Query-driven Document-level Scientific Evidence Extraction from Biomedical Studies

Massimiliano Pronesti[1,2], Joao Bettencourt-Silva[1], Paul Flanagan[2], Alessandra Pascale[1], Oisín Redmond[2], Anya Belz[2†], Yufang Hou[1,3†]
[1]IBM Research Europe - Ireland, [2]Dublin City University,
[3]IT:U Interdisciplinary Transformation University Austria
Correspondence: massimiliano.pronesti@ibm.com, yufang.hou@it-u.at

Abstract

Extracting scientific evidence from biomedical studies for clinical research questions (e.g., *Does stem cell transplantation improve quality of life in patients with medically refractory Crohn's disease compared to placebo?*) is a crucial step in synthesising biomedical evidence. In this paper, we focus on the task of document-level scientific evidence extraction for clinical questions with conflicting evidence. To support this task, we create a dataset called COCHRANEFOREST leveraging forest plots from Cochrane systematic reviews. It comprises 202 annotated forest plots, associated clinical research questions, full texts of studies, and study-specific conclusions. Building on COCHRANEFOREST, we propose URCA (Uniform Retrieval Clustered Augmentation), a retrieval-augmented generation framework designed to tackle the unique challenges of evidence extraction. Our experiments show that URCA outperforms the best existing methods by up to 10.3% in F1 score on this task. However, the results also underscore the complexity of COCHRANEFOREST, establishing it as a challenging testbed for advancing automated evidence synthesis systems.

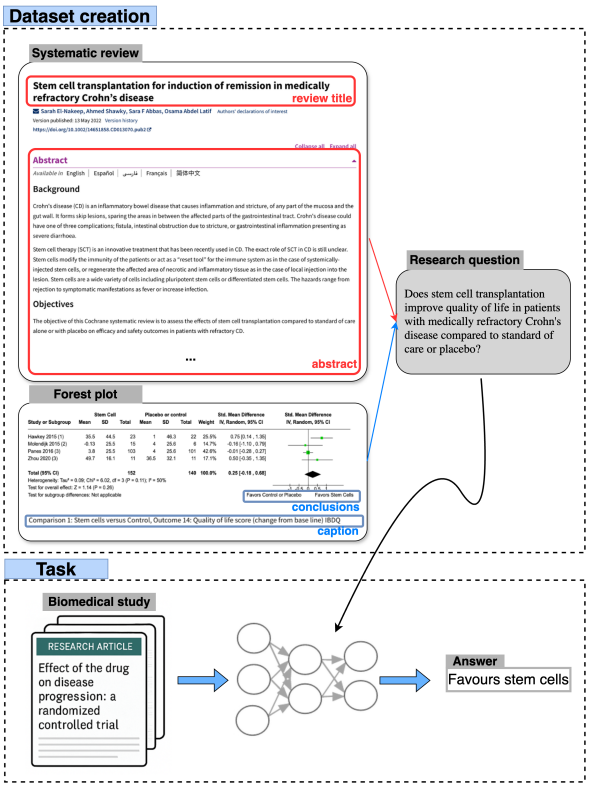


Figure 1: An example from COCHRANEFOREST. The research question is annotated based on systematic review context and forest plot, where each row represents a study and its conclusion about the question.

1 Introduction

Medical practitioners face the challenge of staying current with the ever-growing volume of medical research, making it increasingly difficult to discern meaningful findings from irrelevant ones. Systematic reviews aim to address this challenge by synthesising all relevant evidence for a specific clinical question (Higgins et al., 2024), providing clear, up-to-date answers derived from high-quality research. Systematic reviews are widely regarded as the gold standard in evidence-based medicine, heavily influencing medical decisions made by doctors, health authorities, and patients.

However, the process of producing systematic reviews is both time-consuming and costly. A 2019 study estimated that on average conducting a systematic review takes 1–2 years and costs over $141,000 (Michelson and Reuter, 2019), due to the extensive effort required to sift through vast amounts of potentially relevant studies and perform rigorous statistical analysis before drawing a final conclusion. Given the substantial resources required, there is growing interest in automating various steps of the process (Marshall and Wallace, 2019; Khraisha et al., 2024; Yun et al., 2023, 2024).

In this paper, we focus on document-level sci-

[†]These authors jointly supervised this work.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28034–28051

entific evidence extraction for clinical questions from biomedical studies that come to contradictory conclusions. This involves identifying relevant information from the papers comprising a study and deriving what the study concludes in relation to a question. To facilitate research on this task, we construct COCHRANEFOREST, a new dataset containing 202 human-annotated forest plots extracted from 48 real-world systematic reviews, along with the full texts of the corresponding 263 unique included studies. A forest plot, as illustrated in Figure 1, is the cornerstone of biomedical systematic reviews, consolidating diverse study results into a single axis. It facilitates direct comparison of findings from different studies and provides a visual representation of statistical analyses. In our work, we leverage forest plots to annotate random control trial (RCT) studies containing contradicting conclusions about a research question (Section 3).

To tackle the challenge of extracting scientific evidence from RCTs, which often involve multiple papers reporting results at various stages of the study, we propose **URCA (Uniform Retrieval Clustered Augmentation)**, a retrieval-augmented generation framework designed to address the unique challenges of evidence extraction from individual studies. By uniformly retrieving relevant passages from different papers within the same study and leveraging large language models (LLMs) to distil query-specific information from the clusters of the retrieved passages, URCA addresses common challenges in traditional RAG systems, such as noisy retrieval and the failure to incorporate relevant information from diverse sources (Section 4).

Through extensive experimentation, we demonstrate that URCA outperforms strong baselines, achieving up to a 10.3% improvement in F1 score for predicting the conclusions of RCT studies for the given questions. Our results and analyses demonstrate the potential of the new task and dataset to support significant future research in this domain, as well as the robustness of the proposed method when applied to other question answering (QA) tasks and datasets (Section 5).

Our main contributions are as follows: (1) we introduce and formalise the task of document-level evidence extraction from RCT studies that present contradictory findings in response to a research question; (2) we propose COCHRANEFOREST, a novel dataset derived from forest plots in systematic reviews, encompassing research questions, the corresponding full-text studies, and their conclusions regarding the research question; (3) we develop a novel RAG-based approach, URCA, and establish it as a robust baseline for this task.

2 Background and Task Definition

Systematic review. A systematic review is a rigorous process for identifying, evaluating, and synthesising evidence from multiple studies to address a specific research question. Biomedical systematic reviews often compare clinical interventions by analysing data from trials or studies to inform clinical guidelines.

Papers and studies. A key distinction exists between *papers* and *studies*. Medical RCT studies often generate substantial data across long time periods and multiple publications. During the systematic review process, numerous studies and papers are assessed, but only a subset of studies and their corresponding papers are included in the final synthesis of evidence; these are referred to as *included studies*.

Forest plots. A forest plot is a graphical representation used in systematic reviews to summarise the results of multiple studies regarding a particular research question. It visually displays the effect sizes of individual studies along with their confidence intervals, allowing for easy comparison across studies. As shown in Figure 1, the plot typically includes a central vertical line representing the null effect (i.e. no association), with each study shown as a point estimate and a horizontal line indicating the confidence interval. A summary estimate, often represented by a diamond shape, provides an overall effect size based on all included studies.

Task definition. The input to our task is a research question q and a set of studies $\mathcal{S}$ under evaluation in a forest plot $\mathcal{F}$ belonging to a systematic review (Figure 1). The research question q is formulated to compare the effects of two interventions (e.g., *stem cell transplantation* and *placebo*) for a specific condition (e.g., *patients with medically refractory Crohn's disease*) with regard to a specific outcome measure (e.g., *quality of life*). Each study $s \in \mathcal{S}$ contains one or more papers $S = \{p_1, ..., p_n\}$ and is associated with a conclusion c with respect to the research question q. For each study $s \in \mathcal{S}$, the system must predict the correct conclusion c (e.g., *favours interventions, favours placebo, no difference*) given q and the papers $\{p_1, ..., p_n\}$.

3 The COCHRANEFOREST Dataset

The COCHRANEFOREST dataset consists of 202 forest plots derived from 48 Cochrane[1] medical systematic reviews. Each forest plot is characterised by a research question, a set of studies under assessment, and a list of conclusions. In the following sections, we describe the process used to create and annotate this corpus.

3.1 Dataset Preparation

To construct COCHRANEFOREST, we leveraged the Cochrane Database of Systematic Reviews (CDSR),[2] the leading resource for systematic reviews in healthcare containing 9,301 systematic reviews encompassing over 220,000 included studies as of September 2024.

Our construction process was guided by the dual goals of ensuring diversity in the dataset while addressing practical challenges such as accessibility and data quality. To this end, we began by downloading the CDSR database and then applied a series of systematic filtering steps to refine the dataset, as shown in Figure 5 (Appendix A.1).

First, we excluded withdrawn systematic reviews, which comprised approximately 0.5% of the database, and retained only the latest version of each review to eliminate redundancy caused by prior versions. Reviews with fewer than two included studies were also discarded to ensure that every forest plot in the dataset represented meaningful evidence synthesis. A significant challenge lay in ensuring that all studies referenced in the systematic reviews were accessible in full text. In fact, systematic reviews often include diverse sources such as journal articles, clinical trial reports, and even PhD theses, and only a subset of these are openly available. To address this issue, we retained only reviews for which all included studies were openly accessible. We refer to this subset of reviews as *complete* reviews, and these represent the backbone of our corpus. Since we are interested in contradictory scientific evidence, we further filter the corpus to retain only systematic reviews with forest plots that contain at least two studies with contradicting conclusions. A forest plot is considered to come to contradictory conclusions if at least one study presents findings that differ from another study. For instance, we consider that a study favouring stem cell transplantation contradicts another study reporting no difference between stem cell transplantation and placebo.

After these filtering steps, the final COCHRANEFOREST dataset consists of 202 annotated forest plots extracted from 48 systematic reviews, 263 unique studies, and 923 total records (research question-study pairs). It is worth noting that since COCHRANEFOREST was built from the whole Cochrane database of systematic reviews, it represents the biggest possible dataset of systematic reviews that complies with the open accessibility requirement for the included studies. Moreover, its scale is comparable to established benchmarks such as BioASQ (Tsatsaronis et al., 2012) (618), PubMedQA (Jin et al., 2019) (500), and MMLU-Med (Hendrycks et al., 2021) (1089). Distribution of studies and papers is shown in Table 1.

	Mean	Max	Min
Studies per review	5.67	13	2
Papers per review	6.87	15	2
Papers per study	1.82	5	1
Studies per forest plot	4.57	10	2

Table 1: Statistics for the dataset.

3.2 Forest Plot Annotation

For each forest plot, annotators were provided with the corresponding Cochrane systematic review, a corresponding research question, the set of studies included in the forest plot analysis, and the possible conclusions for each study. Each research question was generated by prompting `llama-3.1-70b` given the title and abstract of the systematic review as well as the caption and the conclusions set of the forest plot. See Appendix D for more details.

The annotation tasks were as follows. First, annotators were asked to verify and, when necessary, edit the automatically generated research question to ensure it was consistent with the analysis shown in the forest plot (**Task 1**, Figure 8b in the appendix). This involved ensuring that the question accurately captured the target population, intervention, comparator, and outcome.

Next, annotators were presented with three predefined labels (i.e., *favours left intervention, favours right intervention, show no difference between left and right interventions*) for annotating the conclusion of each study, with one label preselected based on the automatically extracted 95% confidence interval (CI) reported in the forest plot (**Task 2**, Figure 8b in the appendix). See Appendix

[1] https://www.cochranelibrary.com

[2] https://www.cochranelibrary.com/cdsr/reviews

B for more details. Notably, no annotator modified the pre-selected label.

Finally, annotators reviewed and, when necessary, revised the two intervention names extracted directly from the axes of the forest plot (**Task 3**, Figure 8b in the appendix). We asked annotators to rephrase when the original text was deemed insufficiently clear or explicit out of context (e.g. by expanding acronyms or clarifying ambiguous terms). Combining the annotation results from Tasks 2 and 3, we can effectively derive concrete conclusions for each study in a forest plot.

Annotators included two experts with NLP backgrounds, two graduate students studying computer science, and a medical domain expert, all of whom are co-authors of this paper. The annotation interface is shown in Appendix E.

3.3 Inter-annotator Agreement

To assess the reliability of the annotations, 15 forest plots (63 studies in total) were randomly selected and assigned for independent re-annotation by four of the annotators. We exclude Task 2 from the analysis as no annotator modified the pre-selected labels. We computed two kinds of agreement metrics: (1) **per-annotator metrics**, which evaluate individual annotators' performance relative to the original annotation content; and (2) **aggregated pairwise metrics**, which assess the consistency between pairs of annotators across all annotation tasks, providing an aggregated measure of agreement. For per-annotator evaluation, we measure the average proportion of items changed φ, the character-based Levenshtein distance (Levenshtein, 1966) and HTER (Snover et al., 2006) against the original annotation item (Table 7 in the appendix). Aggregated agreement is computed as mean pairwise HTER and mean pairwise semantic similarity, where the latter is defined as the cosine similarity between the embedding vectors of the annotation items. Additionally, we report Fleiss' κ (Fleiss, 1971) on annotators' binary change decisions (any change vs. none per item).

Results (Table 2) indicate that, while edit-based metrics yield relatively low scores due to their sensitivity to minor textual changes, the semantic similarity between annotations is notably high, with cosine similarity scores of 0.95 for Task 1 and 0.90 for Task 3. This suggests strong agreement on the texts' underlying meaning. Full metric definitions and additional analyses are provided in Appendix A.2.

Metric	Score	
	Task 1	Task 3
Cosine Similarity ($\bar{s}$)	0.95	0.90
HTER	0.34	0.36
Fleiss κ (Fleiss, 1971)	0.06	0.55

Table 2: Aggregated metrics across all annotators for Task 1 and Task 3.

4 URCA: Uniform Retrieval Clustered Augmentation

In this section, we present URCA (Uniform Retrieval Clustered Augmentation), a novel retrieval-augmented generation (RAG) approach tailored for document-level evidence extraction defined in Section 2.

4.1 Framework Overview

As detailed in Algorithm 1 and illustrated in Figure 2, URCA starts from uniformly distributing the retrieval size over the sources of interest. Then the retrieved passages are clustered and aggregated by extracting the relevant information to the query using a language model. Eventually, the final answer is generated based on the aggregated information produced in the previous step.

Algorithm 1 URCA

Require: Query q, Desired number of retrieval passages k, Scaling factor β, Sources $\mathcal{S} = \{s_1, ..., s_S\}$, Embedding Retriever $\mathcal{R}$, Language model $\mathcal{M}_\theta$, Prompt templates $p_{\text{extr}}, p_{\text{ans}}$

$k_s \leftarrow \lceil \min(k + \beta \cdot \log(S), N_{\max})/S \rceil$
$E_T \leftarrow []$
for each source $s \in \mathcal{S}$ **do** ▷ Uniform retrieval
  $[(e^s_1, t^s_1), ..., (e^s_{k_s}, t^s_{k_s})] \leftarrow \mathcal{R}(q, s_i, k_s)$ ▷ Retrieve embeddings and texts
  $E_T \leftarrow E_T \oplus [(e^s_1, t^s_1), ..., (e^s_{k_s}, t^s_{k_s})]$
end for
$[c_1, ..., c_n] \leftarrow \text{Cluster}(E_T)$ ▷ cluster embeddings with UMAP + GMM
for each cluster c **do**
  $D_i \leftarrow \mathcal{M}_\theta(p_{\text{extr}}, q, c)$ ▷ Extract knowledge from clustered texts
end for
$a \leftarrow \mathcal{M}_\theta(p_{\text{ans}}, q, \langle D_1, ..., D_n \rangle)$ ▷ Generate the final answer
return a

4.2 Uniform Retrieval

In the first step, we retrieve the top-k most relevant documents for the query q, ensuring balanced representation across the sources of interest (i.e., the study papers). To achieve this, we allocate a portion of k to each source. Specifically, for each

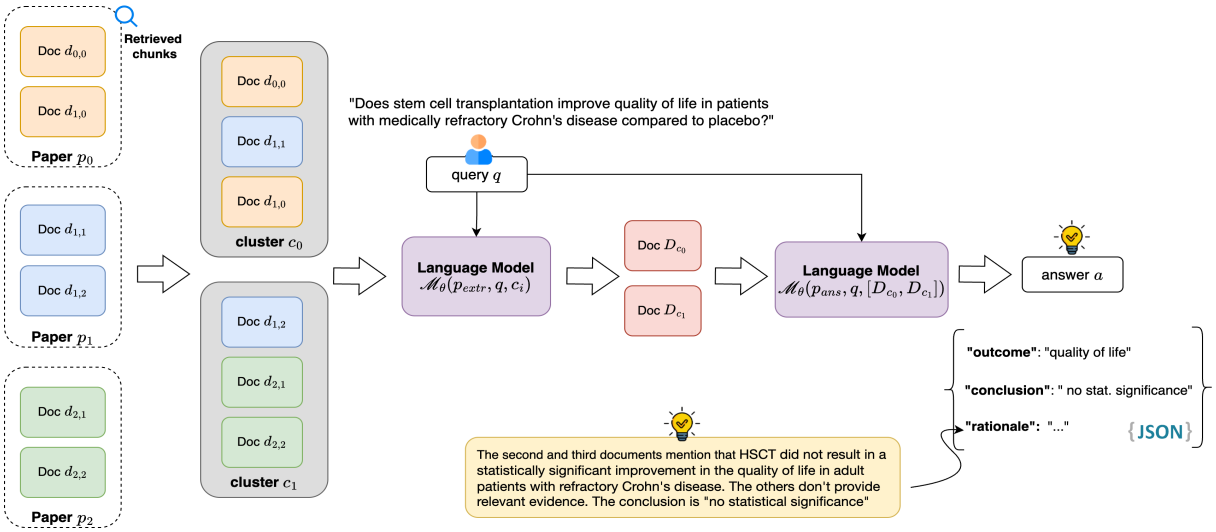


Figure 2: Overview of the URCA framework. Retrieval size is first distributed across all papers. Retrieved chunks are then clustered and aggregated into passages. Finally, the obtained passages are used to generate the final answer.

source, we retrieve k_s entries as follows:

$$k_s = \lceil \min(k + \beta \cdot \log(S), N_{\max})/S \rceil$$

where S is the number of sources under consideration and β is a scaling factor that controls the impact of the logarithmic adjustment. A higher β ensures more results are retrieved per source when the number of sources S is large, which is particularly important when $k < S$. Conversely, a lower β reduces oversampling, keeping the allocation closer to an even distribution of k/S. This approach ensures proportionality while maintaining sufficient representation from each source. We find that this approach is beneficial regardless of the subsequent steps of the pipeline.

4.3 Clustering and Knowledge Extraction

In the second step, we cluster the embeddings of the retrieved documents and employ an LLM to extract the relevant information from each cluster given the query. Our clustering approach follows the methodology introduced by RAPTOR (Sarthi et al., 2024), which employs Gaussian Mixture Models (GMM) for clustering, along with Uniform Manifold Approximation and Projection (McInnes et al., 2018) for dimension reduction of the dense embeddings, and the Bayesian Information Criterion (BIC) for model selection.

To aggregate knowledge, we prompt the LLM $\mathcal{M}_\theta$ (with p_{extr} given in Appendix D) to identify the relevant evidence for the research question q from each cluster of texts c_i. This process discards information irrelevant to q, particularly valuable for addressing complex questions. The extracted information for each cluster is represented as:

$$D_i = \mathcal{M}_\theta(p_{extr}, q, c_i), \quad i = 1, ..., n$$

The resulting set of D_i passages can be viewed as distilled, query-aware evidence of each cluster.

4.4 Answer Finalisation

In the last step, we prompt the LLM (with prompt p_{ans} in Appendix D) to generate the final answer given the research question q and the passages $\langle D_1, ..., D_n \rangle$ produced in the previous step:

$$a = \mathcal{M}_\theta(p_{ans}, q, \langle D_1, ..., D_n \rangle)$$

This step combines the distilled insights from each cluster into a coherent and concise answer. By deferring final synthesis until after clustering and extraction, we ensure that the model operates on already-filtered and semantically aligned information, reducing noise and improving accuracy.

5 Experiments

5.1 Experimental Setup

Problem formulation. We adopt the standard RAG setting where the LLM M_θ has access to an external knowledge base through an off-the-shelf retriever $\mathcal{R}$. In contrast to existing approaches, we assume we have pre-filtered sources S relevant to a research question q. These are the papers comprising the studies under assessment. Given $q, \mathcal{R}$ and a

set of possible conclusions $\mathcal{C}$, the goal is to predict the correct conclusion $c \in \mathcal{C}$ to q.

Notably, we directly employ off-the-shelf retrievers instead of training our own, and prepend all retrieved documents to the question as input to the model, without any re-ranking. This setting is orthogonal to existing research efforts centered on improving the retriever or performing adaptive retrieval (Wang et al.; Asai et al., 2024).

Baselines. To evaluate URCA on COCHRANEFOREST, we compare it against a diverse set of baselines testing different retrieval and synthesis strategies. As a starting point, we include two baselines that eliminate active retrieval: (1) *No RAG* which relies solely on the model's internal knowledge, and (2) *Abstracts* which injects study abstracts directly into the context. These test the importance of external evidence acquisition. Moreover, we include *vanilla RAG*, a straightforward retrieval-augmented generation approach, which we evaluate both with and without uniform retrieval to highlight the limitations of simpler methods and the potential benefits of incorporating a more balanced retrieval strategy. Additionally, we include INSTRUCTRAG (Wei et al., 2025), a more advanced variant that prompts the model to provide rationales connecting answers to the retrieved evidence in passages. For a fair comparison, we use the instructions without training or in-context learning. RAPTOR (Sarthi et al., 2024), the state-of-the-art method that uses recursive clustering and summarisation to synthesise information, serves as a robust comparison for tasks requiring multi-source evidence synthesis. Lastly, we include GRAPHRAG (Edge et al., 2024), which builds a graph-based text index by summarising closely related entities from the source documents.

Other datasets. While URCA is designed to leverage the structured nature of systematic reviews, its underlying approach can be applicable to broader question-answering tasks. To explore this possibility, we conduct experiments on two widely used datasets: PubMedQA (Jin et al., 2019) and MedQA-US (Jin et al., 2021). In these evaluations, we removed the uniform retrieval step for two key reasons. First, uniform retrieval is most beneficial when the relevant sources are known in advance, as is the case for systematic reviews, but this assumption does not hold for open-domain QA tasks. Second, clustering provides greater performance gains than uniform retrieval (Table 5).

LLM and RAG settings. We conduct experiments on both open and closed-source LLMs of different sizes, including `Llama-3.1-70B`, Mistral Large (`Mistral-Large-Instruct-2407`), GPT 3.5 Turbo (`gpt-3.5-turbo-0613`), and GPT 4 (`gpt-4-0613`). The generation temperature is set to 0, and maximum output tokens is set to 1,024. By default, we use the top 10 retrieved passages in all the approaches under comparison.

5.2 Main Results

Performance on COCHRANEFOREST. Table 3 presents the results on COCHRANEFOREST for each model and approach under evaluation. Comparing No RAG and the various retrieval-based methods, we observe that relying solely on the model's internal knowledge is insufficient for this task, highlighting the necessity of external evidence to address domain-specific questions. In contrast, the Abstracts approach emerges as a surprisingly strong baseline. The improvements achieved by standard RAG or even more sophisticated approaches like RAPTOR and INSTRUCTRAG are relatively small in comparison, underscoring the value of abstracts as a compact and effective evidence source. However, abstracts alone are ultimately not enough to address the level of granularity required for outcome-specific questions, which often rely on access to tables, analyses, and other fine-grained evidence beyond the abstract.

Interestingly, RAG and RAG + Uniform Retrieval demonstrate clear improvements over the previous baselines and over more sophisticated approaches such as INSTRUCTRAG. This can be attributed to the latter's reliance on carefully curated demonstrations, a limitation also noted by Wang et al. (2024) for domain-specific datasets like BioASQ (Tsatsaronis et al., 2012). RAPTOR, despite sharing conceptual similarities with URCA, also performs poorly overall. This behaviour can be explained by its offline clustering and synthesis processes which are not guided by the query. As a result, it risks focusing on irrelevant outcome measures or introducing information loss, thus misleading the LLM, ultimately limiting its effectiveness. Similar considerations can be made for GRAPHRAG, which relies on multiple steps of entity linking and offline summarisation.

This underscores a critical insight from our findings: more sophisticated RAG variants are not inherently better in domain-specific biomedical tasks unless their design is explicitly aligned with the

Method	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
	Mistral-Large (2407), 128B				**LlaMa 3.1, 70B**			
No RAG	46.1	47.4	44.8	53.1	49.1	46.1	52.4	45.2
Abstracts	62.6	60.3	65.0	62.6	60.7	60.4	61.1	64.1
RAG	60.9	60.7	61.0	64.3	62.1	60.2	64.1	62.9
+ Uniform Retrieval	63.3	60.9	65.8	63.2	63.4	61.5	65.5	64.2
RAPTOR	61.7	59.1	64.6	61.0	60.6	58.2	63.2	60.0
InstructRAG	62.6	61.5	63.6	64.4	60.9	59.6	62.3	62.1
GraphRAG	64.9	63.8	66.0	66.3	65.6	**64.8**	66.3	63.4
URCA (Ours)	**67.3**	**65.2**	**69.5**	**67.0**	**66.1**	64.0	**68.4**	**64.6**
	GPT-4 (0613)				**GPT-3.5-Turbo (0613)**			
No RAG	47.5	55.9	41.2	59.5	24.1	18.9	33.2	56.4
Abstracts	61.0	60.2	61.9	62.6	56.0	55.2	56.9	58.8
RAG	61.6	59.3	64.0	61.6	59.1	57.6	60.6	60.7
+ Uniform Retrieval	62.0	59.7	64.5	61.8	61.8	60.0	63.8	62.8
RAPTOR	60.1	58.2	62.0	60.8	53.6	52.0	55.3	54.1
InstructRAG	61.6	61.4	61.9	63.3	57.4	57.1	57.7	61.1
GraphRAG	63.8	61.6	66.1	63.7	56.6	58.4	54.9	59.3
URCA (Ours)	**65.7**	**63.0**	**68.7**	**64.1**	**62.4**	**60.6**	**64.4**	**63.5**

Table 3: Overall results of URCA and seven baselines on COCHRANEFOREST on four LLMs, showing micro-F1 score, micro-precision, micro-recall, and accuracy. Best scores are reported in bold.

Method	MedQA-US	PubMedQA
No RAG	72.1	77.5
RAG	82.3	79.6
GraphRAG	84.5	80.6
URCA	**85.9**	**81.1**

Table 4: Accuracy on MedQA-US and PubMedQA.

evidence structure of the data.

Overall, URCA delivers the strongest performance across all models and metrics on COCHRANEFOREST. In particular, URCA improves the best baseline (GraphRAG) by at least 3% relative F1 on Mistral Large and GPT-4, reaching a peak of 10.3% on `gpt-3.5-turbo`.

Performance on open-domain medical QA. Table 4 presents the results obtained using GPT-4 under the same experimental settings on two popular medical QA datasets: MedQA-US and PubMedQA. We observe that URCA outperforms the baselines under comparison, demonstrating the robustness and generalisability of our approach beyond the domain of systematic reviews. These results demonstrate that while URCA is tailored for structured biomedical evidence synthesis, its underlying principles can generalise well to broader biomedical QA tasks. On PubMedQA, which focuses on yes/no/maybe questions grounded in abstracts, URCA achieves 81.1% accuracy, surpassing both GraphRAG and standard RAG by a notable margin. This suggests that even in cases where the input context is less structured than full-text reviews, the clustering mechanism still facilitates better information aggregation and reasoning.

The gains are even more pronounced on MedQA-US, a challenging dataset based on US medical licensing exam questions. These questions are longer, more nuanced, and require multi-hop reasoning across biomedical knowledge. URCA achieves 85.9% accuracy, outperforming GraphRAG by 1.4% and standard RAG by 3.6%.

We attribute URCA's strong cross-domain performance to its design, which inherently filters and restructures the retrieved content, disentangling overlapping concepts and offering a more interpretable intermediate representation to the model.

5.3 Ablation Studies and Analyses

Influence of components. As shown in Table 5, we ablate URCA's design from two aspects: (1) *w/o Uniform Retrieval*, where the retriever only fetches the top chunks from the knowledge base regardless of the source; (2) *w/o Clustering*, where the knowledge extraction step is performed without clustering the retrieved passages. We observe that both uniform retrieval and clustering play important roles in URCA's design, with clustering having a more pronounced impact. Removing clustering consistently leads to larger performance drops ($4-6\%$ in F1), underscoring its critical role in organising retrieved content for effective knowledge extraction. In contrast, the absence of uniform re-

Model	Method	F1	Precision	Recall	Accuracy
Llama-3.1-70B	URCA	66.1	64.0	68.4	64.6
	w/o Uniform Retrieval	64.5 (↓ 2.4%)	63.3	65.9	64.6
	w/o Clustering	63.2 (↓ 4.5%)	61.9	64.5	63.6
Mistral-Large-2407	URCA	67.3	65.2	69.5	67.0
	w/o Uniform Retrieval	66.17 (↓ 1.6%)	64.4	68.1	66.5
	w/o Clustering	64.6 (↓ 3.9%)	63.0	66.4	64.9
GPT-3.5-turbo-0613	URCA	62.4	60.6	64.4	63.5
	w/o Uniform Retrieval	61.0 (↓ 2.2%)	61.5	60.5	62.6
	w/o Clustering	59.4 (↓ 4.8%)	59.1	59.8	61.2
GPT-4-0613	URCA	65.7	63.0	68.7	64.1
	w/o Uniform Retrieval	64.7 (↓ 1.6%)	62.1	67.5	63.2
	w/o Clustering	61.9 (↓ 5.7%)	59.4	64.7	60.3

Table 5: Ablation studies on the impact of uniform retrieval and clustering on URCA.

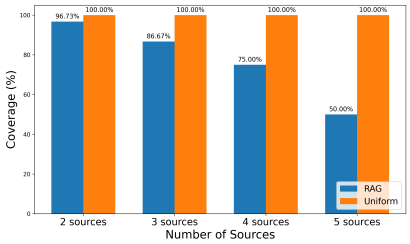


Figure 3: Coverage rate of multiple sources with and without uniform retrieval.

trieval results in smaller but noticeable declines ($1.0 - 2.5\%$), highlighting its role in ensuring diverse coverage. These results confirm that clustering is the primary driver of URCA's performance, while uniform retrieval further enhances its robustness. Figure 3 further supports this by showing that uniform retrieval improves the coverage rate of multiple sources, mitigating biases where one or two papers dominate the retrieved context. This is particularly important in our setting, where individual studies may be documented by a varying number of papers.

Influence of cluster ordering. Following the passage ordering strategies defined by Alessio et al. (2024), we assess the impact of cluster reordering in URCA. In particular, we compare five strategies: ascending or descending similarity to the query, random ordering, and two ping-pong variants. The obtained results are reported in Table 6. We observe that the performance improvement of URCA compared to baseline methods is significantly greater than the performance variations observed due to changes in cluster order within URCA. In addition, thanks to the knowledge extraction step, URCA proves to be robust to the chosen ordering approach, with only minor performance variations of $\pm 1\%$.

Benefits of clustering. We perform a quantitative analysis in which we replace clustering with contiguous grouping, i.e. after randomly shuffling the retrieved chunks, we naively group contiguous chunks. Figure 4 shows that grouping the chunks without a criterion worsens the overall performance. This highlights the importance of structured grouping in preserving information quality, as naive contiguous grouping disrupts coherence making it harder for the model to identify consistent signals, whereas clustering helps identify meaningful relationships between chunks and better model judgment in the information extraction step.

Qualitative Example. Figure 6 presents a qualitative example from CochraneForest in which we show the intermediate steps of URCA and how the clustering steps improve the final prediction. Without clustering, the LLM doesn't properly handle the context and extracts information regarding unrelated outcomes, thus reaching a wrong conclusion. In contrast, by grouping and filtering chunks within clusters, URCA provides better judgment of ambiguous or mixed data and discards irrelevant information early. Further details about this example can be found in Appendix C.2.

6 Related work

Medical evidence extraction. Previous studies on medical evidence extraction mainly focus on the abstract or individual passage level, such as inferring the effectiveness of a treatment for a given condition (Nye et al., 2020), extracting contradictory claims about COVID-19 drug efficacy (Sosa et al., 2023), and summarising single or multiple

Model	Ordering Strategy	F1	Precision	Recall	Accuracy
Llama-3.1-70B	Ascending	66.1	63.4	68.4	64.6
	Descending	67.1	64.7	69.6	66.0
	Random	66.1	63.9	68.6	65.2
	Ping-pong Descending Top-to-bottom	66.0	63.7	68.5	64.6
	Ping-pong Descending Bottom-to-top	66.8	64.5	69.4	65.7
Mistral-Large-2407	Ascending	67.3	65.2	69.5	67.0
	Descending	68.1	65.7	70.7	67.2
	Random	67.6	65.1	70.2	66.6
	Ping-pong Descending Top-to-bottom	67.5	65.1	70.1	66.5
	Ping-pong Descending Bottom-to-top	67.5	65.1	70.2	66.5
GPT-3.5-turbo-0613	Ascending	62.4	60.6	64.4	63.5
	Descending	62.5	61.2	64.0	63.6
	Random	61.2	59.8	62.7	61.8
	Ping-pong Descending Top-to-bottom	61.9	60.4	63.5	62.3
	Ping-pong Descending Bottom-to-top	61.9	60.4	63.5	62.3
GPT-4-0613	Ascending	65.7	63.0	68.7	64.1
	Descending	64.7	61.7	67.9	62.4
	Random	64.4	61.5	67.5	62.0
	Ping-pong Descending Top-to-bottom	64.5	61.6	67.8	62.0
	Ping-pong Descending Bottom-to-top	64.6	61.7	67.8	62.0

Table 6: Ablation study on cluster ordering strategies for URCA. “Ascending" is used in our main experiments.

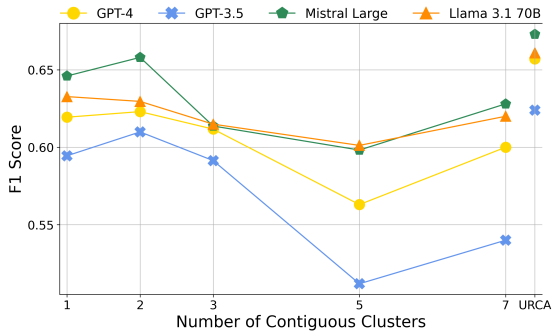


Figure 4: Performance with contiguous grouping versus URCA on different numbers of contiguous clusters.

RCT studies (Shaib et al., 2023; O’Doherty et al., 2024). Similar to our task, Lehman et al. (2019) propose a task to infer reported findings from a full-text RCT article given a research question. In contrast, our work can be seen as an example of scientific argument mining at the global discourse level (Al Khatib et al., 2021). It focuses on extracting medical evidence for a given question across multiple full-text documents, emphasising cross-document synthesis at the full-document level. Additionally, our research questions are naturally derived from systematic reviews, aligning closely with real-world clinical research interests.

RAG. Existing RAG methods (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2023) have shown promise in improving the factual accuracy of large language models (LLMs) by leveraging up-to-date information and specialised (non-parametrised) knowledge from external sources (Vu et al., 2024; Kasai et al., 2024). However, these methods face significant challenges due to the inclusion of irrelevant or erroneous content from retrieval systems (Khattab et al., 2022; Chen et al., 2024; Xiang et al., 2024) and the inherent noise in retrieval corpora (Izacard et al., 2022; Shao et al., 2025; Dai et al., 2024). Critically, most existing approaches do not take into account the importance of pre-selecting and diversifying sources. URCA addresses these limitations by ensuring all relevant sources are represented at the retrieval stage.

7 Conclusion

We formalised the complex and challenging task of document-level evidence extraction for clinical questions, and released a dataset (CochraneForest) to support further work on this task. We also proposed URCA, a novel RAG-based framework that retrieves uniformly from multiple sources, clusters relevant evidence, and generates conclusions. Our empirical results show that URCA consistently outperforms state-of-the-art RAG approaches though performance still leaves room for improvement, highlighting the task’s complexity. A future direction is to take a more quantitative approach, extracting numerical data to compute 95% confidence intervals for final conclusions.

8 Limitations

Despite the contributions of this work, there are some limitations that must be acknowledged and addressed in future research. First, in the proposed CochraneForest dataset, we annotate research questions and study conclusions but do not provide rationales identifying which specific passages within the included studies support these conclusions. This limits the interpretability of the annotations and prevents a fine-grained understanding of the reasoning underlying the conclusions. The absence of rationale annotations also limited our ability to conduct a thorough error analysis. Without identifying supporting evidence, we could not examine where and why models fail when extracting conclusions.

References

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

M Alessio, G Faggioli, N Ferro, FM Nardini, R Perego, et al. 2024. Improving RAG systems via sentence clustering and reordering. In *CEUR WORKSHOP PROCEEDINGS*, volume 3784, pages 34–43.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447. ACM.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. *OpenAI Blog. Available online: https://openai. com/blog/new-and-improved-embedding-model (accessed on 28 November 2023)*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors. 2024. *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.5 (updated august 2024) edition. Cochrane. Available from www.training.cochrane.org/handbook.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. REALTIME QA: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. 2024. Can large language models replace humans in systematic reviews? evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4):616–626.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet Physics Doklady*, 10(8):707–710.

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

Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8:1–10.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Matthew Michelson and Katja Reuter. 2019. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications*, 16:100443.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace. 2020. Trialstreamer: Mapping and browsing medical evidence in real-time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69, Online. Association for Computational Linguistics.

James O'Doherty, Cian Nolan, Yufang Hou, and Anya Belz. 2024. Beyond abstracts: A new dataset, prompt design strategy and method for biomedical synthesis generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 358–377, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *the Twelfth International Conference on Learning Representations*.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2025. Scaling retrieval-based language models with a trillion-token datastore. *Advances in Neural Information Processing Systems*, 37:91260–91299.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Daniel Sosa, Malavika Suresh, Christopher Potts, and Russ Altman. 2023. Detecting contradictory COVID-19 drug efficacy claims from biomedical literature. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–713, Toronto, Canada. Association for Computational Linguistics.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. BioASQ: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshness-LLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2024. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *Preprint*, arXiv:2410.07176.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. RAT: Retrieval augmented thoughts elicit context-aware reasoning and verification in long-horizon generation. In *NeurIPS 2024 Workshop on Open-World Agents*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *the Thirteenth International Conference on Learning Representations*.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. *Preprint*, arXiv:2405.15556.

Hye Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Hye Sun Yun, David Pogrebitskiy, Iain James Marshall, and Byron C Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. In *Machine Learning for Healthcare Conference*. PMLR.

A Dataset Construction and Annotation Protocols

A.1 Filtering Systematic Reviews

Figure 5 illustrates the filtering process used to construct the CochraneForest dataset. After downloading the full Cochrane CDSR archive, we applied multiple filtering stages described in Section 3.1 to ensure that only reviews with complete data and high-quality forest plots were retained.

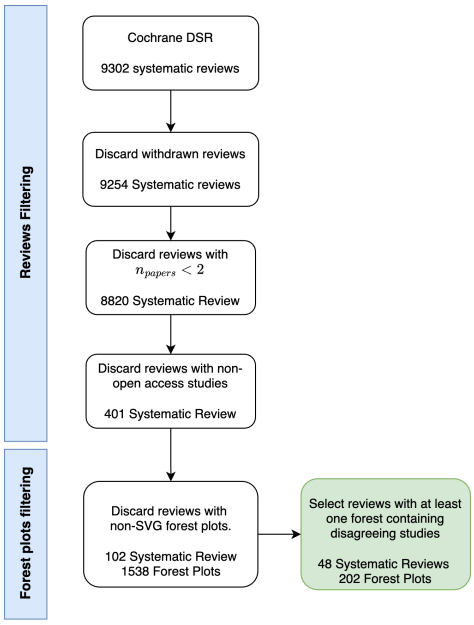


Figure 5: Filtering steps.

A.2 Inter-Annotator Agreement

Per-annotator metrics. For each annotator a_i and set of annotation tasks $\mathcal{T}$, we compute the average proportion of items changed φ, the character-based Levenshtein distance (Levenshtein, 1966) and HTER (Snover et al., 2006) against the original annotation item. Specifically, for an individual annotator a_i, let $\mathcal{T}$ denote the set of annotation tasks (e.g. $\mathcal{T}_1$ for **Task 1**, $\mathcal{T}_3$ for **Task 3**) and $\mu(a_i)$ be one of such metrics. Then $\mu(a_i)$ is defined as the average over all annotation items $t_k \in \mathcal{T}$:

$$\mu(a_i) = \frac{1}{|\mathcal{T}|} \sum_{t_k \in \mathcal{T}} f(a_i, t_k)$$

where $f(a_i, t_k)$ is the function specific to the metric being evaluated. For instance, $f(a_i, t_k) = 1$ if annotator a_i changed the item t_k, 0 otherwise. Table 7 presents the per-annotator results.

Aggregated pairwise metrics. Let $\{a_1, ..., a_m\}$ be the set of annotators who annotated the annotation tasks in $\mathcal{T}$. For each pair of annotators (a_i, a_j) and metric $f(a_i, a_j, t_k)$, we define the pairwise agreement as the average of f across all annotation tasks $t_k \in \mathcal{T}$:

$$\mu(a_i, a_j) = \frac{1}{|\mathcal{T}|} \sum_{t_k \in \mathcal{T}} f(a_i, a_j, t_k)$$

The final aggregated agreement for a task set $\mathcal{T}$ is computed as the mean across all annotator pairs:

$$\overline{\mu}(\mathcal{T}) = \frac{1}{\binom{m}{2}} \sum_{1 \le i \le j \le m} \mu(a_i, a_j)$$

Annotator	Task	φ	Levenshtein	HTER
A	Task 1	0.33	17.60	0.10
	Task 3	0.07	7.73	0.00
B	Task 1	0.87	41.93	0.49
	Task 3	0.40	13.20	0.62
C	Task 1	1.00	50.73	0.50
	Task 3	0.40	13.27	0.70
D	Task 1	0.87	33.07	0.34
	Task 3	0.47	14.93	0.66

Table 7: Per-annotator scores for Task 1 and Task 3 for the original text; φ is the average proportion of items changed.

For this evaluation, we compute again HTER and the semantic similarity expressed as cosine similarity of the embedding vectors $e^i_{t_k}, e^j_{t_k}$:

$$s(a_i, a_j) = \frac{1}{|\mathcal{T}|} \sum_{t_k \in \mathcal{T}} \frac{e^i_{t_k} \cdot e^j_{t_k}}{\|e^i_{t_k}\| \|e^j_{t_k}\|}$$

Each pairwise similarity $s(a_i, s_j)$ is computed averaging the similarities obtained from 3 embedding models of different sizes selected from the MTEB benchmark (Muennighoff et al., 2023): `all-MiniLM-L12-v2` (Reimers and Gurevych, 2019), `NV-Embed-v2` (Lee et al., 2025), and `text-embeddings-ada-002` (Greene et al., 2022).

Average pairwise metrics are reported in Table 8 (cosine similarity), Table 9 (fraction of changes φ), Table 10 (Levenshtein distance) and Table 11 (HTER).

A.3 Annotation Examples

Research question annotation. For research question refinement, annotators were asked to clarify and expand automatically generated questions. For instance, an initial question such as "Does FMT increase the resolution of recurrent *Clostridioides difficile* infection in immunocompetent individuals?" was rewritten expanding acronyms and adding the missing comparator as "Does fecal microbiota transplantation (FMT) increase the resolution of recurrent *Clostridioides difficile* infection in immunocompetent individuals compared to placebo or no treatment?" (Figure 8c).

Study label annotation. In labeling studies, annotators examined each forest plot and determined whether individual studies supported the intervention, the control, or remained inconclusive. These decisions were made by visually assessing whether the 95% confidence intervals crossed the effect threshold indicated in the plot.

Conclusion annotation. For conclusion annotation, annotators revised and expanded the textual conclusions provided by the original review authors. In cases where the conclusions included acronyms or vague references, these were clarified. For example, "Favours FMT" was revised to "Favours fecal microbiota transplantation (FMT)", and "Favours control" became "Favours control group" (Figure 8c).

B Extracting Study-Level Conclusions from Forest Plots

In this section, we describe the method used to determine the overall conclusion of each study in a forest plot, specifically whether the study supports the *intervention*, *control*, or indicates no significant difference, also known as the *null hypothesis*. This process is based on extracting and interpreting the 95% confidence interval (CI) for each study, which is typically represented as a point estimate (e.g., mean difference, odds ratio, or risk ratio) accompanied by a horizontal line indicating the 95% confidence interval. The threshold value plays a crucial role in this analysis; it is generally set at **1** for ratios (e.g., odds ratios, risk ratios) and **0** for mean differences, representing the null hypothesis, where there is no difference between the intervention and control groups.

To determine the conclusion for each study in a forest plot, we followed a systematic approach. First, we extracted the 95% confidence interval and the point estimate effect for each study. These values are typically presented in the format:

Point Estimate [Lower Bound, Upper Bound]

Next, we identified the appropriate threshold value for comparison. For ratios, this threshold is **1**, as values greater than 1 indicate a favouring of the intervention and values less than 1 favour the control. For mean differences, the threshold is **0**, with positive values suggesting favouritism towards the intervention group and negative values favouring the control.

The final step involved comparing the confidence interval to the threshold to determine the conclusion. If the entire confidence interval lies entirely below the threshold (for example, [0.2, 0.8] with a threshold of 1), the study is interpreted as supporting the **intervention group**. Conversely, if the

entire confidence interval is above the threshold (e.g., [1.2, 1.8]), the study supports the **control group**. If the confidence interval crosses the threshold (e.g., [0.8, 1.2]), the result is not statistically significant, meaning the study supports the **null hypothesis**.

C Addition results

C.1 Additional evaluations for URCA

We report precision, recall and F1 score for type of conclusion (*Intervention*, *Control*, *Inconclusive*) in Table 12.

C.2 Qualitative Example

Figure 6 presents a real example from our dataset COCHRANEFOREST. URCA semantically clusters the retrieved chunks into three groups: the first one contains relevant information on the outcome under assessment (*fistula closure*), the second one focuses on numerical aspects related to the time to closure and systemic markets, and the third one on a different outcome (*Crohn's Disease Active Index*). Coloured text indicates the extracted information. By grouping and filtering chunks within clusters, irrelevant information is discarded early. In addition, clustering keeps related chunks together, allowing for better judgment of ambiguous or mixed data (e.g., in Cluster 2). In contrast, without clustering, the LLM doesn't properly handle the context and extracts information regarding the non-statistical significance of PDAI and CDI ("There was no difference in Crohn's Disease Activity Index (CDAI) score between stem cell transplantation and placebo", "[...] we noted no significant differences between treatment groups"), thus incorrectly concluding the overall study is inconclusive with respect to the research question.

D Prompt templates

We provide the prompts we used to execute all our experiments in Figure 7.

E Annotation Interface

The annotation interface used to annotate COCHRANEFOREST is shown in Figure 8. Specifically, Figure 8a presents an overview of the interface, which contains a visualisation of the forest plot, title and abstract of the systematic review, as well as links to the Cochrane review; Figure 8b presents the annotation tasks; Figure 8c illustrates an annotation example for Task 1 and Task 3.

F Hyperparameters and APIs

We executed all the experiments either via API or on our own cluster. We used the paid-for OpenAI API to access GPT-3.5.-turbo and GPT-4. On the other hand, we hosted the open-source models used in this paper on a distributed cluster containing a total of 176 NVIDIA H100 GPUs and served them with vllm (Kwon et al., 2023). As explained in Section 5.1, we set the temperature to 0, $\beta = 2$, and the maximum number of tokens to 1,024 for all models. We left all the other hyperparameters to the default value.

G Scientific artefacts and licensing

In this work, we used the following scientific artefacts. LLaMa 3.1 is licensed under a commercial license.[3] GPT-3.5 Turbo and GPT-4 are licensed under a commercial license.[4] Mistral Large is licensed under the Mistral Research License.[5]. Mining text and data from the Cochrane library is permitted for non-commercial research through the Wiley API.[6] The usage of the listed artefacts is consistent with their licenses.

[3] https://llama.meta.com/doc/overview
[4] https://openai.com/policies/terms-of-use/
[5] https://mistral.ai/licenses/MRL-0.1.md
[6] https://www.cochranelibrary.com/help/access

	A	B	C	D
A		0.94	0.93	0.95
B	0.94		0.98	0.94
C	0.93	0.98		0.93
D	0.95	0.94	0.93	

(a)

	A	B	C	D
A		0.90	0.82	0.89
B	0.90		0.92	0.97
C	0.82	0.92		0.91
D	0.89	0.97	0.91	

(b)

Table 8: Pairwise cosine similarity $s(a_i, a_j)$ between annotators for Task 1(a) and Task 3 (b).

	A	B	C	D
A		1.00	1.00	0.87
B	1.00		0.87	1.00
C	1.00	0.87		1.00
D	0.87	1.00	1.00	

(a)

	A	B	C	D
A		0.53	0.6	0.53
B	0.53		0.6	0.6
C	0.6	0.6		0.67
D	0.53	0.6	0.67	

(b)

Table 9: Pairwise φ between annotators for Task 1 (a) and Task 3 (b).

	A	B	C	D
A		51.27	63.53	33.27
B	51.27		30.53	51.33
C	63.53	30.53		60.53
D	33.27	51.33	60.53	

(a)

	A	B	C	D
A		15.67	17.87	16.73
B	15.67		9.60	4.33
C	17.87	9.60		8.47
D	16.73	4.33	8.47	

(b)

Table 10: Pairwise $lev(a_i, a_j)$ between annotators for Task 1 (a) and Task 3 (b).

	A	B	C	D
A		0.47	0.36	0.19
B	0.47		0.19	0.39
C	0.36	0.19		0.43
D	0.19	0.39	0.43	

(a)

	A	B	C	D
A		0.62	0.42	0.30
B	0.62		0.26	0.22
C	0.42	0.26		0.31
D	0.30	0.22	0.31	

(b)

Table 11: Pairwise HTER(a_i, a_j) between annotators for Task 1 (a) and Task 3 (b).

Model	***Intervention*** **(F1/P/R)**	***Inconclusive*** **(F1/P/R)**	***Control*** **(F1/P/R)**
Llama-3.1-70B	57.4/49.3/68.6	65.3/77.1/56.6	72.0/65.5/80.0
Mistral-Large-2407	63.1/58.3/68.7	68.6/76.9/62.0	67.9/60.4/77.8
GPT-3.5-turbo-0613	54.4/53.9/54.9	67.9/72.7/63.7	64.1/56.9/73.3
GPT-4-0613	61.3/56.6/66.7	63.9/76.5/54.9	67.3/55.9/84.5

Table 12: URCA performance metrics by conclusion category.

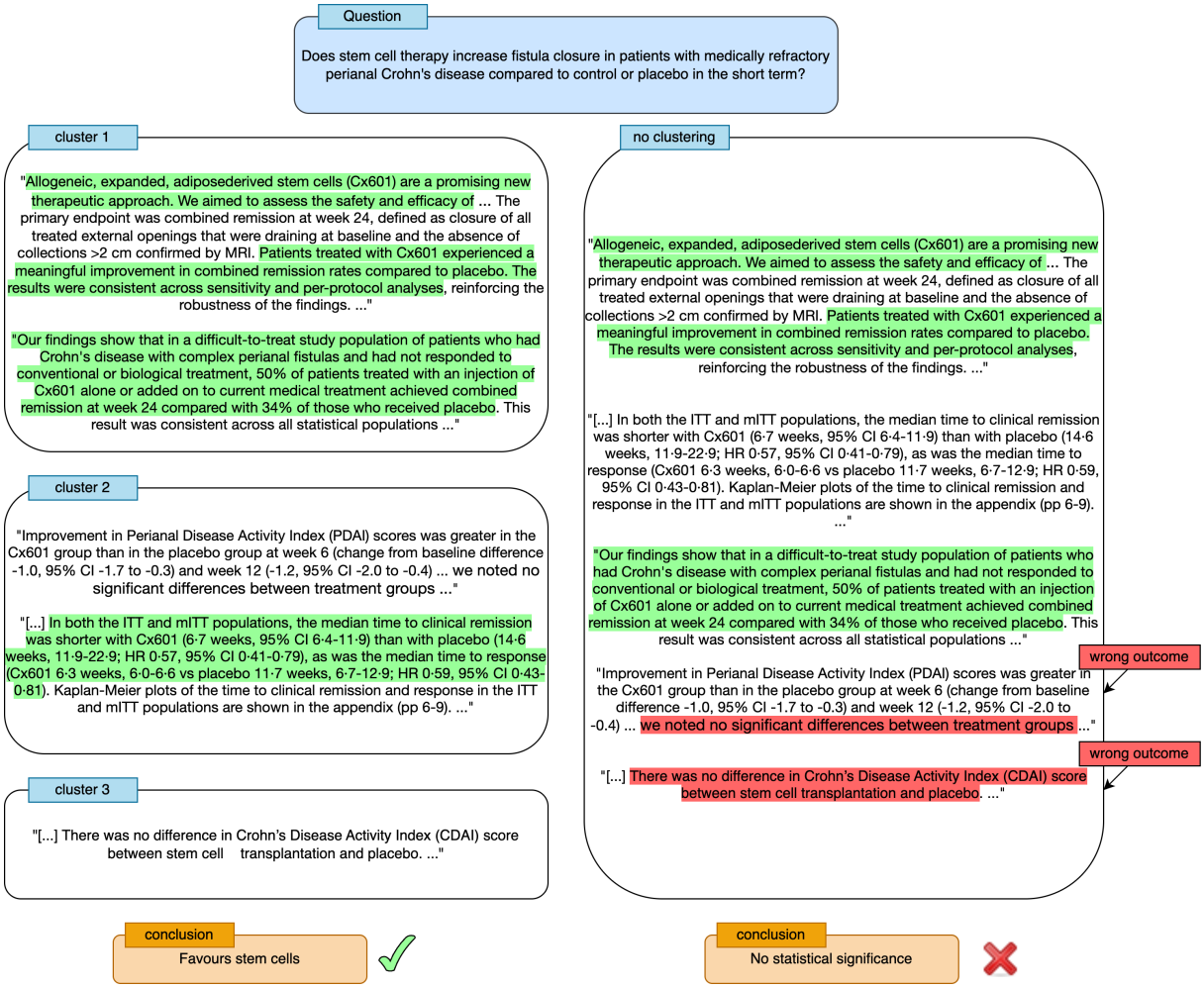


Figure 6: Qualitative example from COCHRANEFOREST. URCA clusters the retrieved passages and filters out irrelevant information from each cluster.

Prompt for research question generation

Task: Generate a research question in natural language for a forest plot belonging to a Cochrane systematic review given:
* the forest plot caption
* the possible conclusions
* the systematic review title
* the systematic review abstract.

The research question must be phrased to be binary and end with a question mark.

Title: {title}
Abstract:
{abstract}

Caption: {caption}
Conclusions: {conclusions}

Research question:

(a)

Prompt for knowledge extraction (p_{extr})

Task: Extract relevant information from externally retrieved documents in response to the given question. These documents come from medical research papers part of a systematic review.

* Report the relevant parts of the original text to the question.
* Focus on the numerical results, statistical significance, conclusions, and any relevant information that can help answer the question.
* If two documents provide conflicting information, report both.
* Make sure your response is long enough to include all the relevant information.
* Do not report documents that only contain irrelevant information.
* Group documents containing similar information together.

Follow this format:
```
Document i:
<relevant information>

Document j, k:
<relevant information from Document j and k if similar or conflicting>
...
```

Initial Documents:
{context}

Question: {question}

Final Documents:

(b)

Prompt for answer finalisation (p_{ans})

Task: Predict the conclusion of a medical study (a collection of papers) for a given research question. You are provided with:
* a research question
* the possible conclusions
* the meaningful chunks of text from the study and their source.

Research question: {question}

Chunks:
{top_chunks}

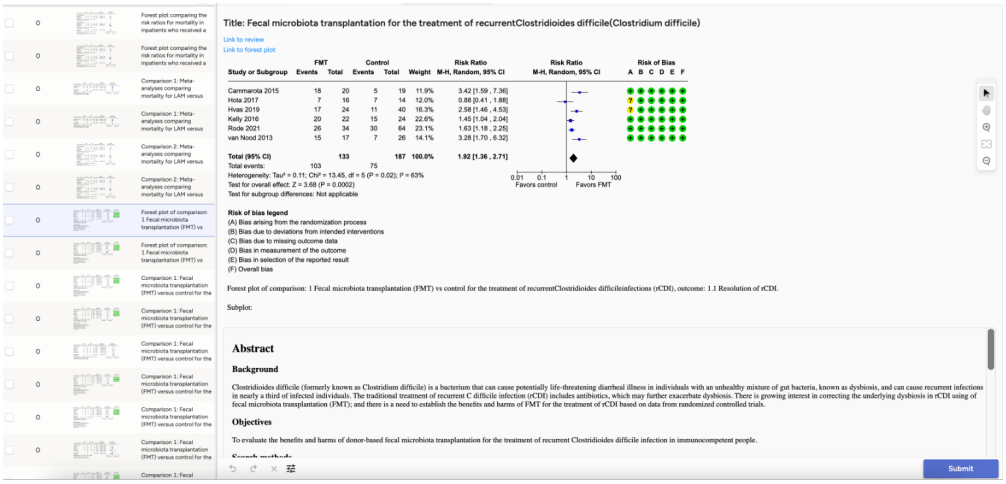
The possible conclusions are: {conclusions}.
“No statistical significance" means that the study is overall inconclusive with respect to the outcome measured.

You must respond with a JSON that respects the following format:
```
“outcome_measured": "the outcome you are assessing"
“rationale": "The portion of the chunks that led you to the conclusion on that outcome."
“conclusion": "The conclusion of the study". Must be one of: {conclusions}.
“conclusion_id": The number of the conclusion (0-indexed)
```

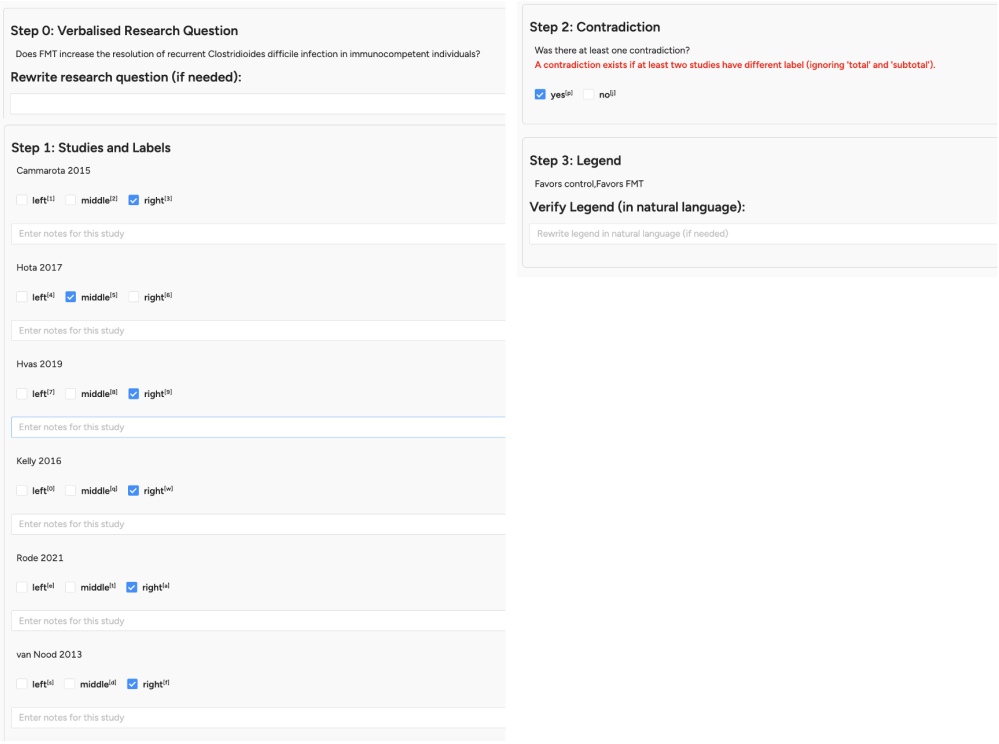
If there is no overall agreement in the context, then the result must be no statistical significance, i.e. 1.

(c)

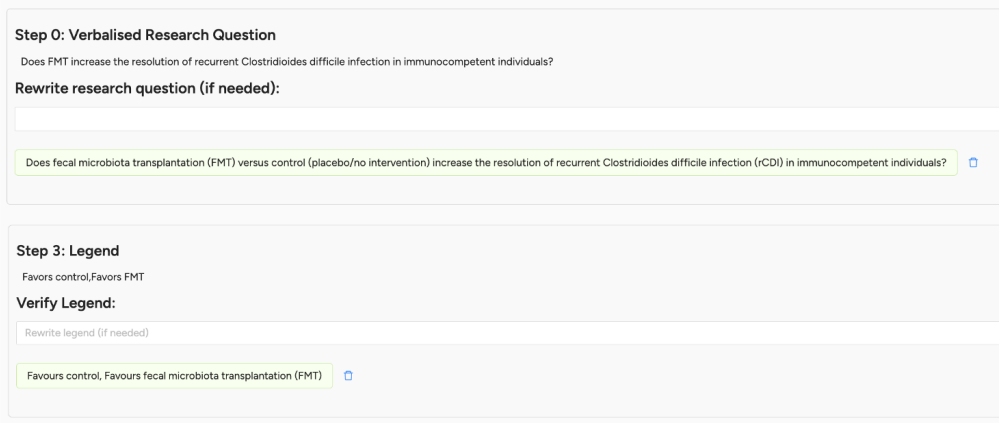
Figure 7: Prompt templates for research question generation (a), knowledge extraction (b), and answer finalisation (c).



(a) Main part of the interface, containing a visualisation of the forest plot, title, and long abstract of the systematic review.



(b) Annotation tasks.



(c) Example annotation for Task 1 (expanded acronyms and added missing comparator) and Task 3 (expanded acronyms).

Figure 8: The forest plot annotation interface.