BEYOND RELEVANCE: BROAD-TO-DEEP CONTEXTUAL EVIDENCE REASONING FOR RETRIEVAL-AUGMENTED MEDICAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical question answering increasingly relies on retrieval-augmented generation to provide large language models with external evidence, yet most existing methods still treat evidence construction as a largely static relevance-ranking problem. Inspired by how people often approach unfamiliar problems, first surveying a broad range of potentially useful references, then organizing the most informative evidence, and finally following concepts that remain unclear, we formulate evidence acquisition as a broad-to-deep process. Towards this end, we propose a novel approach named Broad-to-deep Contextual Evidence Reasoning (**BRACE**) for medical question answering with two complementary stages. The broad stage constructs a compact evidence context through determinantal resolution packing, where candidate documents are assessed by both their relevance to the question and the uncertainty they resolve, while the determinantal objective preserves complementary information across the selected set. Building on this context, the deep stage performs progressive uncertainty resolution by iteratively localizing high-uncertainty concepts within the evolving question and evidence context and converting them into targeted retrieval cues. Newly acquired evidence reshapes the context and may expose more specific information needs, allowing retrieval to progressively move from broad semantic coverage toward deeper supporting knowledge. In this way, **BRACE** lets the evolving context determine both which evidence should be retained and where subsequent retrieval should proceed. Experiments on five medical question-answering benchmarks with three open-source readers consistently demonstrate the effectiveness of our framework over strong retrieval, generation, and hybrid knowledge-augmentation baselines. Our code is available at https://anonymous.4open.science/r/BRACE-6EC4.

1 INTRODUCTION

Medical question answering (QA) is a cornerstone of biomedical natural language processing (Jin et al., 2022). Although large language models (LLMs) have advanced this task, general-purpose models still lack specialized medical knowledge and up-to-date clinical evidence. Continued pretraining or fine-tuning on medical corpora can partially alleviate this issue (Balde et al., 2024; Xie et al., 2025; Touchent et al., 2026), yet it requires high-quality annotated data and substantial computational resources, making it extremely costly. Therefore, knowledge augmentation paradigms that introduce external knowledge as auxiliary evidence offer a more flexible and interpretable alternative (Cheng et al., 2025). Among representative knowledge augmentation paradigms, retrieval-augmented generation (RAG) retrieves evidence from external corpora (Xiong et al., 2024), generation-augmented generation (GAG) leverages the model's parametric knowledge to generate background documents (Yu et al., 2023; Frisoni et al., 2024), and hybrid frameworks combine retrieved and generated evidence (Abdallah & Jatowt, 2023; Zhang et al., 2023; Li et al., 2026).

Whichever the source, most existing methods treat evidence construction as a static relevance-ranking problem. Typically, candidates are scored by criteria such as relevance, source reliability, or retrieval–generation compatibility, and the top-ranked ones form a fixed context before the reader reads anything. A preliminary experiment on MedQA (Appendix B) shows that relevance tracks usefulness on average, as single-document accuracy drops from 69.6% for ranks 1–10 to 61.7% for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ranks 21–30, but not for individual documents, as 11.3% of top-10 documents turn a correct answer wrong, while 9.6% of documents ranked 21–30 turn a wrong answer right. These results lead to our core insight: the value of evidence resides not only in the document itself, but also in the context in which it is placed. Evidence construction should therefore be a reasoning process over context rather than a one-shot ranking, which raises our central question: *How can we let the context itself decide which evidence is retained and where retrieval should proceed next?*

To address this question, we draw inspiration from how humans approach an unfamiliar problem. First, humans survey broadly and organize the most informative references. Then they follow up on what remains unclear. Thus, we propose a broad-to-deep contextual evidence reasoning framework (**BRACE**) with two stages, driven by a hierarchy of reader uncertainty: the initial uncertainty when the reader first meets a question and the residual uncertainty after it has read some evidence, which lets the reader acquire external knowledge progressively deeper and more autonomously. The *broad stage* gathers candidates from retrieval and generation and composes a compact initial context by *Determinantal Resolution Packing*, which weighs each candidate by its relevance and by the initial uncertainty it resolves while favoring complementary sets. The *deep stage* performs *Progressive Uncertainty Resolution* for a fixed number of rounds. Each round converts the concepts carrying the most residual uncertainty into retrieval cues appended to the original question, and merges the retrieved evidence into the context. The original question is retained in every query, so that retrieval does not drift from the target topic. We further prove a $(1-1/e)$ approximation guarantee for greedy packing. On five medical QA benchmarks with three open-source readers, **BRACE** achieves the best average accuracy with every reader, exceeding the strongest baseline by 4.03% on average. Our contributions are threefold:

❶ ***New Formulation of Evidence Acquisition.*** We show that evidence usefulness cannot be characterized solely by query-document relevance, and formulate evidence construction as a contextual reasoning process where the value of evidence depends on its interaction with the evolving context under dynamic uncertainty.

❷ ***Novel Framework with Theoretical Analysis.*** We propose **BRACE**, a training-free framework in which a hierarchy of reader uncertainty guides evidence construction. In the framework, Determinantal resolution packing composes a compact and complementary initial context guided by the initial uncertainty, and progressive uncertainty resolution traces the residual uncertainty over a fixed number of rounds. We prove a $(1-1/e)$ approximation guarantee for the packing.

❸ ***Empirical Validation.*** Across five benchmarks and three readers, **BRACE** achieves the best average accuracy against retrieval, generation, and hybrid-augmentation baselines, exceeding the strongest one by 4.03 percentage points on average. Ablations show that each stage contributes to the gains.

2 Related Work

Medical Question Answering. Medical question answering (QA) constitutes an important task within the field of biomedical natural language processing (Wang et al., 2025; Jin et al., 2022). Advances in large language models have spurred research into leveraging the capabilities of general-purpose models to address these tasks (Hu et al., 2025; Thirunavukarasu et al., 2023). Domain-adaptive strategies, including continued pre-training and parameter-efficient fine-tuning on medical corpora (Wu et al., 2024; Balde et al., 2024; Touchent et al., 2026), depend on high- quality annotated data and incurs substantial resource costs. By contrast, knowledge-augmented paradigms, which incorporate external knowledge as auxiliary reference material, offer a more flexible and interpretable alternative (Cheng et al., 2025).

Medical Knowledge Augmentation. The knowledge-augmented paradigm primarily comprises two representative approaches: Retrieval-augmented generation (RAG) and Generation-augmented generation (GAG). RAG obtains this information from external corpora. A representative framework MedRAG (Xiong et al., 2024) provides a comprehensive RAG benchmark and toolkit, systematically evaluating the effects of corpora, retrievers, and backbone models. Adaptive variants such as Self-RAG (Asai et al., 2024) and i-MedRAG (Xiong et al., 2025) further decide when and what to retrieve. GAG instead constructs contextual documents from a language model's parametric knowledge. Representative work includes GenRead (Yu et al., 2023) and MedGENIE (Frisoni et al., 2024). Hybrid approaches combine retrieved and generated documents within the same answering pipeline.

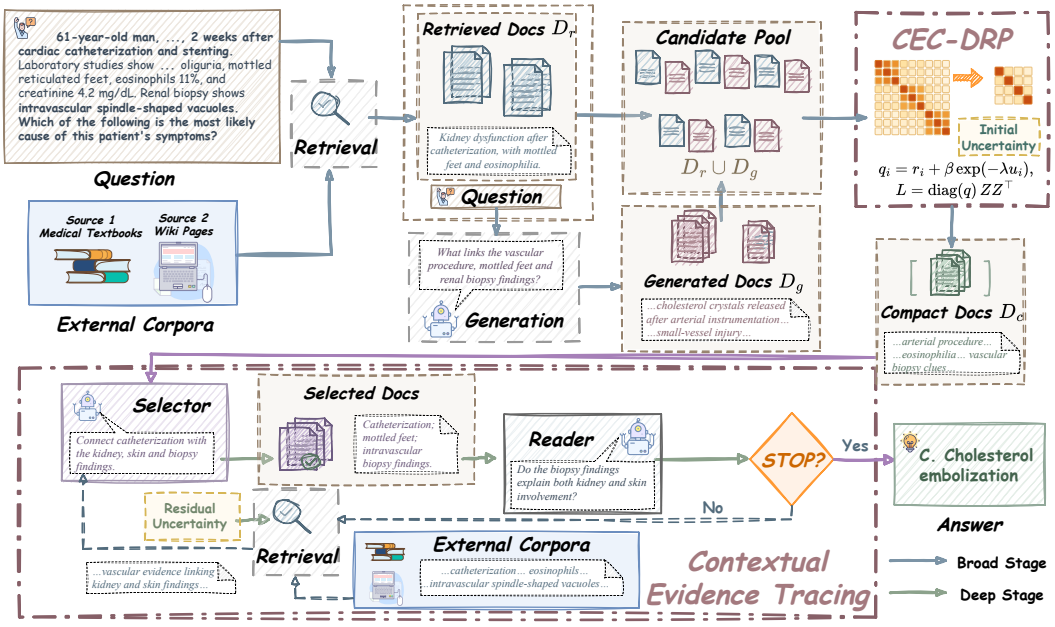

Figure 1: Overview of **BRACE**. In the broad stage, determinantal resolution packing and the selector construct a compact context from the candidate pool of retrieved and generated evidence. In the deep stage, uncertainty-guided retrieval and integration iteratively refine the context for certain rounds. The reader then answers from the updated context.

GRG (Abdallah & Jatowt, 2023) combines LLM-based context generation with dual-encoder retrieval in a single pipeline. COMBO (Zhang et al., 2023) explicitly addresses their compatibility by matching generated passages with retrieved counterparts using discriminators trained on silver compatibility labels. MedRGAG (Li et al., 2026) combines knowledge-guided context completion with knowledge-aware document selection to organize retrieved and generated information. A remaining challenge is to incorporate the downstream reader's own signals when assessing and refining the context. Our work focuses on adapting the evidence context to the downstream reader. Rather than relying solely on the origin or topical relevance of a document, we incorporate reader-derived signals into evidence selection and subsequent refinement.

3 The Proposed BRACE

3.1 Problem Definition

We consider knowledge-augmented medical question answering. Let Q denote a medical question together with its candidate options, and let $a^\star$ be its correct answer. A reader model $\mathcal{M}_r$ answers Q from an ordered evidence context D:

$$\hat{a}(D) = \mathcal{M}_r(D, Q). \tag{1}$$

Candidate evidence comes from a pool $\mathcal{P} = \{d_1, \ldots, d_n\}$ of documents from retrieval and generation processes, and further documents can be retrieved from external corpora $\mathcal{C}$. Conventional knowledge-augmentation methods score each candidate independently with a relevance function, such as a cross-encoder $\mathcal{E}$ that gives $\tilde{r}_i = \mathcal{E}(Q, d_i)$, and pass the top-ranked documents to the reader as a fixed context. Our goal is to design an evidence-selection strategy π that maps $(Q, \mathcal{P})$ to an ordered evidence subset D_π. Under a fixed evidence budget, π seeks an approximately optimal subset in terms of the reader's expected answer accuracy:

$$\pi^\star = \arg\max_{\pi} \mathbb{E}_{(Q)}\big[\mathbf{1}[\hat{a}(D_\pi) = a^\star]\big], \tag{2}$$

where $a^\star$ is the correct answer. Achieving this goal requires two decisions, concerning which evidence to retain and where to retrieve next.

3.2 Framework Overview

We design a broad-to-deep framework **BRACE**, which constructs D_π in two stages driven by a hierarchy of reader uncertainty, as shown in Figure 1. This hierarchy comprises the initial uncertainty the reader has when it first meets Q, and the residual uncertainty that remains after it has read some evidence. In the broad stage (Section 3.3), retrieval and generation provide $\mathcal{P} = D_r \cup D_g$. Each candidate d_i is weighed by an evidence utility q_i that combines its relevance $\tilde{r}_i$ with the initial uncertainty it resolves, ΔU_i. A compact, complementary subset S_g is then selected by determinantal resolution packing, which the selector model $\mathcal{M}_s$ refines into the initial context D^1. In the deep stage (Section 3.4), each round t locates the positions $\mathcal{H}^t$ of highest residual uncertainty in D^t, expands them into spans W^t, and retrieves a new batch B^t from $\mathcal{C}$ with the expanded query $\widetilde{Q}^t = Q \oplus \mathrm{Join}(W^t)$, before merging it into D^{t+1}. The original question Q is retained in every query, so that retrieval does not drift from the target topic. After a fixed number of rounds T, the reader answers from the final context D^{T+1}.

3.3 Contextual Evidence Composition via Determinantal Resolution Packing

The broad stage selects k documents from $\mathcal{P}$ to form an initial evidence set that addresses the question, covers the knowledge required to answer it, and remains compact and non-redundant. Ranking documents by relevance alone cannot achieve this goal for two reasons. First, relevance measures only semantic similarity between a document and the question. It does not indicate whether the document supplies knowledge the reader lacks, which depends on how the document interacts with the question and what the reader already knows. We call this the *question–evidence context*. Second, relevance scores each document independently and may therefore result in redundancy among highly relevant documents. We introduce Determinantal Resolution Packing (DRP), a DPP-based evidence composition mechanism that jointly considers evidence utility and complementarity. Determinantal point processes (DPP) (Kulesza & Taskar, 2012) address both limitations within a single objective. A DPP scores each subset by the determinant of a kernel submatrix, which factorizes into the utility of individual items and the diversity of the set.

We use the question as the query to retrieve documents from external corpora $\mathcal{C}$ such as medical textbooks and Wikipedia with BM25 (Robertson et al., 2009), rerank them by the relevance in Section 3.1, and retain the top k_r as retrieved evidence D_r. Since retrieval may not cover all necessary knowledge, we follow the three-step generation process of MedRGAG (Li et al., 2026). This process includes a summarizer that extracts useful information from the retrieved passages, an explorer that identifies missing knowledge points, and a generator $\mathcal{M}_g$ that produces a background document for each knowledge point, yielding D_g. The two are combined into a candidate pool $\mathcal{P} = D_r \cup D_g = \{d_1, \ldots, d_n\}$, whose size far exceeds the context that the reader can effectively utilize, and whose documents often overlap and are redundant, so a small and precise subset must be selected.

Let X denote the token sequence of (D, Q). We define the reader's uncertainty at position j as the negative log-likelihood, and measure its uncertainty about the question by averaging over all positions of the question body:

$$
\begin{aligned}
u^j(D) &= -\log p_\theta\left(X^j \mid X^{<j}\right), \\
U(D) &= \frac{1}{|\mathcal{T}_Q|} \sum_{j \in \mathcal{T}_Q} u^j(D),
\end{aligned}
\tag{3}
$$

where $\mathcal{T}_Q$ denotes the set of positions in the question body. $U(\varnothing)$ denotes the uncertainty when no evidence is provided. Intuitively, if the evidence provides the knowledge needed to answer, the reader's confidence when reading the question should increase, i.e., $U([d_i]) < U(\varnothing)$.

For each candidate d_i, we assess its evidence utility from two perspectives. The first is its relevance $\tilde{r}_i$ (Section 3.1). The second is the extent to which it resolves the reader's initial uncertainty when placed alone in the context, $\Delta U_i = U(\varnothing) - U([d_i])$. Because the two signals differ in scale, we min–max normalize them within the candidate pool of the current question. Writing $\tilde{u}_i = U([d_i])$, we obtain the normalized r_i and u_i. Since $U(\varnothing)$ is shared by all candidates of the same question, it

cancels under this normalization:

$$\frac{\Delta U_i - \min_j \Delta U_j}{\max_j \Delta U_j - \min_j \Delta U_j} = 1 - u_i, \tag{4}$$

so resolution requires only the n single-document uncertainties and no separate pass for $U(\varnothing)$. We define the evidence utility as:

$$q_i = r_i + \beta \exp(-\lambda u_i), \qquad \lambda \geq 0,\ \beta > 0 \tag{5}$$

where by Eq. (4) the exponential term increases with the normalized resolution. Documents that are relevant to the question and reduce the reader's uncertainty about it therefore receive higher utility. The additive form allows either signal to raise the evidence utility, the exponential term lies in $[\beta e^{-\lambda}, \beta]$ and keeps q_i strictly positive. λ controls the steepness of the penalty on high uncertainty, and β controls the overall contribution of the uncertainty term.

Evidence utility characterizes each document in isolation. To capture complementarity, let z_i be the ℓ_2-normalized embedding of d_i produced by a text encoder, stack these vectors as the rows of Z, and let $q = (q_1, \ldots, q_n)^\top$. We construct the kernel $L = \mathrm{diag}(q)\, ZZ^\top \mathrm{diag}(q)$. For any subset $S \subseteq \mathcal{P}$, with L_S and Z_S the submatrices indexed by S, we have:

$$\det(L_S) = \Big(\prod_{i \in S} q_i^2\Big) \det\left(Z_S Z_S^\top\right), \tag{6}$$

where the first factor rewards documents with high evidence utility and thus reflects relevance and resolution, while the second equals the squared volume spanned by the embeddings, which shrinks as documents become more similar, and thus reflects complementarity. Under the constraint $|S| = k$, we greedily maximize:

$$F(S) = \log\det\left(L_S + \epsilon_{\mathrm{dpp}} I\right), \tag{7}$$

where $\epsilon_{\mathrm{dpp}} > 0$ is a small constant that keeps the logarithm finite. Starting from the empty set, we add the candidate with the largest marginal gain until k documents are selected, and denote the result by S_g. Because the marginal gain of a document depends on the documents already selected, the same document contributes differently depending on the set it joins. The documents in S_g, sorted by relevance, are passed to the selector $\mathcal{M}_s$, an instruction-following LLM that reads them together with the question and options and retains a small subset it judges useful for answering, yielding the initial context $D^1 = \mathcal{M}_s(Q, S_g)$.

3.4 Contextual Evidence Tracing via Progressive Uncertainty Resolution

When critical information is absent from $\mathcal{P}$, packing cannot obtain new information from a fixed context, and re-retrieving with the original question only returns results similar to the first round. Moreover, the information the reader currently lacks depends on what it has already read. We call this the *evolving context*, and it is exactly what the residual uncertainty in Section 3.2 captures. Just as a person exploring new knowledge lets what remains unclear after one round of reading decide where to dig deeper, the deep stage lets the current context decide where retrieval goes next. Concepts about which the reader remains highly uncertain under the current evidence are likely to be exactly the knowledge it has not yet mastered. We introduce Progressive Uncertainty Resolution (PUR), an iterative evidence tracing mechanism that follows the reader's residual uncertainty round by round for a fixed number of rounds T.

At round t, we run one forward pass over the current context D^t, obtain $u^j(D^t)$ (Eq. (3)) at every position of (D^t, Q), and measure the residual uncertainty over all positions of the documents, question, and options, denoted $\mathcal{T}^t$. We take the $\lceil \rho |\mathcal{T}^t| \rceil$ tokens in $\mathcal{T}^t$ with the largest $u^j(D^t)$, where $\rho \in (0, 1]$ is the expansion ratio ($\rho = 0.1$ in our experiments), and denote their positions by

$$\mathcal{H}^t = \arg\max_{\mathcal{H} \subseteq \mathcal{T}^t,\ |\mathcal{H}| = \lceil \rho |\mathcal{T}^t| \rceil} \sum_{j \in \mathcal{H}} u^j(D^t). \tag{8}$$

Unlike the initial uncertainty, which averages over all question positions, the residual uncertainty keeps only the most uncertain fraction of positions. This phenomenon reflects a cognitive process in which, following the formation of an initial understanding of a question, global uncertainty regarding the question is resolved, leaving only a few specific unresolved points. These points, in turn,

determine the direction of subsequent inquiry. The tokens at $\mathcal{H}^t$ play exactly this role. These tokens are often subword pieces, so we use the tokenizer's character offsets to expand them into complete words and merge adjacent words into a span list W^t. The expanded query is

$$\widetilde{Q}^t = Q \oplus \mathrm{Join}(W^t), \tag{9}$$

where $\oplus$ denotes text concatenation and $\mathrm{Join}(\cdot)$ concatenates the spans with a delimiter. The original question is retained so that retrieval does not drift from the target topic, while the spans provide additional discriminative cues. We retrieve documents from $\mathcal{C}$ with $\widetilde{Q}^t$, remove those already seen, and take the top b as the new batch B^t. The selector then incorporates them into the current context under the original question, yielding $D^{t+1} = \mathcal{M}_s(Q, D^t \oplus B^t)$. As the context changes, so does the reader's residual uncertainty in the next round, and retrieval moves from the broad topic of the question toward increasingly specific supporting knowledge.

Algorithm 1 The framework of `BRACE`

Require: Question set $\mathcal{Q}$, corpus $\mathcal{C}$, retriever $\mathcal{R}$, generator $\mathcal{M}_g$, selector $\mathcal{M}_s$, reader $\mathcal{M}_r$, cross-encoder $\mathcal{E}$, maximum rounds T
Ensure: Predicted answers $\widehat{\mathcal{A}}$

for each question $Q \in \mathcal{Q}$ **do**
  $D_r \leftarrow \mathcal{R}(\mathcal{C}, Q)$
  $D_g \leftarrow \mathcal{M}_g(D_r, Q)$
  $S_g \leftarrow \mathrm{DRP}(D_r \cup D_g, Q)$ `// Determinantal resolution packing`
  $D_Q^1 \leftarrow \mathcal{M}_s(Q, S_g)$ `// Form initial context`
  for $t = 1$ to T **do**
    Construct the expanded query $\widetilde{Q}^t$ according to Eq. 8 and Eq. 9.
    $B^t \leftarrow \mathcal{R}(\mathcal{C}, \widetilde{Q}^t)$ `// Top-`b `unseen documents`
    $D_Q^{t+1} \leftarrow \mathcal{M}_s(Q, D_Q^t \oplus B^t)$ `// Merge new evidence into context`
  end for
end for
$\widehat{\mathcal{A}} \leftarrow \{\mathcal{M}_r(D_Q^{T+1}, Q) : Q \in \mathcal{Q}\}$
return $\widehat{\mathcal{A}}$

3.5 Theoretical Guarantees

This section provides a theoretical analysis of our framework; proofs are given in Appendix A. For Contextual Evidence Composition, fix a candidate pool $\mathcal{P}$ of size n, and keep the utility weights q_i and unit embeddings z_i fixed during selection. Let $\varepsilon = \epsilon_{\mathrm{dpp}} > 0$, and shift the objective function F in Equation 7 to:

$$G(S) = F(S) - |S| \log \varepsilon. \tag{10}$$

This shift is constant on subsets of equal size, so it neither changes any greedy choice nor any maximizer of F over subsets of size k. Let S_g be the set of size k returned by the DRP greedy rule, and $S^\star$ be the set that maximizes F over all subsets of $\mathcal{P}$ of size k, and let $\alpha_k = 1 - (1 - 1/k)^k$.

Theorem 3.1. *Let $\varepsilon > 0$ and $1 \le k \le n$. Then G is nonnegative, monotone, and submodular, and*

$$G(S_g) \ge \alpha_k \, G(S^\star), \tag{11}$$

where $\alpha_k \ge 1 - e^{-1}$.

Theorem 3.1 instantiates the classical greedy guarantee of Nemhauser et al. (1978) to the regularized DRP objective. Therefore, greedy selection achieves a constant fraction of $G(S^\star)$ (the value of the optimal subset of size k) without exhaustive search.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on five medical question answering benchmarks spanning examination-based reasoning and biomedical literature understanding: MedQA-US (Jin et al.,

Table 1: Main experimental results (accuracy, %). Best results are in **bold**, and second-best results are underlined.

Reader	Method	MedQA-US	MedMCQA	MMLU-Med	PubMedQA*	BioASQ-Y/N*	Average
Qwen2.5-7B	Direct Response	62.37	55.53	75.76	37.40	73.30	60.87
	Vanilla RAG	61.12	57.33	75.94	34.20	72.01	60.12
	MedRAG	62.61	58.64	76.12	38.00	76.38	62.35
	GenRead	68.89	60.24	78.97	46.00	79.29	66.68
	MedGENIE	69.36	59.91	78.60	49.20	81.88	67.79
	GRG	69.84	60.89	78.24	46.60	81.39	67.39
	MedRGAG	75.49	66.84	84.57	48.00	84.47	71.87
	Ours	**85.94**	**70.26**	**87.70**	**49.40**	**84.79**	**75.62**
Llama-3.1-8B	Direct Response	67.48	58.45	75.48	55.40	76.38	66.64
	Vanilla RAG	67.24	58.38	75.85	50.20	73.30	64.99
	MedRAG	68.42	59.32	76.95	52.00	75.24	66.39
	GenRead	70.07	60.89	78.15	54.60	78.32	68.41
	MedGENIE	68.81	60.24	78.24	56.60	80.26	68.83
	GRG	68.97	60.89	76.77	57.00	82.36	69.20
	MedRGAG	77.30	66.82	84.66	56.60	81.88	73.45
	Ours	**86.33**	**70.88**	**87.33**	**59.00**	**84.14**	**77.54**
Ministral-8B	Direct Response	57.27	51.23	71.63	31.80	72.65	56.92
	Vanilla RAG	60.64	54.96	71.44	26.20	72.49	57.15
	MedRAG	62.29	56.80	76.22	29.80	74.11	59.84
	GenRead	69.36	59.96	77.04	41.60	78.80	65.35
	MedGENIE	68.03	59.45	78.51	42.00	82.36	66.07
	GRG	67.24	60.08	77.41	42.60	81.55	65.78
	MedRGAG	74.86	66.03	83.29	42.00	84.95	70.23
	Ours	**86.17**	**70.40**	**85.86**	**44.80**	**85.11**	**74.47**

2021), MedMCQA (Pal et al., 2022), MMLU-Med (Hendrycks et al., 2021) , PubMedQA* , and BioASQ-Y/N. PubMedQA* denotes the context-removed variant of the original PubMedQA (Jin et al., 2019) dataset. BioASQ-Y/N is constructed by selecting questions with binary answer options (yes/no) from the original BioASQ (Tsatsaronis et al., 2015) benchmark. More details are presented in Appendix C.1.

Baselines. We compare our method with seven baselines covering direct answering, retrieval-based augmentation, generation-based augmentation, and mixed-source augmentation. Direct answering uses only the question and its candidate options without additional contexts. Retrieval-based augmentation gathers supporting medical evidence from external corpora, including Vanilla RAG (Lewis et al., 2020) and MedRAG (Xiong et al., 2024). Generation-based augmentation creates auxiliary context from model's internal knowledge, including GenRead (Yu et al., 2023) and MedGENIE (Frisoni et al., 2024) Mixed-source augmentation, such as GRG (Abdallah & Jatowt, 2023) and MedRGAG (Li et al., 2026) , draws on both generated and retrieved content to construct integrated context. Additional details are provided in Appendix C.2.

Implementation Details. The retrieval and generation steps each contribute 15 documents, together forming the initial candidate pool. Determinantal Resolution Packing selects 10 passages, after which the selector retains at most five for the reader. We adopt Qwen2.5-7B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Ministral-8B (Mistral AI Team, 2024) as readers $\mathcal{M}_r$ and Llama-3.1-8B as the background-document generator $\mathcal{M}_g$. We set the uncertainty penalty to $\lambda = 0.5$ and uncertainty contribution $\beta = 0.1$. We use a top-10% expansion ratio for contextual evidence tracing. Each refinement update retrieves five new passages. The number of refinement rounds T is capped at 3. Additional implementation details are provided in Appendix C.3.

4.2 Performance Comparison

Table 1 presents the main results on five medical QA benchmarks. Our method achieves the highest average accuracy across all benchmarks and consistently outperforms all baselines under three different reader architectures. We highlight three key observations: ***Obs. ❶: Across the five bench-***

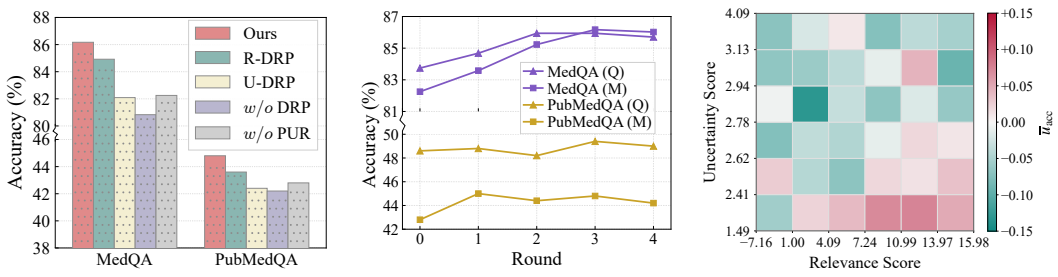

Figure 2: Comprehensive analysis of the proposed method. **Left:** Ablation study on the contribution of key components. **Middle:** Accuracy across refinement rounds under a fixed-round schedule. **Right:** Utility heatmap of uncertainty and relevance scores.

marks and three reader models, our method attains the best average accuracy. This advantage is consistent, indicating that the combination of DRP and PUR generalizes across heterogeneous medical QA tasks and reader architectures. ***Obs. ❷: Compared with the direct-response setting, our method yields an average absolute improvement of 14.40%.*** This gap confirms that external evidence remains essential for medical QA and that our framework delivers it effectively. ***Obs. ❸: Our method outperforms all three families of knowledge-augmented baselines***: retrieval-augmented methods (Vanilla RAG, MedRAG), generation-augmented methods (GenRead, MedGENIE), and hybrid methods(GRG, MedRGAG). Notably, even against the strongest baseline MedRGAG, our method achieves further average absolute gains of 3.75–4.24% (mean 4.03%) across the three reader architectures. It reflects the benefit of reader-adaptive, complementarity-aware evidence selection over static fusion of heterogeneous sources. Overall, these results validate the central hypothesis of this work. A compact, complementary, reader-adaptive evidence set, constructed by Determinantal Resolution Packing and iteratively refined by Progressive Uncertainty Resolution, consistently improves answer accuracy over strong knowledge-augmented baselines.

4.3 Ablation Study

We conduct ablation studies on Ministral-8B over MedQA and PubMedQA to examine the contributions of the key components in our framework. We consider four variants: (1) **R-DRP** retains the DPP diversity component but uses relevance alone as the quality weight, setting $q_i = r_i$. (2) **U-DRP** retains the same diversity component but uses only reader-conditioned uncertainty to determine quality, setting $q_i = \exp(-\lambda u_i)$. (3) ***w/o* DRP** bypasses DRP and directly supplies the mixed retrieved–generated candidate pool to the selector. (4) ***w/o* PUR** disables iterative evidence refinement and produces the answer from the initial context selected by DRP and the selector. The compared results are in Figure 2 (left). From the results, we can observe that the full method outperforms all four ablated variants in average accuracy, supporting the contribution of each examined component. The degradation of *w/o* DRP indicates the benefit of selecting the mixed-source evidence before selection. The lower performance of R-DRP and U-DRP suggests that retrieval relevance and reader-conditioned uncertainty provide complementary signals for assessing candidate quality. Finally, the decline of *w/o* PUR demonstrates the benefit of refining the evidence beyond the initial context. Together, these compared results sufficiently support combining reader-aware evidence selection with iterative refinement.

4.4 Further Analysis

Sensitivity Analysis. We analyze accuracy versus refinement rounds on Qwen2.5-7B and Ministral-8B over MedQA and PubMedQA (Figure 2 Middle). Across all datasets and readers, accuracy generally improves during the early refinement rounds. This indicates that additional retrieval can supply useful evidence beyond the initial context, especially for questions whose key concepts are not fully covered by the broad stage. However, the gains diminish as the number of rounds increases. The curve reaches a peak at a moderate number of rounds and then plateaus or slightly declines. This pattern is consistent with the intuition that the reader's residual uncertainty decreases as more relevant evidence is incorporated, so that further retrieval yields diminishing marginal benefit, while excessive refinement may introduce redundant or weakly related evidence that dilutes the

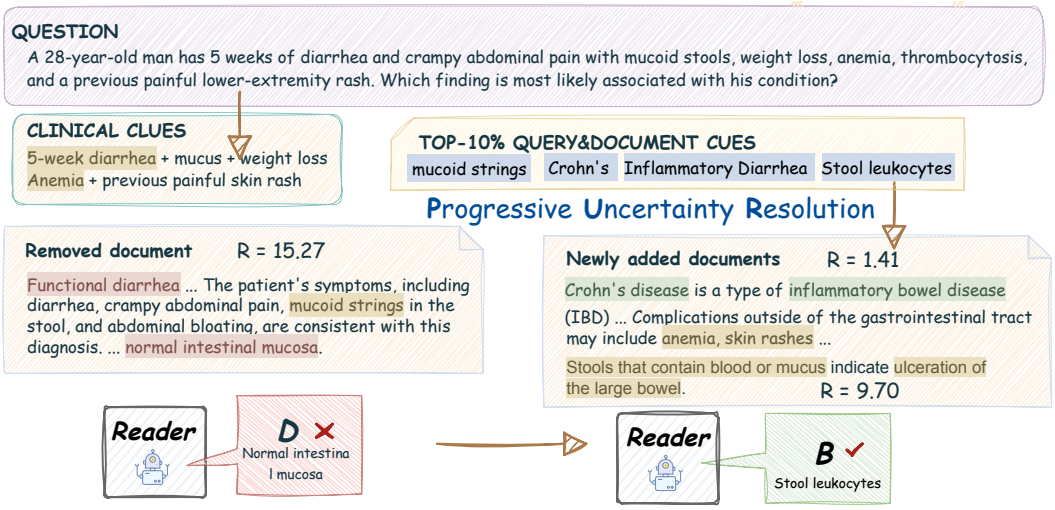

Figure 3: Case study showing how PUR effectively corrects a misleading initial context by using residual uncertainty to retrieve discriminative evidence.

context.We therefore adopt a moderate number of rounds as a fixed budget to keep the evidence tracing procedure simple and effective.

Evidence Utility Validation Beyond Relevance. To examine whether reader-conditioned uncertainty provides complementary information beyond relevance ranking, we conduct a preliminary study on MedQA. Figure 2 (Right) shows the average single-document utility (Eq. 19) across relevance and uncertainty regions, where each cell aggregates question-document pairs from the preliminary study (Appendix B) within the corresponding relevance and uncertainty bins. The results reveal that relevance correlates with utility mainly under low-uncertainty conditions, while highly relevant documents may still have limited utility when uncertainty remains high. This suggests that relevance alone is insufficient for characterizing evidence usefulness, and reader-conditioned uncertainty provides complementary signals for identifying unresolved information needs. Based on this observation, DRP jointly considers relevance and uncertainty for evidence composition, while PUR uses residual uncertainty to guide targeted retrieval.

Case Study. We present a representative case to illustrate how progressive uncertainty resolution corrects a misleading initial context. As shown in Figure 3, in Case #0109 from MedQA, the initial evidence strongly overlaps with the vignette but supports a misleading diagnosis of functional diarrhea, favoring the incorrect option "normal intestinal mucosa." Newly added documents, retrieved via high-uncertainty spans, reorganize the same observations into an inflammatory interpretation: chronic diarrhea, mucus, abdominal pain, weight loss, anemia, and prior skin lesion jointly suggest Crohn's disease. The evidence chain links *Crohn/IBD* to *fecal leukocytes*, changing the answer to the correct one. This shows that high relevance does not guarantee high utility and that residual uncertainty can retrieve discriminative evidence. More cases are in Appendix D.

5 Conclusion

In this paper, we studied how to formulate evidence acquisition as a broad-to-deep contextual reasoning process in medical question answering. We proposed a two-stage framework that transforms a heterogeneous candidate pool into a compact, complementary, and reader-adaptive evidence set. Specifically, we introduced Determinantal Resolution Packing (DRP), a DPP-based evidence composition mechanism that jointly considers evidence utility and complementarity. We also introduced Progressive Uncertainty Resolution (PUR), an iterative evidence tracing mechanism that follows the reader's residual uncertainty round by round. Experiments on five medical QA benchmarks with multiple reader models demonstrate consistent improvements over strong baselines. Our approach relies on the reader's token-level uncertainty, which requires access to the reader's output distribution. Plus, the iteration procedure may introduce additional computation. Future work includes extending the framework to other knowledge-intensive QA tasks beyond the medical domain.

AI USE STATEMENT

In this work, we used generative AI tools to polish the paper to improve readability, including correcting grammar, refining wording, and improving the clarity of paragraphs written by the authors. We did not use generative AI tools to propose or revise hypotheses, design the proposed methods or experimental protocols, implement the methods, generate, clean, or reorder datasets, or interpret experimental results. We also did not use generative AI tools to propose or prove mathematical propositions. We reviewed all AI-assisted work. Every AI-suggested passage was checked by the authors to ensure that it retained its original technical meaning; all numbers, claims, and references in the paper were verified against our experimental records and original sources. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the complete proofs of our theoretical results in Appendix A, the preliminary study in Appendix B, detailed dataset descriptions, baseline configurations, and implementation details in Appendix C, and representative case studies in Appendix D. We specify model checkpoints and revisions, retrieval corpora, BM25 and MedCPT configurations, generation parameters, uncertainty ratios, and evaluation metrics in the appendices. All local experiments use publicly available model checkpoints (Qwen2.5-7B, Llama-3.1-8B, and Ministral-8B). The selector is accessed via a fixed API version. Code for reproducing the experiments and figures will be released publicly upon acceptance.

REFERENCES

Abdelrahman Abdallah and Adam Jatowt. Generator-retriever-generator approach for open-domain question answering. *arXiv preprint arXiv:2307.11278*, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. Medvoc: Vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pp. 6180–6188. International Joint Conferences on Artificial Intelligence Organization, 2024.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. A survey on knowledge-oriented retrieval-augmented generation. *ACM Transactions on Information Systems*, 2025.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9878–9919. Association for Computational Linguistics, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Zhenlin Hu, Zhizhi Peng, Zhen Bi, Qing Shen, Zhenfang Liu, Jungang Lou, and Xin Luo. Advancing healthcare with large language models: Techniques and application. *IEEE/CAA Journal of Automatica Sinica*, 12(12):2371–2398, 2025.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 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)*, pp. 2567–2577, 2019.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

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

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

Lei Li, Xiao Zhou, Yingying Zhang, and Xian Wu. From retrieval to generation: Unifying external and parametric knowledge for medical question answering. In *Proceedings of the ACM Web Conference 2026*, pp. 2419–2430, 2026.

Mistral AI Team. Un ministral, des ministraux, 2024. URL https://mistral.ai/news/ministraux.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025. URL https://arxiv.org/abs/2412.15115.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Rian Touchent, Nathan Godey, and Éric Clergerie. Biomed-enriched: Data-efficient biomedical pretraining via paragraph-level annotation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34276–34287, 2026.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

Yinuo Wang, Baiyang Wang, Robert Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. Trustworthy medical question answering: An evaluation-centric survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27477–27490, 2025.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ digital medicine*, 8(1):141, 2025.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251. Association for Computational Linguistics, 2024.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214. World Scientific, 2025.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4710–4728. Association for Computational Linguistics, 2023.

A Proof of Theoretical Results

A.1 Proof of the Greedy Selection Guarantee

Proof of Theorem 3.1. Let d_z be the embedding dimension, write $a_i = q_i z_i \in \mathbb{R}^{d_z}$, and let A_S be the matrix with rows $a_i^\top$ for $i \in S$, so that $L_S = A_S A_S^\top$ by the definition of L. Define

$$B_S = I_{d_z} + \varepsilon^{-1} \sum_{i \in S} a_i a_i^\top \succ 0. \tag{12}$$

Sylvester's determinant identity gives

$$\begin{aligned} \det\big(L_S + \varepsilon I_{|S|}\big) &= \varepsilon^{|S|} \det\big(I_{|S|} + \varepsilon^{-1} A_S A_S^\top\big) \\ &= \varepsilon^{|S|} \det B_S, \end{aligned} \tag{13}$$

hence $G(S) = \log\det B_S$ and $G(\varnothing) = 0$. Since $B_{S\cup\{i\}} = B_S + \varepsilon^{-1} a_i a_i^\top$ for $i \notin S$, the matrix determinant lemma yields

$$G(S \cup \{i\}) - G(S) = \log\big(1 + \varepsilon^{-1} a_i^\top B_S^{-1} a_i\big) \geq 0, \tag{14}$$

so G is nonnegative and monotone. Let $S \subseteq V$ and $i \notin V$. Then $B_V \succeq B_S \succ 0$, and inversion reverses the positive semidefinite order, so $a_i^\top B_V^{-1} a_i \leq a_i^\top B_S^{-1} a_i$. The marginal-gain identity then gives

$$G(S \cup \{i\}) - G(S) \geq G(V \cup \{i\}) - G(V), \tag{15}$$

which is submodularity.

The definition of G gives, for $i \notin S$,

$$F(S \cup \{i\}) - F(S) = G(S \cup \{i\}) - G(S) + \log\varepsilon, \tag{16}$$

so the DRP greedy rule is also greedy for G. Let S_j be the greedy set after j additions, with $S_0 = \varnothing$. For $0 \leq j < k$, monotonicity and submodularity give

$$\begin{aligned} G(S^\star) - G(S_j) &\leq G(S^\star \cup S_j) - G(S_j) \\ &\leq \sum_{i \in S^\star \setminus S_j} [G(S_j \cup \{i\}) - G(S_j)] \\ &\leq k\,[G(S_{j+1}) - G(S_j)]\,. \end{aligned} \tag{17}$$

With $E_j = G(S^\star) - G(S_j)$, the preceding inequality reads $E_{j+1} \leq (1 - 1/k) E_j$. Since $E_0 = G(S^\star)$, iterating over j gives $E_k \leq (1 - 1/k)^k G(S^\star)$. Since $S_g = S_k$,

$$\begin{aligned} G(S_g) &= G(S^\star) - E_k \\ &\geq \big[1 - (1 - 1/k)^k\big]\, G(S^\star) = \alpha_k G(S^\star). \end{aligned} \tag{18}$$

For $k = 1$, $\alpha_1 = 1$. For $k > 1$, $\log(1 - 1/k) \leq -1/k$ gives $(1 - 1/k)^k \leq e^{-1}$, hence $\alpha_k \geq 1 - e^{-1}$. □

B Preliminary Experiment

This preliminary experiment examines the extent to which relevance ranking reflects the actual utility of a single document for the reader. We use the first 200 questions of MedQA-US, each with the full set of 30 candidate documents (15 retrieved passages and 15 generated documents). Using the question stem and all options as the query, we rescore the candidates with MedCPT-Cross-Encoder and sort them in descending order. We use Qwen2.5-7B-Instruct as the reader and run a per-document condition, evaluating 6,000 question–document pairs in total.

Given a question Q and the correct answer $a^\star$, let $\widehat{a}(D)$ denote the option generated by the reader under evidence context D. We use the change in correctness as a discrete utility:

$$u_{\mathrm{acc}}(d) = \mathbf{1}[\widehat{a}([d]) = a^\star] - \mathbf{1}[\widehat{a}(\varnothing) = a^\star]\,, \tag{19}$$

Table 2: Single-document utility across MedCPT ranking intervals. The last column reports the average accuracy when a single document is provided.

MedCPT Rank	Correct→Incorrect	Incorrect→Correct	Single-Doc Accuracy
1–10	226/2,000 (11.30%)	268/2,000 (13.40%)	69.60%
11–20	278/2,000 (13.90%)	258/2,000 (12.90%)	66.50%
21–30	308/2,000 (15.40%)	192/2,000 (9.60%)	61.70%

where -1 indicates a correct-to-incorrect flip, $+1$ an incorrect-to-correct flip, and 0 no change in correctness.

As the results in Table 2 present, higher-ranked documents correspond to higher single-document accuracy on average, but this trend does not guarantee the utility of an individual document. Among rank 1–10 documents, 11.30% still turn a previously correct answer into an incorrect one; among rank 21–30 documents, 9.60% still turn a previously incorrect answer into a correct one. At the significance level of $\alpha = 0.05$, the paired t-test indicates that the correction rates of ranks 1–10 and 11–20 are not significantly different($p = 0.5584$). These observations directly motivate the design of our method. Because relevance alone is not a sufficient statistic for evidence utility, Contextual Evidence Composition incorporates reader-derived uncertainty into the document-quality term while encouraging set-level complementarity.

C Detailed Experimental Settings

C.1 Dataset Details

We evaluate our method on five representative medical QA benchmarks covering clinical reasoning, biomedical knowledge, and professional medical understanding.

MedQA-US (Jin et al., 2021) is derived from the United States Medical Licensing Examination (USMLE), focusing on diagnosis, treatment, and clinical reasoning. It contains 1,273 test questions covering both Step 1 (basic science) and Step 2&3 (clinical application) question types. Each question consists of a medical scenario with four answer options, requiring models to integrate medical knowledge and perform exam-style reasoning.

MedMCQA (Pal et al., 2022) is a large-scale multi-subject medical QA benchmark collected from Indian medical entrance examinations, including AIIMS and NEET-PG. It covers 21 medical subjects and more than 2,400 healthcare topics. Since the test labels are unavailable, we use the official validation split containing 4,183 four-option questions. This dataset evaluates the ability of models to handle diverse medical knowledge domains across a broad range of subjects.

MMLU-Med (Hendrycks et al., 2021) is the medical subset of the Massive Multitask Language Understanding (MMLU) benchmark. It comprises 1,089 four-option questions spanning six medical-related subjects: anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine. It evaluates medical reasoning across both foundational biomedical knowledge and professional-level clinical understanding.

PubMedQA* (Jin et al., 2019) is built from PubMed abstracts with expert-annotated biomedical research questions. The original task requires models to answer scientific questions using corresponding abstracts as supporting evidence. To match the RAG setting, we remove the original supporting contexts and use the 500 official test samples for evaluation. Each question contains three answer options: yes, no, and maybe.

BioASQ-Y/N (Tsatsaronis et al., 2015) is constructed from the BioASQ Task B challenge, which provides expert-annotated biomedical questions. We extract questions with binary yes/no answers from the gold-standard test sets over the five most recent years (2019–2023), resulting in 618 questions. This dataset provides an additional evaluation scenario for biomedical information retrieval and evidence-based answering.

C.2 Baseline Details

We compare BRACE against representative baselines below. The baselines span from no-evidence answering to multi-source evidence fusion, allowing us to isolate the contribution of each knowledge-augmentation strategy. For fair comparison, all methods use the same reader backbone and retrieval corpora unless otherwise specified.

Direct Response. The reader answers using only the question and candidate options without additional context. This baseline measures the parametric medical knowledge of the reader and serves as a reference for assessing the benefit of knowledge augmentation.

Vanilla RAG. Vanilla RAG (Lewis et al., 2020) implements the standard retrieve-then-read pipeline: a dense retriever encodes the question and retrieves the top-k passages from a corpus via vector similarity, and a reader model generates the answer conditioned on the concatenated passages. In our implementation, it uses the same medical corpora as the other retrieval-based baselines and involves no medical adaptation, re-ranking, or evidence selection.

MedRAG. MedRAG (Xiong et al., 2024) extends Vanilla RAG with a medical-domain retrieval framework. In its default configuration, it combines sparse retrieval (BM25) and dense retrieval (Contriever) over MedCorp, and fuses their results via Reciprocal Rank Fusion (RRF) to improve retrieval robustness and coverage. It also provides medical knowledge graphs and question-type-specific prompts. We follow the default retrieval and fusion settings, while keeping the reader and selector identical to the other methods.

GenRead. GenRead (Yu et al., 2023) introduces the generate-then-read paradigm, replacing document retrieval with LLM-based context generation. Given a question, the LLM first generates a set of contextual documents from its parametric knowledge, and a reader then produces the final answer conditioned on these generated documents. It uses no external corpus, so its performance depends on the generator's internal knowledge. We follow the original prompting strategy and generate the same number of contexts as the retrieved passages used by retrieval-based baselines.

MedGENIE. MedGENIE (Frisoni et al., 2024) is a generation-augmented framework for medical multiple-choice QA. It employs a specialized medical LLM, PMC-LLaMA-13B (Wu et al., 2024), to produce multi-view artificial contexts from perspectives such as disease mechanisms, clinical manifestations, diagnostic criteria, and treatment options. These contexts are then combined with the question to prompt the reader. We use the official pipeline with PMC-LLaMA-13B as the generator and keep the reader and selector consistent with other baselines.

GRG. GRG (Abdallah & Jatowt, 2023) adopts a generator–retriever–generator pipeline. The LLM first generates hypothetical documents conditioned on the question, while a dual-encoder retriever retrieves relevant passages from an external corpus in parallel. The generated and retrieved documents are then passed to a second LLM to produce the final answer. We follow the original GRG pipeline, using the same retriever and corpora as the other retrieval-based baselines, and keep the second-stage reader identical across methods.

MedRGAG. MedRGAG (Li et al., 2026) is a unified retrieval–generation framework with two core modules: Knowledge-Guided Context Completion (KGCC) and Knowledge-Aware Document Selection (KADS). KGCC analyzes retrieved documents and steers the generator to produce background documents that complement the gaps, while KADS adaptively selects an optimal combination of retrieved and generated documents for the reader. It is the most direct predecessor of our method, but does not use reader-derived uncertainty, iterative refinement, or a submodular utility objective. We reimplement it under the same auxiliary-model configuration as our method, replacing the original GPT-4o-mini selector with GPT-5.6-Luna for fair comparison.

C.3 Implementation Details

We employ BM25 (Robertson et al., 2009) to retrieve 32 documents from medical textbooks (Jin et al., 2021) and Wikipedia articles (Xiong et al., 2024) respectively. Following the experimental

protocol of MedRGAG (Li et al., 2026), we employ an auxiliary LLM for evidence organization and selection. Since GPT-4o-mini used in the original MedRGAG implementation was unavailable in our experimental environment, we replace it with GPT-5.6-Luna. For a fair comparison, we consistently apply the same replacement when reimplementing MedRGAG and use GPT-5.6-Luna as the selector M_s in BRACE. All methods are therefore evaluated under the same auxiliary-model configuration. A MedCPT (Jin et al., 2023) cross-encoder is applied to retain the top 15 passages. Together with 15 generated passages, these form the initial candidate pool. For DRP, the numerical stabilizers for relevance and uncertainty min–max normalization are both 10^{-8}, and the log-determinant regularization coefficient ϵ_{dpp} is 10^{-6}. The background generator uses a temperature of 1.2, top-p of 0.9, top-k of 50, and a presence penalty of 1.0. The reader uses a temperature of 0.2, top-p of 1.0, and no top-k truncation, with both the engine-level and request-level random seeds set to 42. The selector uses a temperature of 1.2. Local language-model inference uses bfloat16 precision and is implemented with vLLM (Kwon et al., 2023) and PyTorch (Paszke et al., 2019) using eager execution.

D Case Study

To examine how uncertainty-guided iteration refines the evidence context for medical question answering, we present another two representative examples in which an initially incorrect answer becomes stably correct. For clarity, we focus on the documents removed by the iterative process and the corresponding newly added documents. The two cases in 3 further illustrate how top-10% uncertainty-guided iteration refines the evidence context.

In Case #0074, the initial evidence had already identified tinea versicolor, but ambiguity remained in the diagnostic evidence because Wood lamp examination was repeatedly presented as a plausible confirmatory test in the initial evidence set. The newly added documents construct a more task-specific evidence chain by linking hypopigmented truncal lesions and *Malassezia* to the characteristic findings on KOH preparation. Thus, this round of evidence refinement primarily improves *task alignment* by strengthening the evidence for the specific confirmatory test requested by the question.

Finally, Case #0159 illustrates the refinement of an unresolved *contrastive relation*. Before iteration, both Bartter and Gitelman syndromes were already represented in the evidence, and several findings, such as hypokalemia and elevated renin/aldosterone, were compatible with both conditions. The newly added documents explicitly contrast *Bartter* $\rightarrow$ *hypercalciuria* with *Gitelman* $\rightarrow$ *hypocalciuria*, and further identify urinary calcium as a discriminative feature for distinguishing the two conditions. In this case, the iterative process recovers the *relational information required to distinguish between two competing options*.

Taken together, these cases suggest that the role of the uncertainty signal is not merely to increase medical-keyword coverage. Iterative retrieval improves the **answer-discriminative structure of the evidence** by strengthening task-specific support, replacing misleading high-similarity contexts with more useful evidence, and recovering explicit contrastive relations between candidate answers. Overall, these successful trajectories provide qualitative evidence for the complete query–retrieval–selection–answering pipeline.

Table 3: Case studies of Progressive Uncertainty Resolution (PUR). Yellow highlighting denotes information useful for determining the answer, green indicates evidence directly supporting the correct answer, and red represents evidence biasing toward an incorrect answer.

(a) Case 1: #0074 from MedQA

Question: A 21-year-old man comes to the physician because of pruritus and a hypopigmented rash on his upper body for 5 days. He first noticed the symptoms after returning from a business trip last week in the Bahamas. While he was there, he visited a couple of beaches and went hiking with some coworkers. The rash initially started as a single lesion on his upper back but since then has extended to his shoulders. He has a history of type 1 diabetes mellitus controlled with an insulin pump. He works as an office manager and has no known exposure to melanocytotoxic chemicals. He has been sexually active with three female partners over the past year and uses condoms inconsistently. He is 183 cm (6 ft) tall and weighs 80 kg (176 lb); BMI is 23.9 kg/m^2. His temperature is 37.2°C (99°F), pulse is 78/min, and blood pressure is 130/84 mm Hg. A photograph of the rash is shown. One month ago, his hemoglobin A1C was 7.8%. Which of the following is most likely to confirm the diagnosis?

Options: **A.** Wood lamp examination **B.** Skin culture **C.** Potassium hydroxide preparation **D.** Skin biopsy

Related words in the refined query: Hypopigmented Truncal Eruptions; malassezia; Diagnostic Approaches; KOH; Diagnostic Findings; Diagnosis Tinea.

Removed Documents

Doc 1

Hypopigmented Truncal Eruptions: A Comprehensive Approach. Hypopigmented truncal eruptions refer to skin lesions that exhibit a decrease in skin pigmentation, typically occurring on the torso... it is essential to consider conditions such as tinea versicolor, pityriasis alba, and post-inflammatory hypopigmentation... exposure to fungi, particularly Malassezia species, may be relevant... **Diagnostic Approaches for Hypopigmented Truncal Eruptions:** Wood lamp examination (A)...

Doc 2

Dermatological Background: Pruritic Hypopigmented Rash. A pruritic hypopigmented rash is characterized by patches of skin that lose their pigmentation and become itchy... Pityriasis Versicolor (PV) is a common fungal infection caused by Malassezia yeast, characterized by hypopigmentation and pruritus... **Diagnostic Methods:** Wood lamp examination is useful in diagnosing fungal infections such as pityriasis versicolor, but is not specific enough to confirm the diagnosis without clinical correlation...

Newly Added Documents

Doc 1

Patients present with small, scaly patches of varying color, usually on the chest or back. Lesions may be hypopigmented as a result of interference with melanin production... Diagnosed by clinical impression, and confirmed by potassium hydroxide (KOH) preparation of scale that reveals a "spaghetti and meatballs" pattern of hyphae and spores...

Doc 2

Tinea versicolor is caused by a nondermatophytic, dimorphic fungus, Malassezia furfur... The typical lesions are concentrated on the chest, shoulders, and back... On dark skin the lesions often appear as hypopigmented areas... A KOH preparation from scaling lesions will demonstrate a confluence of short hyphae and round spores ("spaghetti and meatballs")...

Answer Transition: A. Wood lamp examination → C. Potassium hydroxide preparation ✓

Table 3: Case studies of PUR (continued). Yellow highlighting denotes information useful for determining the answer, green indicates evidence directly supporting the correct answer, and red represents evidence biasing toward an incorrect answer.

(b) Case 2: #0159 from MedQA

Question: A 9-year-old boy is brought to a physician by his mother for evaluation of generalized weakness, increased urination, and increased thirst. The mother mentions that her boy is always tired, and seems to be getting worse. He prefers watching television rather than going out and playing with other kids. Furthermore, he has had frequent episodes of constipation since birth. Except for frequent cravings for salty foods, the boy eats a regular diet. The patient was delivered healthy and is fully immunized. The medical history is unremarkable, and he takes no medications. He has no siblings. His father is a banker and his mother is a librarian. The pulse is 90/min, the blood pressure is 110/75 mm Hg, and the respiratory rate is 15/min. He is in the bottom 10th percentile for height and weight according to his age. The remainder of the physical examination is unremarkable. He recently had a urinalysis that showed elevated urinary calcium excretion.

Additional lab testing results are as follows:

Serum sodium: 135 mEq/L
Potassium: 3.2 mEq/L
Chloride: 95 mEq/L
Plasma renin activity: 10 ng/mL/hr (normal, 0.5–3.3 ng/mL/hr)
Plasma aldosterone concentration: 20 ng/dL (normal, 2–9 ng/dL)

What is the most likely diagnosis?

Options: **A.** Renal artery stenosis **B.** Bartter's syndrome **C.** Pheochromocytoma **D.** Gitelman's syndrome

Related words in the refined query: elevated urinary; Bartter's; potassium; hypokalaemia; calcium; aldosterone.

Removed Documents

Doc 1

Diagnosis and Pathophysiology Background — Renal Electrolyte Disorders. The case describes a 9-year-old boy presenting with generalized weakness, increased urination, increased thirst, constipation, and elevated urinary calcium excretion... The key laboratory findings are elevated plasma aldosterone and plasma renin activity, indicating hyperaldosteronism... **Urinary Calcium Excretion:** Elevated urinary calcium excretion is associated with conditions that affect renal calcium handling, including Gitelman's syndrome...

Doc 2

Gitelman syndrome. Diagnosis of Gitelman syndrome can be confirmed after eliminating other common pathological sources of hypokalemia and metabolic alkalosis... Other clinical indicators include elevated serum renin and aldosterone and metabolic alkalosis...

Newly Added Documents

Doc 1

Bartter and Gitelman syndromes are both characterized by low potassium and magnesium, normal to low blood pressure, and hypochloremic metabolic alkalosis. However, Bartter syndrome is also characterized by high renin, high aldosterone, and hypercalciuria...

Doc 2

In a urine calcium:creatinine ratio may help distinguish it from Bartter syndrome as the two disorders can be clinically indistinguishable...

Answer Transition: D. Gitelman's syndrome → B. Bartter's syndrome ✓