: Outcomes are surrogate or marginally relevant to the patient.
- 0.00–0.39: Outcomes have little or no relevance to patient care.

Assign scores using a fine-grained scale from 0.00 to 1.00. Be specific in your reasoning, and avoid using templates. Explain what aspects drove each score — such as mismatches in patient features, intervention context, or outcome priorities.

Figure 9: Details of the Extrapolation analysis prompts.

ing our method to such over-provisioned models yields only marginal gains. To fully demonstrate the advantages of our approach, it will be necessary to evaluate on more challenging datasets with less pre-exposure.

Discussion

(1)In fact, existing benchmark tasks are far from real-world clinical settings. Open-domain QA datasets often use model-generated questions and answers. Multiple-choice QA datasets usually adapt items from existing exams. In real clinical encounters, clinicians prefer the most relevant and directly applicable evidence, such as reliable cases or established knowledge. Clinicians do not rely on potentially hallucinated model responses. Evidence utility matters more than answer accuracy. However, current benchmarks offer no dataset that evaluates evidence filtering ability alone.

(2) Fully replicating every detail of the meticulous process involved in a meta-analysis is impossible. To understand the relevant principles of meta-analysis, we consulted several medical experts. We

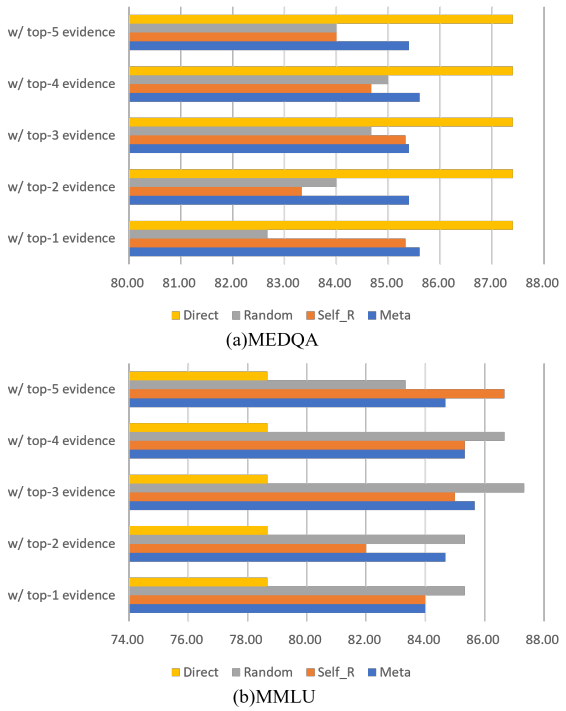

Figure 10: Performance curves of the GPT-4o model under our method.

find that the heterogeneity analysis step typically requires detailed experimental data from each medical paper to perform calculations and comparisons. However, our LLM cannot email each author to obtain this data. Therefore, our heterogeneity analysis is merely a coarse-grained exclusion method based on statistical approaches. In our future work, we will attempt to refine this process to make it more precise and credible.

(3)In fact, it is difficult to define the relationship between the reliability score and the generalizability score in a strict mathematical form. Equation (7) is the most effective and most convincing scheme that we obtained after testing multiple evidence-ranking variants.However, in the medical evidence filtering process, evidence with excessive limitations cannot be applied to users. Unfortunately, the LLMs are hard to make the decision only based on literature abstracts. Therefore, we have to implement this component from other perspectives as a scoring system based on the number of limitations. In our future work, we will optimize this extrapolation evaluation system to make it more reasonable.