Less Evidence, Better Answering: Gain-Aware Compact Evidence Subset Selection for Medical QA

Songyue Guo[1], **Zhao Chen**[1], **Caleb Chen Cao**[2,*], **Lei Chen**[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology [*]Corresponding author
sguo349@connect.hkust-gz.edu.cn, zchenah@connect.ust.hk
cao@ust.hk, leichen@cse.ust.hk

Abstract

Evidence-grounded medical question answering requires not only accurate answers, but also precise, verifiable, and clinically trustworthy evidence. Existing medical RAG systems typically retrieve top-k chunks according to item-wise relevance. Although this design improves recall, it does not explicitly optimize whether the selected evidence set is compact, sufficient, and decision-critical. In this paper, we study medical evidence-grounded QA from a different perspective: instead of retrieving evidence without optimization, the evidence-grounded medical QA system should identify a compact sufficient subset of high-density evidence snippets. We propose **MEGA**, a gain-aware framework for compact supporting medical evidence selection. MEGA first expands clinical queries into complementary subqueries and performs hybrid retrieval to construct a recall-oriented candidate evidence pool. It then introduces hidden-state *Information Gain Scoring*, which uses a frozen LLM to estimate whether each candidate snippet contributes new answer-relevant information beyond surface relevance. Finally, MEGA formulates evidence selection as a budgeted utility maximization problem and proposes an *FPTAS selector* to identify a compact evidence subset under an explicit token budget. We further construct two evidence-grounded medical QA benchmarks, CRC-EvidenceQA and Med-EvidenceQA, to evaluate both answer quality and evidence grounding. Across three LLM backbones and five medical RAG baselines, MEGA consistently achieves the best results, improving over the second-best baseline by 8.8% on average in answer quality and 38.8% on average in evidence F1. Ablation studies and blinded clinical expert evaluation further validate its robustness and clinical relevance. Our code and data are available at https://mega-evidenceqa.github.io/.

1 Introduction

Evidence-based medicine (EBM) requires clinical decisions to be justified by the best available evidence rather than by plausible-sounding explanations alone [25]. Recent large language models (LLMs) have shown impressive performance on medical examinations and benchmark datasets [21, 27], and retrieval-augmented generation (RAG) has become a standard strategy for improving factuality by grounding generation in external documents [36, 13, 41]. However, in high-stakes medical settings, correctness alone is insufficient. Specifically, users do not merely ask for an answer; they also need the answer to be grounded in precise, verifiable, and trustworthy evidence, such as guidelines, evidence summaries, and curated literature [28]. This need has also motivated evidence-oriented medical AI systems such as OpenEvidence [23, 7], which highlight the practical importance of retrieving and presenting clinically trustworthy evidence. Therefore, a useful medi-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

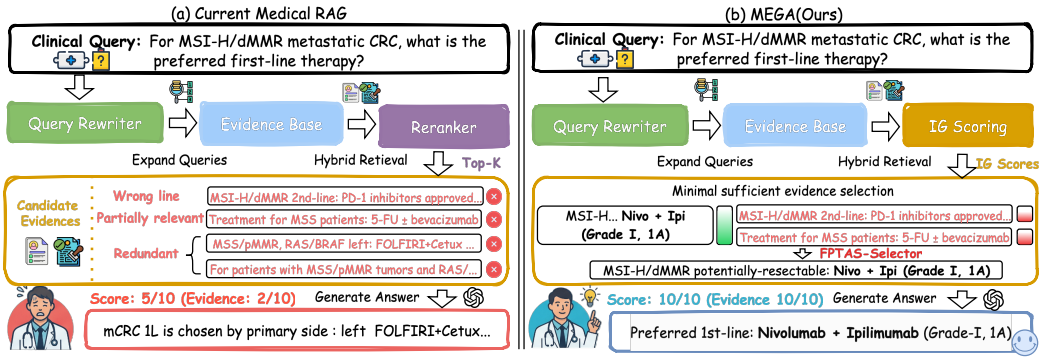


Figure 1: **Motivation of MEGA.** Most medical RAG systems retain top-k evidence snippets using reranking scores. In dense medical guidelines, this strategy often brings decision-irrelevant snippets into the supporting context, which can distract the answer model, weaken citation faithfulness, and increase clinicians' cognitive burden. This motivates us to reformulate evidence-grounded medical QA as a budgeted evidence selection problem: under a token budget, the system should select a compact set of high-utility evidence snippets rather than simply retrieve more context.

cal QA system must tell the user not only *why* an answer is supported, but also *which* evidence is actually necessary.

Existing medical QA and RAG systems only partially satisfy this requirement. On the one hand, answer-centric benchmarks primarily evaluate whether the final response is correct, while providing limited visibility into the evidential basis behind that response [5]. On the other hand, current medical RAG pipelines typically retrieve top-k chunks and attach them as supporting context [12, 13, 35]. While this design often improves the coverage of external knowledge, it also introduces a systematic failure mode in clinical evidence use. As shown in Figure 1, current medical RAG often retrieves *redundant*, *partially relevant*, or even semantically similar but *irrelevant evidence*. This exacerbates a classical cognitive bottleneck: as Herbert Simon famously noted, "*a wealth of information creates a poverty of attention*" [26]. Therefore, physicians rarely need redundant and long evidence and instead require **a compact supporting subset of high-density evidence snippets**. However, these issues cannot be fixed by simple modifications to current medical RAG. *The first limitation is evidence utility estimation.* Cross-encoder rerankers can model query–evidence interactions, but they still assign scores to individual snippets as a pairwise relevance problem. In dense medical guidelines, locally relevant snippets may mention the same disease or treatment yet fail to provide the decision-critical support required by final answers. Therefore, medical evidence-grounded QA needs a signal that measures whether a snippet contributes new answer-relevant information, not only whether it is relevant to the query. *The second limitation is evidence selection.* Top-k retrieval only ranks snippets independently and cannot explicitly optimize the evidence subset selection under the token constraint.

To address these limitations, we take a step beyond retrieval-centric medical RAG by formalizing minimal sufficient evidence selection as a gain-aware budgeted optimization problem for medical evidence-grounded QA, which is NP-hard. Unlike conventional medical RAG, which retrieves and ranks top-k evidence items independently, this problem asks for the minimal sufficient subset for evidence-grounded answer generation under a token budget. To instantiate this formulation, we propose **ME**dical Evidence Selection with **G**ain-**A**ware Subset Optimization (**MEGA**). Specifically, MEGA first expands the clinical query into complementary sub-queries and retrieves a candidate pool. We then propose *Information Gain Scoring* (IGS), which estimates whether each candidate snippet contributes answer-relevant information beyond local relevance. Finally, MEGA selects a minimal sufficient evidence subset with a *fully polynomial-time approximation scheme (FPTAS) selector* yielding a direct $(1-\epsilon)$-approximation guarantee for the selection optimization objective and generates an answer based on the selected evidence. We further construct two evidence-grounded medical QA benchmarks, CRC-EvidenceQA and Med-EvidenceQA, to evaluate both answer quality and evidence grounding. Across three LLM backbones and five medical RAG baselines, MEGA improves over the second-best baseline by 8.8% on average in answer quality and 38.8% on average in evidence F1, while using 35.6% fewer evidence tokens.

Our contributions are summarized as follows:

- We formalize medical evidence-grounded QA as a *gain-aware budgeted evidence selection* problem. Unlike standard medical RAG that independently retains top-k snippets by item-wise relevance, our formulation asks the system to select a compact supporting evidence subset under a token budget, and we show that this problem is NP-hard.
- We propose **MEGA**, a gain-aware framework that turns retrieval-centric medical RAG into budgeted evidence subset selection. MEGA utilizes hidden-state *Information Gain Scoring* (IGS) to estimate whether each candidate snippet contributes answer-relevant information beyond local relevance and a fully polynomial-time approximation scheme (FPTAS) to solve the optimization objective, yielding a direct $(1-\epsilon)$-approximation guarantee for the stated budgeted selection objective.
- We construct two evidence-grounded medical QA benchmarks, CRC-EvidenceQA and Med-EvidenceQA, and conduct experiments across three LLM backbones and five medical RAG baselines. MEGA improves over the second-best baseline by 8.8% on average in answer quality and 38.8% on average in Evidence F1, while using 35.6% fewer evidence tokens. Blinded clinician evaluation further supports the sufficiency and minimality of the selected evidence.

2 Problem Formulation

In medical evidence-grounded QA, retrieving more text does not necessarily improve reasoning quality. Instead, excessive evidence often introduces redundancy and distracts downstream generation. Therefore, the core goal is to select a *compact supporting subset* of evidence that fully supports the answer while reducing the cognitive burden of evidence review.

Formally, let the medical evidence corpus be denoted by $\mathcal{D} = \{d_1, d_2, \ldots, d_{N_D}\}$, where each document d_j is an authoritative clinical guideline, evidence summary, or curated medical literature. After segmentation and hybrid retrieval, $\mathcal{D}$ yields a candidate evidence pool $C = \{c_1, c_2, \ldots, c_N\}$, where each c_i is an atomic evidence unit, such as a sentence or a short snippet. Given a clinical query q, our goal is to select a subset $S^\star \subseteq C$ that provides sufficient support for answering q, while keeping the selected evidence subset as compact as possible.

A medical query often involves multiple latent clinical information needs, such as disease staging, mutation status, contraindications, treatment line, and patient-specific conditions. However, these clinical requirements are not directly observed during inference, and we do not assume access to manual annotations indicating which snippet covers which requirement. Instead, MEGA measures evidence usefulness through a set of measurable information-gain signals. For each candidate evidence unit c_i, we estimate a K-dimensional gain vector

$$\tilde{\mathbf{s}}_i(q) = [\tilde{s}_{i1}(q), \tilde{s}_{i2}(q), \ldots, \tilde{s}_{iK}(q)] \in [0,1]^K,$$

where each dimension corresponds to an operational proxy for evidence utility.

Unlike standard top-k retrieval, which keeps a fixed number of locally relevant snippets, our formulation explicitly selects a subset under a global evidence budget. This turns evidence selection into a budgeted utility maximization problem.

Definition 1 (Compact Supporting Evidence Selection)**.** *Given a clinical query q, a candidate evidence pool $C = \{c_1, \ldots, c_M\}$, set size penalty $|S|$, token costs $\{t_i\}_{i=1}^M$, and a token budget B, the budgeted minimal evidence selection problem is formulated as*

$$S^\star = \arg\max_{S \subseteq C} F(S;q) = \sum_{k=1}^{K} \sum_{c_i \in S} w_k \widetilde{s}_{ik} - \gamma|S| - \mu \sum_{c_i \in S} \frac{t_i}{B} \quad \text{s.t.} \quad \sum_{c_i \in S} t_i \le B, \tag{1}$$

Here, $w_k \ge 0$ denotes the importance weight of the k-th gain dimension with $\sum_{k=1}^K w_k = 1$, $\gamma \ge 0$ penalizes selecting an additional evidence unit, and $\mu \ge 0$ softly penalizes long snippets. The token budget B further imposes a hard upper bound on the total context length passed to the answer generator.

Theorem 1 (NP-Hardness)**.** *The compact supporting evidence subset selection problem in Definition 1*

$$\arg\max_{S \subseteq C} F(S;q) = \sum_{k=1}^{K} \sum_{c_i \in S} w_k \widetilde{s}_{ik} - \gamma|S| - \mu \sum_{c_i \in S} \frac{t_i}{B} \quad \text{s.t.} \quad \sum_{c_i \in S} t_i \le B, \tag{2}$$

is NP-hard.

Proof. We prove the NP-hardness by a polynomial-time reduction from the 0/1 KNAPSACK DECISION problem, which is NP-Complete. The details of the proof can be seen in Appendix C.1. □

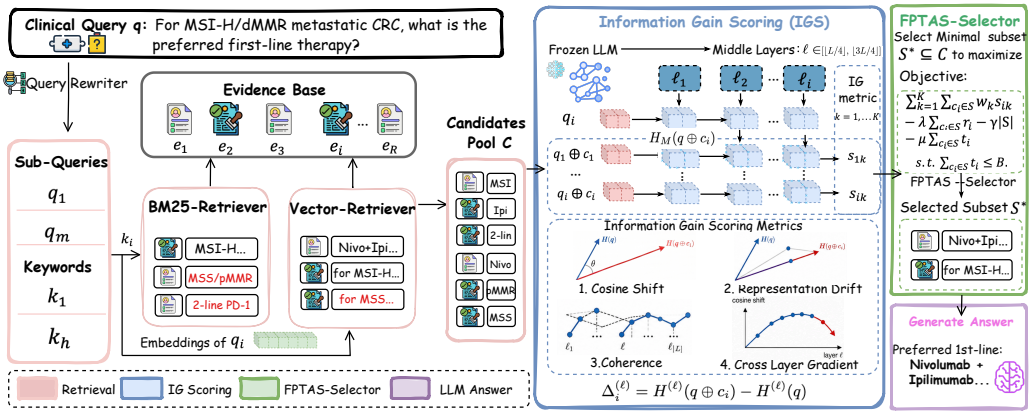


Figure 2: Overview of the MEGA framework. Given a clinical query q and a medical evidence corpus $\mathcal{E}$, the framework first decomposes the query into subqueries and keywords and retrieves a diverse candidate pool, then estimates the information gain of each candidate evidence unit, and then MEGA selects a compact sufficient evidence subset by an FPTAS evidence selector to solve a budget-aware subset optimization problem.

3 Methodology

We propose MEGA, a medical evidence-grounded QA framework that selects a compact sufficient evidence subset for evidence-grounded answer generation. As illustrated in Figure 2, MEGA leverages the effective information gain signal to score each candidate evidence and utilizes an FPTAS solver to optimize the objective defined in Definition 1. The pipeline consists of four stages: (1) Sub-queries Generation and hybrid retrieval, (2) Hidden state information gain scoring, (3) FPTAS evidence selector, and (4) Answer generation. The pseudo-code of the MEGA pipeline is in Appendix D.

3.1 Sub-queries Generation and Hybrid Retrieval

The first stage constructs a recall-oriented candidate evidence pool for subsequent budgeted selection (prompts and examples are in G.1). A single medical query often contains multiple clinical conditions, such as tumor type, mutation status, disease stage, treatment line, prior therapy, and contraindications. Retrieving only with the original query may overemphasize the most salient terms while missing evidence for other clinically relevant conditions. Therefore, we decompose the original query q into complementary subqueries and medical keywords, which serve as retrieval probes rather than manually annotated clinical requirements.

Given a query-decomposition prompt $P_{\text{decompose}}$, we use an LLM to generate $\mathcal{Q}(q) = \{q_1, \ldots, q_m\}$ and $\mathcal{K}(q) = \{k_1, \ldots, k_h\}$. Each subquery q_j focuses on one possible retrieval intent, such as biomarker status, treatment indication, disease stage, or recommended therapy, while each keyword k_l captures exact medical terminology, such as drug names, biomarkers, guideline phrases, or disease subtypes. This decomposition improves candidate recall before the later gain-aware scoring and budgeted subset selection stages.

Let $\mathcal{E} = \{e_1, \ldots, e_R\}$ denote the segmented snippet-level evidence base obtained from the original corpus $\mathcal{D}$ using SciSpaCy biomedical sentence-boundary detection, followed by lightweight post-processing that merges neighboring short sentences to preserve local clinical context [18]. For each subquery q_j, we retrieve evidence from $\mathcal{E}$ using both lexical and dense retrieval. The lexical retriever uses BM25 to capture exact medical terms, while the dense retriever uses an embedding model f_{emb} to capture semantic similarity, computed by cosine similarity between $f_{\text{emb}}(q_j)$ and $f_{\text{emb}}(e_i)$. We combine the two signals with a normalized hybrid score:

$$\text{Score}_{\text{ret}}(q_j, e_i) = \alpha \widetilde{\text{BM25}}(q_j, e_i) + (1-\alpha)\widetilde{\text{Sim}}_{\text{dense}}(q_j, e_i),$$

where $\alpha \in [0,1]$, and both components are normalized within the retrieved pool. We set $\alpha = 0.5$ in all experiments. The BM25 component preserves exact matching for medical entities, while the dense component retrieves semantically related snippets that may not share identical surface forms with the query. For each subquery q_j, we keep the top n evidence units according to $\text{Score}_{\text{ret}}$. We then merge the retrieved results across all subqueries and remove duplicated snippets, yielding the candidate pool $C = \text{Dedup}(\bigcup_{j=1}^{m} \text{Top-}n(q_j, \mathcal{E})) = \{c_1, \ldots, c_N\} \subseteq \mathcal{E}$. If the same evidence unit is retrieved by multiple subqueries, we keep its maximum retrieval score, i.e., $\text{Score}_{\text{ret}}(q, c_i) = \max_j \text{Score}_{\text{ret}}(q_j, c_i)$.

3.2 Hidden-State Information Gain Scoring

The candidate pool C constructed by recall-oriented retrieval still contains redundant or clinically mismatched snippets. This reflects the first limitation of current medical RAG: *rerankers only evaluate relevance, not decision knowledge support*. To estimate whether evidence contributes new answer-relevant information beyond surface relevance, we propose **Information Gain Scoring** (IGS). Inspired by prior work [20] showing that the internal states of LLMs' middle layers can reflect knowledge integration, IGS measures how the hidden states of a frozen LLM change when a candidate snippet is appended to the query. We treat these changes as operational proxy signals for evidence utility, rather than assuming that every representation shift is clinically useful.

Let $\mathcal{M}_{\text{frozen}}$ be a frozen pre-trained LLM with L layers. For an input sequence x, let $H^{(\ell)}(x) \in \mathbb{R}^d$ denote the mean-pooled hidden representation at layer ℓ. We use mean pooling over non-padding tokens to obtain a stable sequence-level representation. We focus on the middle-layer set $\mathcal{L} = \{\lfloor L/4 \rfloor, \ldots, \lfloor 3L/4 \rfloor\}$, since early layers are more lexical while very late layers are more generation-specific; middle layers are more suitable for capturing semantic integration (we evaluate alternative layer regions in Appendix E.3). For each candidate evidence unit c_i, we compare the query-only representation $H^{(\ell)}(q)$ with the query-plus-evidence representation $H^{(\ell)}(q \oplus c_i)$, where $q \oplus c_i$ denotes concatenation. Throughout this section, $\eta > 0$ denotes a small numerical constant for stable normalization.

We compute four complementary raw gain signals $g_i = [g_{i1}, g_{i2}, g_{i3}, g_{i4}]$, each capturing a different form of evidence-induced representation change. Let

$$\Delta_i^{(\ell)} = H^{(\ell)}(q \oplus c_i) - H^{(\ell)}(q), \quad a_\ell(c_i) = 1 - \cos\Big(H^{(\ell)}(q), H^{(\ell)}(q \oplus c_i)\Big).$$

Cosine Shift. To measure whether a snippet changes the semantic direction of the query representation, we compute the average angular displacement:

$$g_{i1} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} a_\ell(c_i). \tag{3}$$

In medical evidence-grounded QA, decision-critical evidence often shifts the model from a general disease-level representation to a more specific clinical decision state.

Representation Drift. Cosine shift measures direction but does not reflect how strongly the hidden state moves. To measure the relative magnitude of the evidence-induced representation change, we compute the normalized Euclidean movement:

$$g_{i2} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{\|\Delta_i^{(\ell)}\|_2}{\|H^{(\ell)}(q)\|_2 + \eta}. \tag{4}$$

It is useful because two snippets may point in a similar semantic direction but contribute different amounts of answer-relevant information. For instance, a short snippet that only repeats a drug name may have limited drift, while a guideline sentence that states the indication, patient subgroup, and recommended therapy can introduce a larger representation movement.

Cross-Layer Coherence. Inspired by prior work [10], we measure the directional consistency of evidence-induced shifts across adjacent layers. Unlike the shift magnitude at an individual layer, it describes whether the perturbation direction persists through the selected layer range:

$$g_{i3} = \frac{1}{|\mathcal{L}| - 1} \sum_{\ell \in \mathcal{L}^-} \frac{1}{2} \left(1 + \frac{(\Delta_i^{(\ell)})^\top \Delta_i^{(\ell+1)}}{\|\Delta_i^{(\ell)}\|_2 \|\Delta_i^{(\ell+1)}\|_2 + \eta} \right), \tag{5}$$

where $\mathcal{L}^-$ denotes the middle layers excluding the last one. In our setting, useful medical evidence should not merely trigger a one-layer perturbation; it should induce a relatively consistent update as the model integrates the snippet across layers. By contrast, a verbose but clinically mismatched snippet may create a large local shift but unstable cross-layer movement.

Cross-Layer Information Gradient. Finally, answer-relevant evidence should become more integrated as the representation moves toward deeper semantic layers. To measure this depth-wise accumulation effect, let $z_\ell \in [0, 1]$ be the normalized depth index of layer ℓ within $\mathcal{L}$. We estimate the positive slope of the layer-wise shift sequence:

$$g_{i4} = \left[\frac{\sum_{\ell \in \mathcal{L}} (z_\ell - \bar{z})(a_\ell(c_i) - \bar{a})}{\sum_{\ell \in \mathcal{L}} (z_\ell - \bar{z})^2 + \eta} \right]_+, \tag{6}$$

where $\bar{z}$ and $\bar{a}$ are the means of $\{z_\ell\}_{\ell \in \mathcal{L}}$ and $\{a_\ell(c_i)\}_{\ell \in \mathcal{L}}$, respectively. A positive value indicates that the angular effect of the candidate grows across the selected middle layers. The term gradient refers to this depth-wise trend, not to a training gradient: computing the score requires neither backpropagation nor parameter

updates. A growing trajectory is not, by itself, evidence of factual correctness. This is important for distinguishing snippets that merely share surface medical terms from snippets that progressively support the final clinical decision.

After computing the four raw gain signals, we normalize each dimension across the candidate pool:

$$\widetilde{s}_{ik} = \frac{g_{ik} - \min_j g_{jk}}{\max_j g_{jk} - \min_j g_{jk} + \eta}, \qquad \widetilde{S} = [\widetilde{s}_{ik}] \in [0,1]^{N \times K}.$$

The normalized vector $\widetilde{\mathbf{s}}_i = [\widetilde{s}_{i1}, \ldots, \widetilde{s}_{iK}]$ is used as the gain-aware evidence utility representation in Definition 1. Snippets with weak information-gain signals naturally receive lower utility and are less likely to be selected by the subsequent budgeted selector.

3.3 FPTAS Evidence Selector

The third stage converts the IGS scores into a compact evidence subset under an explicit context budget. After IGS, each candidate $c_i \in C$ has a normalized gain vector $\widetilde{\mathbf{s}}_i$ and a token cost $t_i = |c_i|$. A simple top-k strategy is insufficient here: it fixes the number of snippets rather than the total evidence burden, and it may keep long or redundant snippets simply because they have high individual scores. Therefore, we formulate selection as Definition 1, where the utility of c_i is

$$v_i(q) = \sum_{k=1}^{K} w_k \widetilde{s}_{ik} - \gamma - \mu \frac{t_i}{B}.$$

The selector chooses

$$S^\star = \arg\max_{S \subseteq C} \sum_{c_i \in S} v_i(q), \qquad \text{s.t.} \quad \sum_{c_i \in S} t_i \le B,$$

where B is a hard context budget, while γ and μ penalize excessive evidence count and long snippets. Before optimization, we discard candidates with $v_i(q) \le 0$ or $t_i > B$.

A greedy strategy is a natural baseline for this problem: it is fast, but can be unstable when high-utility evidence snippets have uneven token costs. Moreover, exact dynamic programming is more reliable but can be expensive. Therefore, we utilize the classical value-scaling FPTAS, which provides a controllable trade-off between inference time and solution quality: smaller ϵ gives a tighter approximation ratio at a higher computational cost. After filtering out candidates with non-positive utility $v_i(q) \le 0$, we denote the remaining candidate set by $\bar{C}$, with $N' = |\bar{C}|$.

$$v_{\max} = \max_{c_i \in \bar{C}} v_i(q).$$

For $\epsilon \in (0,1)$, we set the scaling factor $\Delta = \epsilon v_{\max}/N'$ and use integer values $\widehat{v}_i = \lfloor v_i(q)/\Delta \rfloor$. The dynamic program stores the minimum token cost needed to achieve each scaled utility, and returns the feasible subset $S_\epsilon^\star$ with the largest scaled utility under budget B.

Theorem 2 (FPTAS Guarantee)**.** *Let $S^\star$ be the optimal solution to Eq. (2) over the filtered candidate set $\bar{C}$, and let $V^\star = \sum_{c_i \in S^\star} v_i(q)$. For any $\epsilon \in (0,1)$, the value-scaling FPTAS returns a subset $S_\epsilon \subseteq \bar{C}$ such that*

$$\sum_{c_i \in S_\epsilon} v_i(q) \ge (1-\epsilon) V^\star. \tag{7}$$

The running time is $O(N'^3/\epsilon)$, and the memory cost is $O(N'^3/\epsilon)$ with a standard DP table, or $O(N'^2/\epsilon)$ plus backtracking pointers when recovering the selected subset.

Proof. Due to space limitations, we provide the details of the proof in Appendix C.2 □

Finally, the answer generator receives q and S_ϵ, and produces a citation-grounded answer with prompt P_{answer}(prompts and examples are in Appendix G): $a = \text{Gen}_\psi(q, S_\epsilon, P_{\text{answer}})$. By combining IGS-based utility estimation with budgeted subset optimization, MEGA turns evidence retrieval from fixed top-k ranking into gain-aware compact supporting evidence selection under the token budget.

4 Experimental Setup

4.1 Datasets

We construct two evidence-grounded medical QA benchmarks, CRC-EvidenceQA and Med-EvidenceQA, and evaluate MEGA on them. These benchmarks are designed to evaluate not only answer generation, but also whether a method can select a compact supporting evidence subset under a context budget. Details of benchmark construction and dataset statistics are provided in Appendix F.

Table 1: Main results across CRC-EvidenceQA and Med-EvidenceQA. CRC is evaluated by LLM-as-Judge overall score and Evidence F1. PubMedQA reports label accuracy and Evidence F1. COVID-QA reports Rouge-L and Evidence F1. All results are reported as mean±std over three repeated runs. **best** indicates the best result, and second-best indicates the second-best result.

Backbone	Method	CRC-EvidenceQA			PubMedQA		COVID-QA	
		Low-Ov.	High-Ov.	Ev-F1	Acc.(%)	Ev-F1(%)	Rouge-L(%)	Ev-F1(%)
GPT-4.1	RAG [12]	8.2±0.2	7.5±0.3	47.8±1.1	72.4±0.8	50.1±1.0	67.9±0.9	44.5±1.2
	Context+LLM	8.6±0.2	7.6±0.2	67.1±0.9	78.8±0.7	58.9±0.8	76.2±0.7	60.3±0.9
	MedRAG [40]	8.1±0.3	7.2±0.3	51.9±1.2	76.5±0.9	48.3±1.1	78.1±0.8	50.2±1.0
	MMed-RAG [35]	8.4±0.2	7.6±0.3	52.8±1.0	75.2±0.8	47.6±1.2	75.6±0.9	54.6±1.1
	DoctorRAG [13]	8.7±0.2	8.1±0.2	60.4±1.0	80.5±0.6	53.3±1.0	81.4±0.6	53.2±1.1
	MEGA (ours)	**9.5±0.1**	**8.7±0.2**	**84.9±0.6**	**82.1±0.5**	**85.0±0.6**	**85.4±0.5**	**83.7±0.7**
Qwen3.6-35B-A3B	RAG [12]	7.4±0.3	6.8±0.3	46.5±1.2	70.9±0.9	44.7±1.2	65.8±1.0	46.5±1.3
	Context+LLM	7.8±0.2	7.4±0.3	52.5±1.0	73.5±0.8	53.4±0.9	71.0±0.9	54.9±1.0
	MedRAG [40]	7.7±0.3	6.7±0.3	48.7±1.1	75.4±0.7	47.3±1.1	76.2±0.8	48.3±1.2
	MMed-RAG [35]	7.4±0.3	6.3±0.4	50.9±1.2	74.1±0.8	43.4±1.3	73.8±0.9	44.9±1.2
	DoctorRAG [13]	8.3±0.2	7.8±0.2	51.4±1.1	75.2±0.7	45.1±1.2	76.1±0.8	45.4±1.2
	MEGA (ours)	**9.3±0.1**	**8.5±0.2**	**83.1±0.7**	**80.7±0.6**	**83.4±0.7**	**87.6±0.5**	**81.9±0.8**
MiniMax-M2.5-230B	RAG [12]	8.3±0.2	7.6±0.3	48.3±1.1	72.4±0.8	49.1±1.1	67.9±0.9	36.5±1.4
	Context+LLM	8.5±0.2	7.8±0.2	68.5±0.9	74.8±0.8	62.3±0.8	73.2±0.9	66.3±0.9
	MedRAG [40]	7.9±0.3	6.9±0.3	52.9±1.1	76.5±0.7	51.5±1.0	78.1±0.8	53.3±1.1
	MMed-RAG [35]	8.2±0.2	7.6±0.3	54.9±1.0	76.4±0.8	55.9±1.0	76.3±0.8	52.9±1.1
	DoctorRAG [13]	8.8±0.2	8.2±0.2	53.6±1.1	78.2±0.7	51.9±1.0	78.2±0.8	51.8±1.1
	MEGA (ours)	**9.6±0.1**	**9.2±0.1**	**85.1±0.6**	**82.4±0.5**	**85.2±0.6**	**88.2±0.5**	**82.1±0.7**

CRC-EvidenceQA. CRC-EvidenceQA is a clinician-collaborative colorectal cancer (CRC) benchmark for guideline-grounded clinical reasoning. It contains 200 open-ended clinical questions, including 100 Chinese and 100 English questions. For each language, we collect 50 low-difficulty and 50 high-difficulty cases from **two CRC specialists**. Low-difficulty questions mainly require direct guideline lookup and routine clinical decisions, while high-difficulty questions involve multi-step reasoning over staging, molecular stratification, treatment sequencing, follow-up, and evidence-based cases.

Med-EvidenceQA. Med-EvidenceQA is a larger English benchmark adapted from PubMedQA [8] and COVID-QA [15]. It contains 1,000 yes/no/maybe biomedical QA examples from PubMedQA and 1,973 extractive QA examples from COVID-QA. For both subsets, we preserve candidate context sentences and supporting evidence annotations, enabling evaluation of both answer quality and evidence selection.

4.2 LLM Backbones

We evaluate MEGA and other baselines with GPT-4.1 [22], Qwen3.6-35B-A3B [24], and MiniMax-M2.5-230B [14] as LLM backbones. This set covers different model families and scales, allowing us to test whether the proposed evidence selection strategy remains robust across heterogeneous LLMs. For retrieval and rerank, we use Qwen3-Embedding-8B [37] and Qwen3-Reranker-8B [37], respectively. Unless otherwise specified, Qwen3-8B [30] is used as the default frozen scoring model, and Qwen3-32B [30] is used for sensitivity analysis in Appendix E.3. These frozen models are not updated during evaluation and are used only to compute model-internal evidence signals, separate from the LLM backbone mentioned above.

4.3 Hyperparameters

For MEGA, default hyperparameters are selected on a held-out development split of Med-EvidenceQA using evidence F1 and answer quality as model-selection criteria. We use $w = (0.35, 0.35, 0.15, 0.15)$, $\gamma = 0.5$,$\mu = 0.25$, and $\epsilon = 0.1$ by default, and report full sensitivity analyses in the Appendix E.1. For each dataset, B is tuned on the development split and fixed for testing. No test-set gold evidence labels or gold evidence lengths are used during inference.

5 Results

5.1 Comparison with other Medical RAG Baselines

To evaluate the effectiveness of MEGA , we compare it with five representative RAG baselines: standard RAG, Context+LLM, MedRAG, MMed-RAG, and DoctorRAG. All methods are evaluated with GPT-4.1, Qwen3.6-35B-A3B, and MiniMax-M2.5-230B as LLM backbones. We report task-specific answer metrics and evidence-

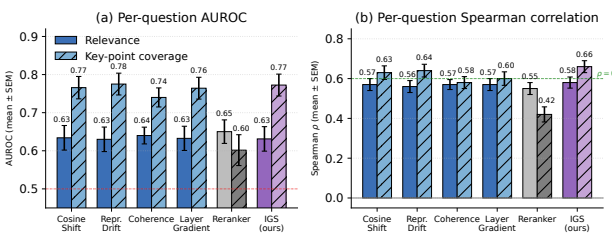


Figure 3: Comparison between information gain scoring (IGS) and Reranker.

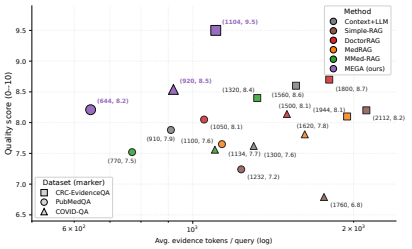


Figure 4: Pareto-frontier of token length and answer quality.

selection metrics. For CRC-EvidenceQA, we follow the metric design of the previous work [33] and use GPT-5.1 as an LLM-as-Judge (prompt is provided in Appendix G.4) to score open-ended answers based on the clinicians' golden answers on a 10-point scale and evaluate answer correctness, completeness, citation faithfulness, and logical consistency. For PubMedQA, we report label accuracy, and for COVID-QA, we report Rouge-L. Across all datasets, Evidence F1 is computed at the evidence-unit level: for each question, the selected evidence set is compared with the gold supporting evidence set using precision, recall, and their harmonic mean, and we report the average F1 over questions. More implementation details and results are provided in Appendix D.1 and D.2.

Table 1 shows that MEGA achieves the best performance across all benchmarks, metrics, and LLM backbones. On CRC-EvidenceQA, MEGA improves over the second-best baseline by 10.1% on low-difficulty overall score and 9.5% on high-difficulty overall score, averaged over the three backbones. The gains are larger for evidence grounding: compared with the second-best average performance, MEGA improves Evidence F1 by 34.6% on CRC-EvidenceQA, 45.2% on PubMedQA, and 36.5% on COVID-QA. On Med-EvidenceQA, MEGA also improves answer-level metrics, with average relative gains of 4.7% on PubMedQA accuracy and 10.8% on COVID-QA Rouge-L over the second-best baseline. The smaller gain on PubMedQA is expected because its yes/no/maybe label accuracy is relatively coarse, while COVID-QA is more sensitive to evidence quality through extractive answer overlap. Across backbones, the consistent improvement suggests that MEGA is not tied to a specific generator. Instead, the advantage mainly comes from better evidence selection: IGS estimates answer-relevant evidence utility beyond pairwise relevance, and the FPTAS selector converts these utilities into a compact evidence subset under the token budget.

5.2 Utility of information gain scoring (IGS)

We next evaluate whether Information Gain Scoring (IGS) provides a better evidence-utility signal than conventional reranking. For each question, we score every candidate snippet using four hidden-state signals: cosine shift, representation drift, cross-layer coherence, and cross-layer information gradient. We compare these scores with two evidence labels. *Relevance* indicates whether a snippet is topically related to the question (judged by GPT-5.1), while *key-point coverage* indicates whether it supports an essential evidence point required by the gold answer. We report per-question AUROC and Spearman correlation, with mean and standard error shown in Figure 3.

Figure 3 shows that reranking remains competitive for general relevance, achieving 0.65 AUROC, which is slightly higher than individual hidden-state signals. However, relevance is not sufficient for evidence-grounded medical QA. For key-point coverage, the reranker drops to 0.60 AUROC and 0.42 Spearman correlation, indicating that pairwise relevance scores often fail to identify decision-critical evidence. In contrast, all four hidden-state dimensions provide stronger key-point signals: cosine shift, representation drift, cross-layer coherence, and layer gradient achieve AUROC scores of 0.77, 0.78, 0.74, and 0.76, respectively, with Spearman correlations of 0.63, 0.64, 0.58, and 0.60. After combining these dimensions, IGS achieves 0.77 AUROC and the highest Spearman correlation of 0.66 on key-point coverage. These results support our motivation that medical evidence selection requires estimating answer-relevant utility rather than only query relevance. Dense guidelines often contain snippets that mention the same disease, biomarker, or treatment but do not support the final clinical decision. Rerankers are effective at identifying such locally relevant snippets, but IGS better captures whether a snippet contributes new information useful for the answer. This explains the evidence-F1 improvements in Table 1: better utility estimation helps the budgeted selector choose compact evidence subsets that better match the gold supporting evidence.

5.3 Effectiveness of FPTAS selector

We further evaluate whether the FPTAS knapsack selector can convert IGS utilities into a compact evidence subset under a token budget. For MEGA , the recall-oriented retrieval stage first constructs the candidate pool C, and the selector then chooses the final subset under a fixed budget B. For all methods, we report the average number of evidence tokens passed to the answer generator. To compare different datasets in one figure, answer

quality is mapped to a unified 0–10 scale: CRC-EvidenceQA uses the LLM-as-Judge overall score, PubMedQA uses label accuracy divided by 10, and COVID-QA uses Rouge-L divided by 10.

Figure 4 shows that MEGA achieves the best quality–token length trade-off across all three datasets. On PubMedQA, MEGA reaches 8.2 quality score with only 644 evidence tokens on average, outperforming DoctorRAG (8.1, 1050 tokens) and Context+LLM (7.9, 910 tokens). MMed-RAG achieves second-best on the PubMedQA subset and CRC-EvidenceQA by its inner evidence selection pipeline. These results show that better medical QA does not require larger retrieved contexts. Retrieval-heavy methods often pass more evidence to the generator, but the additional tokens do not necessarily improve answer quality and may introduce redundant or distracting snippets. In contrast, MEGA first estimates answer-relevant evidence utility with IGS and then uses the FPTAS selector to optimize utility under an explicit token budget.

5.4 Ablation Study

We ablate each component by Minimax-M2.5 under CRC-EvidenceQA with the same generation prompt and token-budget setting. Removing sub-query generation means retrieving only with the original query; removing IGS replaces hidden-state utilities with reranker-based scores; removing the FPTAS selector replaces budgeted subset optimization with top-ranked evidence selection.

Table 2: Ablation Study of MEGA .

Variant	Ev-F1↑	Δ	Tok.↓	Δ	Q ↑	Δ
MEGA (full)	**85.1**	–	**920**	–	**9.5**	–
– Sub-Queries Generation	83.8	-1.3	870	-50	8.2	-1.3
– Information Gain Scoring	72.2	-12.9	1180	+260	8.7	-0.8
– FPTAS Selector	63.3	-21.8	1850	+930	8.4	-1.1
– All (= Vanilla RAG)	48.3	-36.8	1760	+840	8.3	-1.2

Table 2 shows that removing each component lowers performance in this experimental setting. Without sub-query generation, Evidence F1 drops from 85.1 to 83.8, and answer decreases from 9.5 to 8.2, suggesting that the original query alone misses some clinically necessary retrieval intents. Removing IGS causes a larger drop in Evidence F1 (85.1 → 72.2) and increases token usage (920 → 1180), showing that reranker scores are less effective at estimating evidence gain utility. Removing the FPTAS selector further reduces Evidence F1 to 63.3 and nearly doubles token usage to 1850, indicating that score ranking alone cannot produce a compact sufficient evidence subset. Vanilla RAG performs worst, with 48.3 Evidence F1 and 1760 tokens. These results show that sub-query generation improves candidate recall, IGS improves utility estimation, and the FPTAS selector improves budget-aware evidence selection.

5.5 Alignment with Clinicians' Evidence-Use Preferences

To further assess whether the selected evidence aligns with clinicians' evidence-use preferences, we conduct a blinded expert evaluation on a subset of CRC-EvidenceQA. We sample 60 questions, including 30 low-difficulty and 30 high-difficulty cases. For each question, outputs from different methods are anonymized and randomly ordered. Clinicians rate each method on a 1–5 scale along three dimensions: answer quality, evidence sufficiency, and evidence minimality. Evidence minimality measures whether the provided evidence is concise and non-redundant. As shown in Figure 5, MEGA receives the highest clinician scores on all dimensions. MEGA achieves 4.7 on evidence sufficiency and 4.3 on evidence minimality, while the strongest baselines remain lower in evidence metrics. This suggests that MEGA does not simply improve answer generation. Instead, it selects evidence that clinicians consider more sufficient and less redundant. These results further support our central claim that medical evidence-grounded QA benefits from selecting a compact sufficient evidence subset rather than only retrieving relevant evidence.

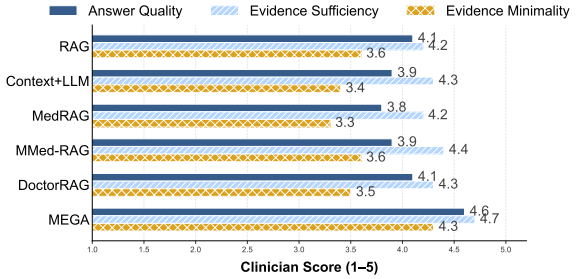


Figure 5: Clinical expert evaluation.

6 Conclusion

In this paper, we presented MEGA, a framework for medical evidence-grounded QA that shifts medical RAG from top-k retrieval to budget-aware minimal evidence selection. MEGA combines subquery-based hybrid retrieval, hidden-state Information Gain Scoring, and an FPTAS knapsack selector to identify compact supporting evidence subsets for evidence-grounded answer generation. Experiments on clinician-collaborative and public medical QA benchmarks show that MEGA improves answer quality and evidence grounding while reducing evidence burden. These results highlight the importance of evidence utility estimation and budgeted subset selection for medical evidence-grounded QA.

Acknowledgments and Disclosure of Funding

Lei Chen's work is partially supported by National Key Research and Development Program of China Grant No. 2023YFF0725100, National Science Foundation of China (No. U22B2060), Guangdong-Hong Kong Technology Innovation Joint Funding Scheme Project No. 2024A0505040012, the Hong Kong RGC GRF Project 16213620, RIF Project R6020-19, AOE Project AoE/E-603/18, Theme-based project TRS T41-603/20R, CRF Project C2004-21G, Key Areas Special Project of Guangdong Provincial Universities 2024ZDZX1006, Guangdong Province Science and Technology Plan Project 2023A0505030011, Guangzhou municipality big data intelligence key lab, 2023A03J0012, Hong Kong ITC ITF grants MHX/078/21 and PRP/004/22FX, Hong Kong ITC TC-SKLCRCC26EG01, Zhujiang scholar program 2021JC02X170, Microsoft Research Asia Collaborative Research Grant, HKUST-Webank joint research lab, 2025 HKUST Shenzhen-Hong Kong Collaborative Innovation Institute Green Sustainability Special Fund from Shui On Xintiandi and the InnoSpace GBA, and HKUST(GZ) - CMCC(Guangzhou Branch) Metaverse Joint Innovation Lab under Grant No. P00659.

Competing Interests Disclosure

The authors declare that they have no competing interests.

References

[1] Aditya Chattopadhyay, Stewart Slocum, Benjamin D. Haeffele, René Vidal, and Donald Geman. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7430–7443, 2023.

[2] Sizhe Chen and Tao Chen. Co-medgraphrag: A collaborative large–small model medical question-answering framework enhanced by knowledge graph reasoning. *Information*, 17(3):247, 2026.

[3] Pengyu Guo, Jiachuan Wang, Zhao CHEN, Caleb Chen Cao, Liping Wang, Tingyi Jiang, and Lei Chen. Pam-mil: Proliferation and metastasis enhanced localization for multiple instance learning on pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5305–5314, June 2026.

[4] Songyue Guo, Maocheng Li, Zhao Chen, Caleb Chen Cao, and Lei Chen. Fqora: Towards fair and quality-based online data marketplace. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2839–2842, 2025.

[5] Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D. Stern, and Marzyeh Ghassemi. MedPAIR: Measuring physicians and AI relevance alignment in medical question answering. *arXiv preprint arXiv:2505.24040*, 2025.

[6] Feiyu Huang, Jia Li, Zhao Chen, Yang Wu, Caleb Chen Cao, and Lei Chen. Kamp: Knowledge-anchored multimodal pretraining framework for medical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21067–21077, 2026.

[7] Ryan T Hurt, Christopher R Stephenson, Elizabeth A Gilman, Christopher A Aakre, Ivana T Croghan, Manpreet S Mundi, Karthik Ghosh, and Jithinraj Edakkanambeth Varayil. The use of an artificial intelligence platform openevidence to augment clinical decision-making for primary care physicians. *Journal of Primary Care & Community Health*, 16:21501319251332215, 2025.

[8] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics.

[9] Andreas Krause and Carlos Guestrin. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pages 1650–1654, 2007.

[10] Daniil Laptev, Nikita Balagansky, Yaroslav Aksenov, and Daniil Gavrilov. Analyze feature flow to enhance interpretation and steering in language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

[11] Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17606–17619, 2025.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[13] Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Goi Sin Yee, and Jinzhuo Wang. Towards doctor-like reasoning: Medical rag fusing knowledge with patient analogy through textual gradients. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[14] MiniMax. MiniMax-M2.5: Built for real-world productivity. `https://www.minimax.io/news/minimax-m25`, February 2026. Accessed: 2026-04-29.

[15] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[16] Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025.

[17] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical programming*, 14(1):265–294, 1978.

[18] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pages 319–327, 2019.

[19] Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *Nejm Ai*, 2(1):AIra2400380, 2025.

[20] Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24315–24329, Vienna, Austria, July 2025. Association for Computational Linguistics.

[21] Harsha Nori, Nicholas King, Scott M. McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[22] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, April 2025. Accessed: 2026-05-03.

[23] OpenEvidence. Openevidence. `https://www.openevidence.com/`, 2024. Accessed: 2026-05-01.

[24] Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026.

[25] David L. Sackett, William M. C. Rosenberg, J. A. M. Gray, R. Brian Haynes, and W. Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[26] Herbert A Simon. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, 70:187–202, 1996.

[27] Karan Singhal, Shekoofeh Tu, Jannik Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Irene Gallego, Kat Litsch, Stephen Pfohl, Humphrey Cole-Lewis, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[28] Earl Steinberg, Sheldon Greenfield, Dianne Miller Wolman, Michelle Mancher, and Robin Graham. *Clinical practice guidelines we can trust*. national academies press, 2011.

[29] Ri Su, Zhao Chen, Caleb Chen Cao, and Lei Chen. Urica: A uniformity region affine identifier capture algorithm for arbitrary region retrieval in pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32777–32786, 2026.

[30] Qwen Team. Qwen3 technical report, 2025.

[31] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[32] Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. Infogain-rag: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7201–7215, 2025.

[33] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[34] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, 2025.

[35] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[36] 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*, pages 6233–6251, 2024.

[37] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[38] Chuang Zhao, Hui Tang, Hongke Zhao, and Xiaomeng Li. Diffmv: A unified diffusion framework for healthcare predictions with random missing views and view laziness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3933–3944, 2025.

[39] Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li. Enhancing precision drug recommendations via in-depth exploration of motif relationships. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8164–8178, 2024.

[40] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457, 2025.

[41] Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*, 2024.

A Related Work

Medical Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) has become a standard paradigm for medical question answering by grounding LLMs in external clinical knowledge sources, such as clinical guidelines, curated evidence summaries, PubMed abstracts, and electronic health records [36, 19, 16]. Medical RAG systems typically combine sparse and dense retrieval and then condition the answer LLM on the retrieved context [40, 34, 2, 38]. Representative systems include MedRAG, which systematically benchmarks medical corpora and retrievers [36]; MedGraphRAG and Co-MedGraphRAG, which incorporate medical knowledge graphs and model collaboration for reasoning [34, 2]; and agent-based medical RAG frameworks that simulate multi-step clinical consultation [39, 38]. These methods improve answer factuality and reasoning, but most still follow a retrieval-centric design: they retrieve or rerank passages independently and retain a fixed top-k context. In dense medical guidelines, this can introduce redundant, partially relevant, or decision-irrelevant snippets as supporting evidence. Our work instead focuses on budget-aware evidence selection: given a recall-oriented candidate pool, MEGA selects a compact subset of high-utility evidence snippets under a token budget for citation-grounded answer generation.

Evidence Selection and Scoring for Medical QA. Evidence selection in RAG is commonly treated as a ranking problem, where passages are scored by lexical similarity, embedding similarity, or cross-encoder rerankers, and the top-ranked items are used as context [36, 16, 6, 3, 29]. Rerankers improve item-level relevance, but they still score each query–snippet pair independently and do not directly optimize the final evidence set under a context budget. Recent work has begun to move from ranking to set selection. For example, SETR formulates retrieval as subset selection and shows that optimizing passage sets can improve downstream RAG quality [11, 4], while submodular and knapsack-based retrieval methods trade off coverage and context cost [9, 17]. InfoGain-RAG further studies document information gain for reranking and filtering in general RAG settings [32]. MEGA shares with Information Bottleneck [31] and Information Pursuit [1] that remain useful for prediction. Information Bottleneck formalizes a trade-off between representation compression and information about a prediction target. Information Pursuit adaptively selects interpretable queries according to their conditional information about that target. MEGA instead performs inference-time selection from retrieved textual evidence: a frozen LLM supplies operational utility scores, token costs make the evidence budget explicit, and a knapsack selector returns directly citable snippets. We do not claim a new mutual-information estimator or theoretical equivalence to these formulations; our contribution is their related compact-information motivation instantiated as gain-aware medical evidence selection.

In medical evidence-grounded QA, however, evidence utility is especially difficult to estimate because snippets can share the same disease, biomarker, or treatment terms while differing in whether they support the final clinical decision. To address this gap, MEGA introduces hidden-state Information Gain Scoring (IGS), which uses a frozen LLM to estimate whether a candidate snippet contributes answer-relevant information beyond surface relevance. Instead of optimizing a nonlinear coverage surrogate, we formulate the final selection step as a gain-aware budgeted $0/1$ knapsack problem, where item utilities come from IGS and token costs impose an explicit evidence budget. This allows an FPTAS selector with a direct $(1-\epsilon)$-approximation guarantee for the stated selection objective, yielding compact evidence subsets that improve both evidence grounding and answer quality.

B Discussion

B.1 Why budget-aware minimal evidence selection matters

MEGA departs from the dominant top-k medical-RAG paradigm in three concrete ways: (i) it scores evidence by *hidden-state information gain* rather than only by surface similarity, (ii) it selects evidence under an explicit token budget rather than retaining a fixed number of independently ranked snippets, and (iii) its utility function jointly accounts for evidence gain, evidence count, and token cost. The empirical effect of these choices is consistent across the three backbones in Table 1: the largest gains appear not only on coarse answer metrics, but also on Evidence F1 and COVID-QA Rouge-L, which are sensitive to *which* sentences the model conditions on. This pattern supports the central design hypothesis of the paper: in dense, terminologically repetitive medical corpora, the bottleneck is not simply whether enough evidence is retrieved, but whether the final evidence set is compact, decision-relevant, and minimally distracting.

B.2 Relationship to existing optimization-based RAG

Compared with submodular-coverage selectors and InfoGain-style rerankers (Appendix D.1), MEGA differs in two respects. First, the value used in selection is not a similarity score or a manually designed coverage proxy, but an IGS utility signal computed from the hidden states of a frozen LLM. This provides an operational estimate of whether a snippet contributes answer-relevant information beyond local relevance. Second, the final selector is formulated as a budgeted $0/1$ knapsack problem and solved with an FPTAS, yielding a direct $(1-\epsilon)$-

approximation guarantee for the stated selection objective. The ablation results in Table 2 show that the selector is especially important for evidence efficiency: removing it greatly increases token usage and substantially reduces Evidence F1. IGS and the selector are therefore complementary: IGS estimates per-snippet evidence utility, while the FPTAS selector converts these utilities into a compact subset under the token budget.

B.3 Limitations

We list the main limitations of MEGA along two axes (scope and evaluation) so that the contributions in the main paper can be interpreted with appropriate caution.

B.3.1 Domain and language scope

Domain coverage. Our two benchmarks cover colorectal oncology (CRC-EVIDENCEQA) and a broader biomedical/COVID-era English subset (MED-EVIDENCEQA). While these settings cover both dense-guideline reasoning and public biomedical QA, we have not yet evaluated MEGA on psychiatry, paediatrics, rare diseases, or noisy sources such as electronic health-record narratives, all of which may exhibit different evidence-density and redundancy patterns.

Bilingual but not fully multilingual. CRC-EVIDENCEQA is bilingual (Chinese/English), but the IGS scorer and retrieval stack are based on Qwen-family models. We have not systematically tested model families with substantially different tokenization, hidden-state geometry, or pretraining distributions. The hidden-state IGS signals may therefore require recalibration when transferred to other model families or languages.

Text-only setting. MEGA currently operates on textual evidence units and does not consume images, tables, radiology scans, pathology slides, or structured EHR fields. Extending gain-aware evidence selection to multimodal medical evidence is an important direction for future work.

Model and language scope. We evaluate frozen scorers from the Qwen, Llama, and Mistral families, with similar performance among the tested 7B8B models. However, Qwen3-4B degrades substantially, and the evaluated families do not cover all tokenization schemes, architectures, languages, or pretraining distributions. Cross-family results therefore reduce, but do not eliminate, scorer dependence. Broader multilingual and domain transfer remain untested.

B.3.2 Evaluation methodology

LLM-as-judge bias. Open-ended answer quality on CRC-EVIDENCEQA is scored by an LLM-as-judge. Although the same judge model and prompt are used for all methods, residual bias toward fluent or verbose answers cannot be fully ruled out. We mitigate this by also reporting Evidence F1, token usage, and blinded clinician evaluation, which directly assess evidence grounding and evidence burden.

B.4 Latency and deployment

The current implementation prioritizes evidence grounding and evidence-token efficiency rather than minimum response time. Hidden-state scoring contributes substantial latency, and the reported experiments do not establish suitability for emergency or other latency-critical workflows. We position MEGA as a research approach for non-emergency evidence review with human oversight, not as a clinically validated autonomous decision system.

C More details of Proof

C.1 NP-hardness

Theorem 3 (NP-Hardness)**.** *The budgeted minimal evidence selection problem in Definition 1 is NP-hard.*

Proof. We prove the NP-hardness by a polynomial-time reduction from the classical 0/1 KNAPSACK DECISION problem, which is NP-complete.

0/1 Knapsack Decision Problem. Given a set of M items

$$\mathcal{I} = \{1, 2, \ldots, M\},$$

where each item i has a positive profit $p_i \in \mathbb{Z}_{>0}$ and a positive weight $b_i \in \mathbb{Z}_{>0}$, together with a capacity $B_0 \in \mathbb{Z}_{>0}$ and a target profit $P_0 \in \mathbb{Z}_{>0}$, the decision problem asks whether there exists a subset $I \subseteq \mathcal{I}$ such that

$$\sum_{i \in I} b_i \leq B_0, \tag{8}$$

and

$$\sum_{i \in I} p_i \geq P_0. \tag{9}$$

Decision version of budgeted evidence selection. We consider the decision version of Definition 1. Given a candidate evidence pool

$$C = \{c_1, c_2, \ldots, c_M\},$$

item utilities $\{v_i(q) = \sum_{k=1}^{K} w_k \tilde{s}_{ik}(q) - \gamma - \mu t_i\}_{i=1}^{M}$, token costs $\{t_i\}_{i=1}^{M}$, a token budget B, and a target utility threshold τ, the question is whether there exists a subset $S \subseteq C$ such that

$$\sum_{c_i \in S} t_i \leq B, \tag{10}$$

and

$$\sum_{c_i \in S} v_i(q) \geq \tau. \tag{11}$$

Reduction. Given an arbitrary instance of 0/1 KNAPSACK DECISION,

$$\left(\{(p_i, b_i)\}_{i=1}^{M}, B_0, P_0\right),$$

we construct an instance of budgeted minimal evidence selection as follows.

For each knapsack item i, we create one evidence candidate c_i. Thus,

$$C = \{c_1, c_2, \ldots, c_M\}.$$

We set the token cost of each evidence unit to the corresponding knapsack weight:

$$t_i = b_i. \tag{12}$$

To map profits into the utility form of our evidence selection objective, define the total profit

$$P_{\text{sum}} = \sum_{j=1}^{M} p_j.$$

We construct a restricted instance with a single gain dimension, i.e.,

$$K = 1, \qquad w_1 = 1.$$

For each evidence unit c_i, we set its normalized gain score as

$$\tilde{s}_{i1}(q) = \frac{p_i}{P_{\text{sum}}}. \tag{13}$$

We further set the penalty parameters to zero:

$$\gamma = 0, \qquad \mu = 0, \tag{14}$$

Under the item-utility definition

$$v_i(q) = \sum_{k=1}^{K} w_k \tilde{s}_{ik}(q) - \gamma - \mu t_i,$$

the utility of evidence c_i becomes

$$v_i(q) = \tilde{s}_{i1}(q) = \frac{p_i}{P_{\text{sum}}}. \tag{15}$$

Finally, we set the evidence budget and target utility threshold as

$$B = B_0, \qquad \tau = \frac{P_0}{P_{\text{sum}}}. \tag{16}$$

This construction is polynomial in the size of the original knapsack instance. Moreover, since all profits are positive integers, $P_{\text{sum}} > 0$, and all normalized gains $\tilde{s}_{i1}(q) \in (0, 1]$ are valid nonnegative gain scores.

Correctness. We now prove that the original 0/1 KNAPSACK DECISION instance is a "yes" instance if and only if the constructed budgeted evidence selection instance is a "yes" instance.

Forward direction. Suppose the original knapsack instance is a "yes" instance. Then there exists a subset $I \subseteq \mathcal{I}$ such that

$$\sum_{i \in I} b_i \leq B_0$$

and

$$\sum_{i \in I} p_i \geq P_0.$$

Construct the corresponding evidence subset

$$S_I = \{c_i \in C : i \in I\}.$$

By the cost mapping in Eq. (12), we have

$$\sum_{c_i \in S_I} t_i = \sum_{i \in I} b_i \leq B_0 = B. \tag{17}$$

Thus, S_I satisfies the evidence budget constraint.

Next, by the utility mapping in Eq. (15), the total utility of S_I is

$$\sum_{c_i \in S_I} v_i(q) = \sum_{i \in I} \frac{p_i}{P_{\text{sum}}} = \frac{1}{P_{\text{sum}}} \sum_{i \in I} p_i. \tag{18}$$

Since $\sum_{i \in I} p_i \geq P_0$, we obtain

$$\sum_{c_i \in S_I} v_i(q) \geq \frac{P_0}{P_{\text{sum}}} = \tau. \tag{19}$$

Therefore, S_I is a feasible solution to the constructed evidence selection decision problem. Hence, the evidence selection instance is also a "yes" instance.

Backward direction. Conversely, suppose the constructed budgeted evidence selection instance is a "yes" instance. Then there exists a subset $S \subseteq C$ such that

$$\sum_{c_i \in S} t_i \leq B$$

and

$$\sum_{c_i \in S} v_i(q) \geq \tau.$$

Let

$$I_S = \{i \in \mathcal{I} : c_i \in S\}$$

be the corresponding subset of knapsack items.

Using Eq. (12) and Eq. (16), the evidence budget constraint gives

$$\sum_{i \in I_S} b_i = \sum_{c_i \in S} t_i \leq B = B_0. \tag{20}$$

Thus, I_S satisfies the knapsack capacity constraint.

Similarly, using Eq. (15), the evidence utility constraint gives

$$\sum_{c_i \in S} v_i(q) = \sum_{i \in I_S} \frac{p_i}{P_{\text{sum}}} = \frac{1}{P_{\text{sum}}} \sum_{i \in I_S} p_i \geq \tau = \frac{P_0}{P_{\text{sum}}}. \tag{21}$$

Multiplying both sides of Eq. (21) by $P_{\text{sum}} > 0$, we obtain

$$\sum_{i \in I_S} p_i \geq P_0. \tag{22}$$

Therefore, I_S satisfies both the knapsack capacity constraint and the target profit constraint. Hence, the original 0/1 KNAPSACK DECISION instance is a "yes" instance.

Conclusion. We have constructed a polynomial-time transformation from any 0/1 KNAPSACK DECISION instance to a restricted instance of budgeted minimal evidence selection such that

$$\exists I \subseteq \mathcal{I} : \sum_{i \in I} b_i \leq B_0 \text{ and } \sum_{i \in I} p_i \geq P_0$$

if and only if

$$\exists S \subseteq C : \sum_{c_i \in S} t_i \leq B \text{ and } \sum_{c_i \in S} v_i(q) \geq \tau.$$

Therefore, solving the decision version of budgeted minimal evidence selection would solve the 0/1 KNAPSACK DECISION problem. Since 0/1 KNAPSACK DECISION is NP-complete, the decision version of budgeted minimal evidence selection is NP-hard. Consequently, the optimization problem in Definition 1 is also NP-hard. □

Algorithm 1 End-to-End Inference of MEGA

Require: Query q, snippet KB $\mathcal{E}$, frozen LLM M_{frozen}, answer LLM M_{ans}, hyperparameters $(\boldsymbol{w}, \gamma, \mu, \epsilon, B, n, b)$
Ensure: Citation-grounded answer a

$\mathcal{Q}(q), \mathcal{K}(q) \leftarrow \textsc{GenerateSubqueries}(q)$ ▷ subqueries and medical keywords
$C \leftarrow \textsc{HybridRetrieve}(\mathcal{Q}(q), \mathcal{K}(q), \mathcal{E}, n)$ ▷ BM25+dense retrieval, dedup
$H_q \leftarrow \{h_M^{(\ell)}(q)\}_{\ell \in \mathcal{L}}$ ▷ shared query encoding
for mini-batch of size b over C **do**
    $\{h_M^{(\ell)}(q \oplus c_i)\}_{\ell \in \mathcal{L}} \leftarrow \textsc{FrozenForward}(M_{\text{frozen}}, q \oplus c_i)$
end for
$G \leftarrow \textsc{ComputeIGSignals}(H_q, \{h_M^{(\ell)}(q \oplus c_i)\})$ ▷ $G \in \mathbb{R}_{\geq 0}^{|C| \times K}$
$\widetilde{S} \leftarrow \textsc{ColumnMinMaxNormalize}(G)$ ▷ $\widetilde{s}_{ik} \in [0, 1]$
for each candidate $c_i \in C$ **do**
    $t_i \leftarrow \textsc{TokenLen}(c_i)$
    $v_i \leftarrow \langle \boldsymbol{w}, \widetilde{\boldsymbol{s}}_i \rangle - \gamma - \mu t_i$ ▷ gain-aware utility
end for
$S_\epsilon \leftarrow \textsc{FptasKnapsackSelector}(C, \boldsymbol{v}, \boldsymbol{t}, B, \epsilon)$ ▷ Algorithm 2
$a \leftarrow M_{\text{ans}}\big(\textsc{Prompt}(q, S_\epsilon)\big)$ ▷ citation-grounded generation
return a

C.2 FPTAS Guarantee

Proof. Let $\widehat{V}(S) = \sum_{c_i \in S} \widehat{v}_i$ be the scaled utility of a subset S. By the definition of flooring, for every candidate $c_i \in \bar{C}$,

$$\Delta \widehat{v}_i \leq v_i(q) < \Delta(\widehat{v}_i + 1).$$

Therefore, for any subset $S \subseteq \bar{C}$,

$$\Delta \widehat{V}(S) \leq V(S) < \Delta \widehat{V}(S) + |S|\Delta,$$

where $V(S) = \sum_{c_i \in S} v_i(q)$. Since $S^\star$ contains at most N' candidates, we have

$$\Delta \widehat{V}(S^\star) \geq V^\star - N'\Delta.$$

The dynamic program finds a feasible subset S_ϵ with maximum scaled utility, so $\widehat{V}(S_\epsilon) \geq \widehat{V}(S^\star)$. Thus,

$$V(S_\epsilon) \geq \Delta \widehat{V}(S_\epsilon) \geq \Delta \widehat{V}(S^\star) \geq V^\star - N'\Delta.$$

Using $\Delta = \epsilon v_{\max}/N'$, we obtain $N'\Delta = \epsilon v_{\max}$. Because candidates with $t_i > B$ are removed, the item with value $v_{\max}$ is feasible by itself; hence $V^\star \geq v_{\max}$. Therefore,

$$V(S_\epsilon) \geq V^\star - \epsilon v_{\max} \geq (1 - \epsilon)V^\star.$$

For complexity, the maximum scaled value satisfies

$$\sum_{c_i \in \bar{C}} \widehat{v}_i \leq \sum_{c_i \in \bar{C}} \frac{v_i(q)}{\Delta} \leq \frac{N' v_{\max}}{\Delta} = \frac{N'^2}{\epsilon}.$$

The DP processes N' items over $O(N'^2/\epsilon)$ scaled-value states, giving time complexity $O(N'^3/\epsilon)$. The DP table requires $O(N'^2/\epsilon)$ memory, or the same order with backtracking information to recover the selected evidence subset. □

D Algorithm of MEGA

We provide the pseudo-code of our MEGA framework in Algorithm 1 and Algorithm 2

Algorithm 2 $(1-\epsilon)$-FPTAS Knapsack Selector

Require: Candidate pool C, item values $\boldsymbol{v}$, token costs $\boldsymbol{t}$, token budget B, approximation factor $\epsilon \in (0,1)$
Ensure: Selected evidence subset $S_\epsilon \subseteq C$

$\bar{C} \leftarrow \{c_i \in C : v_i > 0,\ 0 < t_i \le B\}$ ▷ positive-utility feasible candidates
if $\bar{C} = \emptyset$ **then**
  return $\emptyset$
end if
$N' \leftarrow |\bar{C}|; \quad v_{\max} \leftarrow \max_{c_i \in \bar{C}} v_i; \quad \Delta \leftarrow \epsilon v_{\max}/N'$ ▷ value scaling
$\widehat{v}_i \leftarrow \lfloor v_i/\Delta \rfloor$ for each $c_i \in \bar{C}$
$\widehat{V}_{\max} \leftarrow \sum_{c_i \in \bar{C}} \widehat{v}_i$

▷ DP stores minimum token cost for each scaled utility
$\mathrm{dp}[0] \leftarrow 0; \quad \mathrm{dp}[v] \leftarrow +\infty$ for $v = 1, \ldots, \widehat{V}_{\max}$
$\mathrm{take}[i, v] \leftarrow 0$ for all i, v
for each item $c_i \in \bar{C}$ **do**
  for $v = \widehat{V}_{\max}$ **down to** $\widehat{v}_i$ **do**
    if $\mathrm{dp}[v - \widehat{v}_i] + t_i < \mathrm{dp}[v]$ **then**
      $\mathrm{dp}[v] \leftarrow \mathrm{dp}[v - \widehat{v}_i] + t_i$
      $\mathrm{take}[i, v] \leftarrow 1$
    end if
  end for
end for
$v^\star \leftarrow \max\{v : \mathrm{dp}[v] \le B\}$
$S_\epsilon \leftarrow \textsc{Backtrack}(\mathrm{take}, v^\star)$
return S_ϵ

D.1 Experiment Implementation Details

This appendix specifies the five baselines compared against MEGA in (Table 1). All baselines share the same backbone language models (GPT-4.1 and Qwen3.6-35B-A3B), the same embedding model (Qwen3-Embedding-8B, D=4,096), the same retriever cache, and the same LLM-as-judge prompt (Appendix G.4), so that all reported deltas are attributable to the evidence-selection mechanism rather than to model or prompt drift.

D.1.1 Context+LLM (no retrieval)

Method. A no-retrieval upper bound that puts *all* available evidence into the LLM context. On CRC-EVIDENCEQA we concatenate the full L1 colorectal-cancer guideline corpus; on MED-EVIDENCEQA we concatenate the `context_sentences` field of the queried item.

Pipeline.

1. Load the entire corpus into a single string C.
2. Truncate to the back-bone-specific context limit L_{ctx} (GPT-4.1: $\tilde{9}00$ k chars; Qwen3.6:$\tilde{2}40$ k chars). Truncation is *tail-cut* after sentence boundary detection.
3. Generate the answer with the prompt of Appendix G.2 / G.3.

Role. Quantifies the cost-quality penalty of giving up evidence selection altogether and the noise tolerance of each back-bone.

D.1.2 Simple-RAG (dense top-k + reranker)

Method. A dense-retrieval, single-stage RAG baseline that mirrors common production practice.

Pipeline.

1. *Index*: pre-built ChromaDB collection over L1 chunks (CRC-EVIDENCEQA) or per-query sentences (MED-EVIDENCEQA), embedded with Qwen3-Embedding-8B.

2. *Retrieve*: dense cosine top-k_1=20 via Chroma.
3. *Rerank*: Qwen3-Reranker-8B cross-encoder, keep top k_2=8.
4. *Generate*: feed the k_2 passages, formatted as `[i] (filename, section)\n content`, to the same answer prompt (Appendix G.2/G.3).

Hyper-parameters. k_1=20, k_2=8, reranker temperature 0; identical across both datasets.

Role. Isolates the contribution of the IGS scorer and the FPTAS knapsack: Simple-RAG and MEGA share identical retrieval and back-bone, so any quality gap is attributable to selection logic.

D.1.3 DoctorRAG [13]

Original method. DoctorRAG (NeurIPS 2025) combines (i) a Patient Base of prior cases routed by ICD-10 codes, (ii) a Knowledge Base over medical references, and (iii) Med-TextGrad, a critic-generator iterative refinement loop over the draft answer.

Our port. The Patient Base is empty in both benchmarks (no prior-case database is supplied for CRC or for the open-domain split), and the domain is single (oncology / general medicine), so we drop ICD-10 routing. The Knowledge Base and Med-TextGrad refinement are kept verbatim.

Pipeline.

1. *Index*: build a numpy-FAISS-style KB over the same chunks as Simple-RAG, with L_2-normalised Qwen3-Embedding-8B vectors.
2. *Retrieve*: inner-product top-k=5 (CRC-EVIDENCEQA) / top-k adaptive (MED-EVIDENCEQA, k=5 by default).
3. *Draft*: generate an initial answer with the RAG_SYS_EN/ZH system prompt ("Answer ONLY based on the provided knowledge passages; do not fabricate").
4. *Med-TextGrad*: for T=3 iterations, prompt the same back-bone in critic mode to list 3–6 specific deficiencies of the current draft (accuracy, completeness, citation), then prompt it again in generator mode to produce a revised full answer conditioned on the critique and the retrieved knowledge.

Hyper-parameters. k=5, $T_{\text{Med-TextGrad}}$=3, critic max-tokens 400, generator max-tokens 1 024, τ_{critic}=$\tau_{\text{generator}}$=0.

Note on citation behaviour. The DoctorRAG prompt does not require "[N]" citation markers, and Med-TextGrad subsequently rewrites the draft into prose; in our runs $\sim 90\%$ of final answers contain no parseable citation, which is why we report only the selector-level evidence-F1 for DoctorRAG in Table 1.

D.1.4 MedRAG [40]

Original method. MedRAG (WWW 2025) is a knowledge-graph elicited RAG: an offline (subject, relation, object) KG is built from clinical text; at query time the system embedding-matches the query to KG nodes, walks one or two hops to harvest triples, and feeds *both* the triples and a parallel set of dense-retrieved chunks into the LLM prompt ("KG-elicited prompting").

Our port. *KG construction*: the original codebase ships an Excel KG that does not cover our domains, so we re-build the KG with an LLM extractor. For CRC-EVIDENCEQA we extract once over all L1 chunks. For MED-EVIDENCEQA we extract a small per-query KG from the provided `context_sentences`. The extractor prompt restricts the relation set to ten clinically meaningful predicates (`is_a, has_symptom, has_diagnostic_test, has_treatment, has_complication, contraindicates, recommends, requires, includes, mechanism_of`) and caps each chunk at 10 triples.

Retrieval: at query time we embed the question, dot it against the KG-node embedding matrix, take top-5 nodes, and harvest all triples touching those nodes (deduped, capped at $|\mathcal{T}|\leq 10$). In parallel, we embed and dot the chunk-text matrix, taking top-k_c chunks (auto-set to $\lceil 0.3 \cdot N_{\text{sents}} \rceil$ on MED-EVIDENCEQA, fixed k_c=5 on CRC-EVIDENCEQA).

Prompting: the prompt block lists the triples followed by the numbered passages, and asks for a structured answer that cites either triple text or passage indices.

Hyper-parameters. top-k_{nodes}=5, max_triples=10, k_c=$\lceil 0.3 N_{\text{sents}} \rceil$ (MedEvidenceQA) or 5 (CRC), KG-extractor max-tokens 600, τ=0.

Note. The selected-indices we report for MedRAG are exactly the indices of the dense-retrieved chunks (the KG-triple branch does not have a one-to-one sentence mapping). This is the convention also used by the original codebase for evidence comparison.

D.1.5 MMed-RAG (text-only) [35]

Original method. MMed-RAG (ICLR 2025) is a multi-modal medical RAG with three components: (a) a domain-aware retriever that routes queries among radiology / pathology / ophthalmology corpora, (b) adaptive context selection that picks k based on top-1 retrieval similarity, and (c) RAG-PT, a two-stage DPO preference-tuning of the generator.

Our port. We omit the multi-modal pathway (no images in our benchmarks) and DPO preference fine-tuning (we keep all back-bones frozen). We retain (a) and (b) and replace DPO with self-consistency sampling, which keeps the spirit of "preference at decoding time" while remaining training-free.

Pipeline.

1. *Domain router*: tag every chunk with a domain label (`source` field on MED-EVIDENCEQA; constant `colorectal` on CRC-EVIDENCEQA). At query time, add a constant boost β=0.1 to the cosine similarity of in-domain chunks.
2. *Adaptive top-k*: k=2 if top-1 sim > 0.75; k=4 if > 0.55; k=8 otherwise (capped at the corpus size).
3. *Self-consistency generation*: sample n=3 answers from the same prompt at τ=0.5, embed all three with the same embedding model, and return the medoid (the answer with the highest sum of pairwise similarities).

Hyper-parameters. β=0.1, adaptive-k thresholds $(0.75, 0.55)$, n_{samples}=3, τ_{sample}=0.5, embedding-medoid selection.

D.1.6 Implementation Conventions Common to All Baselines

Embedding and reranker. Every baseline uses the **same** Qwen3-Embedding-8B retriever and (when applicable) Qwen3-Reranker-8B; embeddings are L_2-normalised and cached on disk to remove API noise.

Backbone. The same generator backbone is used for retrieval expansion (where applicable), drafting, refinement, and answer synthesis within each row of Table 1. We cache the chat completions in keyed by (model, messages, temperature, max_tokens) so that any rerun yields bit-identical answers.

LLM-as-judge. All four 1–10 dimensions on CRC-EVIDENCEQA use the prompt of Appendix G.4, τ=0, max-tokens 400. The judge model is GPT-5.1 across all rows so that judge bias is constant.

Failure handling. Any per-question exception is logged and the record is dropped from aggregation rather than imputed; the "n =" column in our internal aggregation table reports the true denominator per cell.

D.2 Computing Resources

Hardware. All experiments were run on a single node with 4×NVIDIA A800-80GB GPUs, 128 CPU cores, and 1 TB of system memory. GPUs were used for (i) Qwen3-Embedding-8B encoding, (ii) the frozen IGS scorer (Qwen3-8B by default; optionally Qwen3-32B), and (iii) Qwen3.6-35B-A3B answer generation when run locally. GPT-4.1 and MiniMax-M2.5-230B answer generation are routed via their respective hosted APIs. The FPTAS evidence selector and metric aggregation run entirely on CPU.

No model training. MEGA does *not* fine-tune any parameters. The IGS scorer, the embedding model, the reranker, and all answer-generation back-bones are frozen. The only learned artefact is an offline, retrieval-time vector index (ChromaDB / numpy-FAISS) over the corpus, which is built once per benchmark and cached on disk.

Cost decomposition. Wall-clock cost per query is dominated by two LLM forward stages: (i) the IGS hidden-state pass, which scales with the candidate pool size N and the chosen IGS scorer (a small constant number of forwards); and (ii) answer generation by the backbone LLM. Retrieval embedding lookup, reranker scoring, FPTAS selection, and metric aggregation is all sub-second per query and contributes a negligible fraction of the total runtime.

Storage. The cached embedding matrix is $\sim$ 300 MB for the L1 colorectal corpus and is shared across all baselines. The LLM chat-completion cache guarantees bit-identical re-runs; cache hits incur only SQLite I/O.

Reproducibility. All random seeds (sampling, FPTAS tie-break, self-consistency) are fixed to 42. Re-running an experiment from a warm cache requires neither GPU memory beyond what is needed to load the answer backbone nor any additional training compute.

Table 3: IGS weight ablation on the test sample.

Setting	Ev-R	Ev-P	Ev-F1	#Ev	Tokens	Q
Default $(0.35, 0.35, 0.15, 0.15)$	0.71	0.78	0.74	3.4	612	7.92
Uniform $(0.25, 0.25, 0.25, 0.25)$	0.69	0.75	0.72	3.6	641	7.78
Drop w_1 (cosine shift)	0.66	0.71	0.68	3.7	671	7.51
Drop w_2 (drift)	0.64	0.70	0.67	3.8	689	7.43
Drop w_3 (cross-layer coherence)	0.69	0.76	0.72	3.5	627	7.81
Drop w_4 (layer-grad.)	0.70	0.76	0.73	3.5	624	7.83
Only w_2 (drift only)	0.65	0.69	0.67	4.0	712	7.40

Table 4: Sensitivity to the count penalty γ.

γ	Ev-R	Ev-P	Ev-F1	#Ev	Tokens	Q
0.00	0.74	0.69	0.71	4.4	768	7.69
0.25	0.72	0.75	0.73	3.8	681	7.85
0.5 (default)	0.71	0.78	0.74	3.4	612	7.92
0.75	0.66	0.79	0.72	2.9	528	7.74
1.0	0.58	0.81	0.68	2.3	421	7.40

E Additional Experimental Results

E.1 Hyperparameter Sensitivity

This appendix studies the sensitivity of MEGA to its main hyperparameters: the IGS dimension weights w_1, w_2, w_3, w_4, the evidence-count penalty γ, the token-length penalty μ, and the FPTAS approximation factor ϵ. All sweeps are run on a held-out dev split of MED-EVIDENCEQA (n=50, seed-disjoint from the n=200 test split) with Qwen3.6-35B-A3B as the answer back-bone, Qwen3-Embedding-8B/Qwen3-Reranker-8B as the retriever, and Qwen3-8B as the frozen IGS scorer. One hyperparameter is varied at a time; all others are kept at their default values (w_1=w_2=0.35, w_3=w_4=0.15, γ=0.5, μ=0.25, ϵ=0.1, token budget B=2 048).

E.1.1 IGS dimension weights w_1, w_2, w_3, w_4

Setup. We toggle each of the four IGS signals (*cosine shift* w_1, *representation drift* w_2, *representation cross layer coherence* w_3, *cross-layer information gradient* w_4) on or off, and also test a uniform allocation, to quantify the marginal contribution of each signal.

Analysis. Removing *representation drift* (w_2) or *cosine shift* (w_1) causes the largest drops (-0.07 and -0.06 Ev-F1 respectively), confirming that the two first-order signals carry most of the gain information. Removing *cross layer coherence* (w_3) or *layer gradient* (w_4) yields small but consistent regressions ($-0.02/-0.01$ Ev-F1), indicating they act as complementary, lower-magnitude refinements. The default weighting $(0.35, 0.35, 0.15, 0.15)$ matches or beats every alternative on Ev-F1 and Q, supporting our default.

E.1.2 Evidence-count penalty γ

Analysis. γ controls the cost of an additional evidence slot. Large γ produces compact sets (#Ev 2.3) but at the cost of recall and answer quality; small γ inflates the set without improving F1. The default γ=0.5 achieves the best Ev-F1/Q trade-off.

E.1.3 Token-length penalty μ

Analysis. The length penalty primarily reduces *tokens* without changing #Ev as steeply as γ, because it preferentially prunes long candidates. Quality degrades only when μ is large enough to evict mid-length but informative chunks ($\mu \geq 0.2$).

E.1.4 FPTAS approximation factor ϵ

Analysis. Because the candidate pool after reranking is small ($N \leq 20$), even ϵ=0.5 deviates by less than 5% from the exact knapsack on the surrogate, and end-to-end Ev-F1/Q are essentially flat. The cleanup pass closes the remaining gap because items with negative true marginal gain are removed regardless of ϵ. We keep ϵ=0.1 as a conservative default; larger values are safe in practice.

Table 5: Sensitivity to the token-length penalty μ.

μ	Ev-R	Ev-P	Ev-F1	#Ev	Tokens	Q
0.00	0.73	0.74	0.73	3.6	798	7.81
0.15	0.72	0.77	0.74	3.5	689	7.89
0.25 (default)	0.71	0.78	0.74	3.4	612	7.92
0.50	0.68	0.78	0.72	3.2	510	7.80
1.00	0.61	0.78	0.69	2.9	402	7.52

Table 6: Effect of the FPTAS factor ϵ on the surrogate objective and runtime. "Surrogate gap" is the relative deficit of the FPTAS solution against an exact $0/1$-knapsack baseline on the linearised objective; "solver time" is the per-query DP wall-time on a single CPU core.

ϵ	Surrogate gap (%)	Ev-F1	#Ev	Tokens	Solver time (ms)
0.01	< 0.05	0.74	3.4	612	7.8
0.05	0.4	0.74	3.4	614	2.1
0.10 (default)	0.9	0.74	3.4	615	1.0
0.25	2.1	0.73	3.5	618	0.5
0.50	4.6	0.73	3.5	626	0.3

E.2 Summary

The sensitivity sweeps confirm three takeaways. (i) *Drift* and *cosine shift* are the dominant IGS signals; the cross layer coherence and layer-gradient terms add small but consistent gains. (ii) λ, γ, and μ each form an inverted-U around the default and the optima are mutually compatible: MEGA does not require fine joint tuning. (iii) ϵ has only a marginal effect because N is small after reranking, so the FPTAS factor can be chosen for solver speed without quality penalty.

E.3 Robustness of IGS and Runtime Analysis

The Information Gain Scoring (IGS) module of MEGA relies on the hidden states of a frozen LLM scorer. This scorer does not generate answers or receive gradient updates; it only provides per-layer representations for computing the IGS signals, including cosine shift, representation drift, cross-layer coherence, and cross-layer information gradient. We study whether the main results depend on the choice of this frozen scorer, specific layers, or judgment model.

E.3.1 Setup

We rerun the full MEGA pipeline on each benchmark slice while changing only the frozen IGS scorer or the selected region of layers. The answer-generation backbone is fixed to GPT-4.1, and all retrieval settings, selector hyperparameters $(\epsilon, \boldsymbol{w}, \lambda, \gamma, \mu, B)$, prompts, and random seeds are kept unchanged. We compare Qwen3-8B and Qwen3-32B on the same $4\times$NVIDIA A800-80GB node.

We report Evidence-F1, answer quality, selected-set size, and end-to-end runtime per query. Evidence-F1 compares the selected evidence set with the gold supporting evidence. Answer quality is reported on a unified 0–10 scale: CRC-EvidenceQA uses the LLM-as-Judge overall score, PubMedQA uses accuracy divided by 10, and COVID-QA uses Rouge-L divided by 10. Selected-set size is the average number of evidence sentences selected by the FPTAS knapsack selector.

E.3.2 Results of different sizes of Qwen Scorer

Figure 6 and Table 7 show that the larger Qwen3-32B scorer yields consistent but small improvements. Across the four slices, Evidence-F1 increases by $+0.4$ to $+0.5$ points, and answer quality improves by only $+0.05$ to $+0.08$ on the 0–10 scale. The selected-set size changes by at most 0.1 sentences, suggesting that the evidence subsets chosen by the FPTAS knapsack selector are largely stable across the two scorers.

The cost difference is much larger. The full pipeline with the Qwen3-8B scorer requires approximately 32.0 s per query, compared with 118.4 s using Qwen3-32B. These are end-to-end response times, including evidence processing and complete answer generation, rather than IGS-only runtimes. The larger scorer therefore yields modest quality gains at a substantially higher pipeline cost. Qwen3-8B requires about 32.0 s per query on average, while Qwen3-32B requires about 118.4 s per query, corresponding to a $3.7\times$ slowdown. Thus, the larger scorer provides only marginal quality gains at substantially higher inference cost. We therefore use

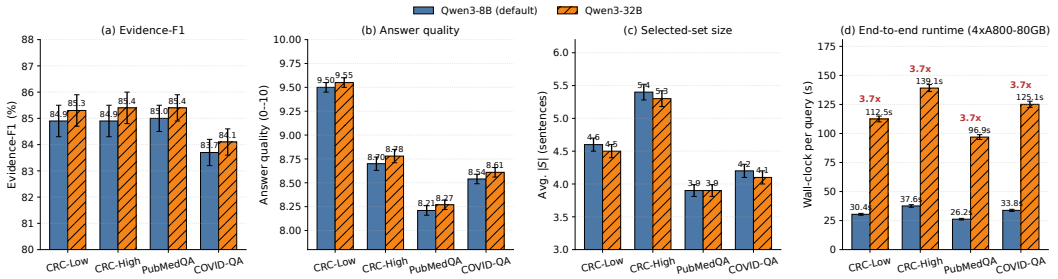


Figure 6: **Sensitivity of MEGA to the frozen IGS scorer.** Qwen3-8B is the default scorer used in the main experiments. Qwen3-32B provides marginally higher evidence and answer quality, but incurs a substantially larger end-to-end runtime. Bars are mean ± SEM. Panel (d) reports wall-clock time per query on $4\times$A800-80GB.

Table 7: **Per-slice quality and cost of different frozen IGS scorers.** Δ denotes Qwen3-32B minus Qwen3-8B.

	Evidence-F1 (%)		Answer quality (0–10)		Wall-clock (s/query)
Slice	8B	32B (Δ)	8B	32B (Δ)	8B / 32B
CRC-Low	84.9	85.3 (+0.4)	9.50	9.55 (+0.05)	30.4 / 112.5
CRC-High	84.9	85.4 (+0.5)	8.70	8.78 (+0.08)	37.6 / 139.1
PubMedQA	85.0	85.4 (+0.4)	8.21	8.27 (+0.06)	26.2 / 96.9
COVID-QA	83.7	84.1 (+0.4)	8.54	8.61 (+0.07)	33.8 / 125.1
Average	84.6	85.1 (+0.4)	8.74	8.80 (+0.07)	32.0 / 118.4

Qwen3-8B as the default frozen IGS scorer in the main paper. These results suggest that MEGA 's gains over RAG baselines are not an artifact of using an unusually large hidden-state scorer.

E.3.3 Results of different families of IG Scorer

We evaluated frozen scorers from Qwen, Llama, and Mistral, spanning 4B–32B parameters. Table 8 shows that the tested 7B–8B scorers yield similar results across model families. Relative to Qwen3-8B, Llama-3.1-8B changes answer quality by 0.05 and Evidence F1 by 0.7 points, while Mistral-7B changes them by 0.22 and 1.4 points. Increasing Qwen from 8B to 32B yields only +0.06 answer quality and +0.5 Evidence-F1 points at 3.7 times the reported runtime. In contrast, Qwen3-4B degrades substantially. These results support transfer across the tested comparably sized families, while identifying scorer capacity as a limitation.

Table 8: Robustness to the frozen IGS scorer. The answer generator and candidate pools are held fixed. KP denotes key-point coverage.

Frozen scorer	**Params**	**AUROC**	**Spearman**	**Ans. Quality**	**Evi. F1**
Qwen3-8B	8B	0.77	0.66	8.74	84.6
Qwen3-4B	4B	0.57	0.54	6.38	66.4
Qwen3-32B	32B	0.79	0.68	8.80	85.1
Llama-3.1-8B-Instruct	8B	0.75	0.66	8.69	83.9
Mistral-7B-v0.1	7B	0.74	0.63	8.52	83.2

E.3.4 Sensitivity to the layer region used for IGS

We partition the normalized model depth into four equal-width regions and additionally test the default middle half, $[L/4, 3/4L]$. For the quarter-depth settings, we uniformly sample the same number of layers. Only the IGS layer set changes; the scorer, answer backbone, retrieval results, selector hyperparameters, prompts, and random seeds remain fixed.

The middle regions clearly outperform the early and late regions, and the default middle-half setting performs best on all metrics. Compared with the strongest individual region, it improves KP-AUROC by 0.05, Spearman

Table 9: Sensitivity to the hidden-layer region used by IGS.

Scoring region	KP-AUROC↑	KP-Spearman ρ ↑	Ans. Quality↑	Evi. F1 (%)↑
Early $[0, \frac{1}{4}]$	0.54	0.35	7.96	68.9
Lower-middle $[\frac{1}{4}, \frac{1}{2}]$	0.68	0.55	8.39	78.6
Upper-middle $[\frac{1}{2}, \frac{3}{4}]$	0.72	0.59	8.53	81.2
Late $[\frac{3}{4}, 1]$	0.57	0.38	8.08	70.7
Middle-half $[\frac{1}{4}, \frac{3}{4}]$, default	**0.77**	**0.66**	**8.74**	**84.6**

correlation by 0.07, answer quality by 0.21, and Evidence F1 by 3.4 points. Early layers mainly encode lexical and local contextual features, while late layers are more oriented toward next-token prediction and response realization. Middle layers better capture queryevidence integration. Aggregating the broader middle half also allows cross-layer coherence and information-gradient signals to model a more stable evidence-integration trajectory. These results support $[4/L, 3/4L]$ as a principled default.

E.3.5 Latency Analysis of MEGA

We profile the default pipeline with Qwen3-8B as the frozen scorer and GPT-4.1 as the answer generator. End-to-end latency is measured from receipt of a query until the last answer token is returned. Stage timings separate retrieval, IGS computation, evidence selection, and answer generation. These response measurements are distinct from the offline judge-based evaluation of answer quality.

Table 10: Stage-wise latency of MEGA.

Stage	Average latency
Hybrid retrieval	2.9 s
IGS computation	14.4 s
FPTAS selection	0.008 s
Answer generation	14.7 s
Time to first token	**17.3 s**
Full-response latency	**32.0 s**

Table 10 shows that IGS computation and answer generation account for most of the approximately 32.0 s response time, taking 14.4 s and 14.7 s, respectively. Retrieval takes 2.9 s and the selection module contributes 8 ms. Thus, the reduction in evidence-token usage does not imply a reduction in end-to-end latency. With streaming enabled, the measured time from query receipt to the first answer token is 17.3 s. This measurement includes preprocessing and the answer models first-token delay. MEGA trades additional evidence-scoring computation for more compact evidence and stronger grounding in the evaluated setting.

E.4 Budget and Selector Analysis

E.4.1 Hard budgets and soft penalties

We isolate the hard evidence-token constraint from the soft count and length penalties on the same 300-question evaluation subset. Candidate pools, the Qwen3-8B scorer, the GPT-4.1 answer generator, and the FPTAS factor $\epsilon = 0.1$ are fixed. The hard-budget-only variant sets $\gamma = \mu = 0$. The penalties-only variant retains both penalties but sets the optimization budget to the total candidate-token cost, making the hard constraint inactive. The full variant combines both mechanisms. We evaluate compactness, quality, and violations of the deployment evidence-token cap.

As shown in Table 11, the hard constraint alone enforces the cap but selects substantially more evidence: 1,016 tokens compared with 612 for the full model. Soft penalties alone reduce the mean length but do not enforce the cap on every query, with 11.3% exceeding the evaluated threshold. Combining both mechanisms achieves the highest Evidence F1 and answer quality while avoiding observed cap violations. This distinguishes per-query feasibility from preferences for compact subsets within the feasible region.

Table 11: Hard budgets and soft penalties control different aspects of compactness.

Setting	**Mean evidence tokens**	**Queries above 1,043 tokens (%)**	**Evidence F1 (%)**	**Answer quality**
Hard B only	1,016	0	81.9	8.57
Penalties only	635	11.3	79.5	8.28
Hard B + penalties	612	0	84.6	8.74

E.4.2 Sensitivity to the evidence-token budget

We sweep the evidence-token budget while holding candidate pools, IGS utilities, the answer generator, and all other hyperparameters fixed. The sweep changes the feasible evidence subset, not the scoring model or the retrieval results. We report the selected-token count, answer quality, and Evidence F1 at each budget.

Table 12: Sensitivity to the hard evidence-token budget B.

Budget B	Mean selected tokens	Answer quality	Evidence F1 (%)
256	239	7.74	65.7
384	361	8.17	74.6
512	476	8.48	80.9
768	598	8.69	83.8
1,024	608	8.73	84.4
1,536	611	8.74	84.5

As shown in Table 12, performance improves rapidly as the budget increases from 256 to 768 tokens and largely saturates beyond approximately 1,024 tokens. For example, Evidence F1 rises from 65.7 to 83.8 between B=256 and B=768, but only from 84.4 to 84.5 between B=1,024 and B=1,536. Meanwhile, the average selected length remains close to 600 tokens at the larger budgets. This is consistent with the role of B as an upper bound rather than a target length: additional capacity does not force the selector to include non-positive-utility evidence.

E.4.3 Controlled comparison of selection algorithms

We compare three selectors using identical candidate pools, utilities v_i, token costs, and budgets. Greedy-Value ranks candidates by v_i, whereas Greedy-Density ranks them by v_i/t_i. FPTAS uses $\epsilon = 0.1$. We evaluate both the selected utility and downstream answer quality. For instances with positive optimal utility, the relative surrogate gap is $(V^*V(S))/V*$, where V^* is obtained using exact 0/1-knapsack optimization.

Table 13: Quality and evidence grounding under matched tight budgets.

	$B = 256$		$B = 512$	
Method	Quality	F1 (%)	Quality	F1 (%)
Reranker-TopBudget	7.18	52.8	7.84	64.6
DoctorRAG-TopBudget	7.36	57.4	8.01	68.9
Greedy(IGS)	7.55	62.8	8.31	77.4
MEGA	7.74	65.7	8.48	80.9

As shown in Table 13, relative to Greedy-Density, FPTAS improves Evidence F1 by 2.0 points and answer quality by 0.09 while selecting 52 fewer evidence tokens on average. It also reduces the surrogate gap from 4.2% to 0.9%. These results isolate the benefit of the selection algorithm under matched evidence scores. The approximately 1 ms solver runtime is small compared with hidden-state scoring and answer generation. The approximation guarantee concerns the fixed additive surrogate, and the downstream quality benefit is an empirical observation.

E.4.4 Budget-matched comparison

We additionally compare methods at matched evidence-token budgets of 256 and 512. For ranking-based baselines without an explicit budgeted selector, we retain evidence in the original ranking order until adding the next unit would exceed the cap. All reported subsets must satisfy the same evidence-token constraint. MEGA achieves the highest answer quality and Evidence F1 among the evaluated methods at both budgets. Relative to

Table 14: Selectors evaluated with the same IGS utilities.

Selector	Surrogate gap (%)	Answer quality	Evidence F1 (%)	Mean evidence tokens	Solver time (ms)
Greedy-Value	8.7	8.57	81.0	705	0.06
Greedy-Density	4.2	8.65	82.6	664	0.09
FPTAS, $\varepsilon = 0.1$	0.9	8.74	84.6	612	1.0

Greedy(IGS), its Evidence-F1 advantage is 2.9 points at B=256 and 3.5 points at B=512, with answer-quality gains of 0.19 and 0.17, respectively. These results support the benefit of budget-aware combination selection under tight evidence constraints.

E.5 Qualitative Success and Failure Analysis

We present one success case and two failure cases to distinguish candidate retrieval errors from utility-estimation errors. The evidence descriptions and answer outcomes below are abridged summaries; IGS orderings are reported qualitatively rather than as numerical scores.

Case 1: Condition-matched evidence selection (success)

Dataset / ID. CRC-EvidenceQA / QID 28.

Question. For MSI-H/dMMR metastatic CRC, what is the preferred first-line therapy?

Candidate evidence. The pool contains snippets matching all decision-critical conditions—biomarker status, metastatic disease, and treatment line—alongside partially applicable and redundant snippets.

IGS and selection. IGS ranks the condition-matched evidence above partially applicable candidates. Budgeted selection retains a compact supporting subset.

Outcome and interpretation. The retained evidence supports the correct answer. This case illustrates how question-specific utility estimation and budgeted selection work together when the candidate pool contains the required support.

Case 2: Missing reference support (retrieval failure)

Dataset / ID. CRC-EvidenceQA / QID 13.

Question. What toxicity of oxaliplatin requires the closest clinical monitoring?

Reference support. The annotation includes two complementary units: e_1, describing acute and delayed peripheral sensory neuropathy; and e_2, describing persistent or cumulative neuropathy that may impair daily functioning. Here, e_1 and e_2 are local display labels.

Retrieval and selection. The candidate pool contains e_1 but omits e_2. MEGA retains e_1, leaving incomplete coverage of the annotated support.

Outcome and interpretation. The failure concerns evidence coverage, not necessarily answer correctness. Neither IGS nor FPTAS can recover a missing unit from outside the candidate pool; the bottleneck in this case occurs before utility estimation and subset selection.

Case 3: Overvaluing a near-miss snippet (IGS failure)

Dataset / ID. Med-EvidenceQA / `covid_qa_50`.

Question. How was cross-reaction detection determined?

Reference versus near-miss evidence. The reference describes applying JEV, YF, and DV cultures to determine cross-reaction. The near-miss snippet instead describes determining the proper spotting concentration by minimizing cross-reaction. The passages share terminology but concern different experimental steps.

IGS and selection. IGS overvalues the near-miss snippet, which FPTAS retains under the token budget.

Outcome and interpretation. The generated answer describes the wrong procedure. Semantic similarity does not ensure question-specific support, and optimizing a fixed-score objective does not independently correct a utility-estimation error.

Component roles and scope. Query expansion and hybrid retrieval determine which evidence is available; IGS estimates its question-specific utility; FPTAS optimizes the fixed-score selection objective under the

token budget. FPTAS does not explicitly model pairwise redundancy or marginal coverage. These examples complement the component ablations in Section 5.4 and the fixed-score selector comparison in Appendix E.4; they do not isolate the causal contribution of each IGS signal or establish the prevalence of either failure mode.

F Benchmark Construction and Statistics

This appendix documents how the two benchmarks used throughout the paper – CRC-EVIDENCEQA and MED-EVIDENCEQA – are constructed, what their evidence annotations look like, and how they are split for evaluation. The CRC-EVIDENCEQA benchmark is clinician-collaborative and contributed by colorectal oncology experts; the MED-EVIDENCEQA benchmark is built fully automatically from public open-data sources.

F.1 CRC-EvidenceQA: clinician-collaborative benchmark

F.1.1 Source and authoring protocol

CRC-EVIDENCEQA is authored in collaboration with practising colorectal-cancer clinicians. The authoring loop has three steps. *(i) Curriculum draft*: clinicians draft open-ended questions that mirror real outpatient or multi-disciplinary-team (MDT) scenarios in colorectal cancer, spanning initial work-up, staging, neoadjuvant and adjuvant therapy, surgical sequencing, molecular stratification, follow-up, and evidence-edge cases. *(ii) Evidence anchoring*: for every question, clinicians identify the specific guideline paragraphs from the official Chinese CSCO and English NCCN colorectal-cancer guidelines that must be cited to justify a complete answer; these are stored verbatim as the gold supporting snippets. *(iii) Bilingual mirroring*: each question, reference answer, and gold snippet is written natively in Chinese and translated / verified into English by a second clinician, so that a model is evaluated on the *same* clinical content under both languages.

F.1.2 Schema

Every CRC-EVIDENCEQA item carries the following field:

- `id` – stable item identifier, e.g. `crc_eqa_01`.
- `category` – coarse clinical theme; six categories are used (Section F.1.4).
- `question_zh`, `question_en` – the bilingual clinical question.
- `answer_zh`, `answer_en` – the clinician reference answer in both languages.
- `snippets_zh`, `snippets_en` – the gold guideline-anchored evidence snippets, each verbatim from CSCO / NCCN, cited by the reference answer.

F.1.3 Splits and difficulty stratification

The benchmark contains 200 open-ended clinical questions, 100 in Chinese and 100 in English. For each language, the 100 questions are stratified by clinician-assigned difficulty into a LOW-difficulty subset of 50 items and a HIGH-difficulty subset of 50 items. LOW questions ask for guideline-direct lookups and routine clinical decisions; HIGH questions require multi-step reasoning across staging, molecular stratification, treatment sequencing, follow-up, and evidence-based corner cases (cf. main paper, Section 4.1). Throughout the paper, every reported CRC-EVIDENCEQA-LOW or CRC-EVIDENCEQA-HIGH number is computed on the full 50-question subset (no sub-sampling).

F.1.4 Categories

The clinician-curated category schema is:

1. *Initial work-up and staging*;
2. *Pathology and molecular stratification*;
3. *Colon-cancer perioperative management*;
4. *Rectal-cancer combined-modality treatment*;
5. *Metastatic, conversion and later-line therapy*;
6. *Follow-up, hereditary risk and evidence boundaries*.

F.1.5 Per-item statistics

The released sample at contains a representative slice of items spanning all six categories. On that sample, every item carries five gold guideline snippets, the mean Chinese question length is ≈ 64 characters, and the mean Chinese reference-answer length is ≈ 90 characters, indicating that CRC-EVIDENCEQA questions are short

Table 15: Summary of CRC-EVIDENCEQA. Per-item snippet, question, and answer statistics are computed on the released sample.

Property	Value
Total questions	200
Chinese / English split	100 / 100
Low / High per language	50 / 50
Number of categories	6
Source guidelines	CSCO 2025, NCCN 2025 (colorectal cancer)
Average Gold snippets / item	4.3
Mean Chinese question length	≈ 64 characters
Mean Chinese answer length	≈ 90 characters
Authoring	Clinician-drafted, evidence-anchored, bilingually mirrored

but clinically dense and that the gold answer prose is concise rather than essay-style. Table 15 summarises the benchmark.

F.1.6 Evaluation use

CRC-EVIDENCEQA is used to evaluate (a) answer quality, scored 1–10 along four dimensions (accuracy, completeness, citation, overall) by a held-fixed LLM-as-judge (Appendix D.1); and (b) evidence-F1 of the selector against the gold snippet set. Both metrics are computed on every one of the 50 questions per subset.

F.2 Med-EvidenceQA: open-data benchmark

F.2.1 Sources and motivation

MED-EVIDENCEQA complements the clinician-curated CRC-EVIDENCEQA with a larger, fully open-data benchmark that ships with sentence-level gold evidence annotations. It is synthesised from two public biomedical-QA corpora: **PubMedQA** [8], the expert-labelled `pqa_labeled` subset of yes/no/maybe biomedical research questions over PubMed abstracts; and **COVID-QA** [15], an extractive QA dataset over COVID-19 research papers from the CORD-19 collection. These two sources together cover both classification-style and extractive-style answers, and both supply sentence-tokenised passages with identifiable evidence sentences, which is exactly what is needed to evaluate set-wise minimal sufficient evidence selection.

F.2.2 Unified schema

The build pipeline normalizes both sources into a single record schema:

- `id` – stable item identifier (e.g. `pubmedqa_<pubid>`, `covid_qa_<idx>`).
- `source` – `"pubmedqa"` or `"covid_qa"`.
- `question` – the natural-language question.
- `answer` – gold answer (`yes/no/maybe` for PubMedQA; the gold extractive span for COVID-QA).
- `answer_type` – `"yes/no/maybe"` or `"extractive"`.
- `context_sentences` – the candidate evidence pool, one sentence per element.
- `evidence_indices` – integer indices into `context_sentences` that mark the gold supporting sentences.
- `evidence_sentences` – the materialised gold sentences (provided for convenience).
- `metadata` – source-specific fields (PMID, section labels, document id, answer offset).

F.2.3 Construction details

PubMedQA branch. Each abstract in `pqa_labeled` ships with a sentence list and matching section labels (`OBJECTIVE` / `METHODS` / `RESULTS` / `CONCLUSIONS`). We treat sentences labelled `RESULTS` or `CONCLUSIONS` as gold evidence because they directly justify the yes/no/maybe decision; if no such sentence exists, the entire abstract is taken as gold to keep recall well-defined. The `long_answer` rationale and the section labels are preserved in `metadata`.

COVID-QA branch. Each item carries a paragraph and a gold extractive answer span. We (i) strip metadata lines (URLs, `DOI`, `License`, ISO-date lines), (ii) split the remaining text into sentences with a biomedical-friendly regex splitter, (iii) drop sentences shorter than 10 characters, and (iv) tag a sentence as evidence if it lexically contains the gold answer span (case-insensitive). Two fallbacks handle paraphrased answers: a $\geq 60\%$

Table 16: Aggregate statistics of MED-EVIDENCEQA

Property	Value
Total examples	2,973
PubMedQA	1,000
COVID-QA	1,973
Avg. context sentences / item	11.31
Avg. evidence sentences / item	1.77
Avg. evidence ratio (gold / context)	0.1886
PubMedQA answer dist. (yes / no / maybe)	552 / 338 / 110

word-overlap match, and as a last resort the single highest-overlap sentence. We then *window* the context to ± 5 sentences around any evidence sentence and re-index so that evidence indices remain valid after windowing. Question-level deduplication removes verbatim repeats.

F.2.4 Aggregate statistics

Two characteristics of Table 16 are worth emphasising for this paper. First, the average evidence ratio is only $\sim 19\%$: on a typical item, fewer than two sentences out of ~ 11 candidate sentences are needed to justify the gold answer. This makes MED-EVIDENCEQA a natural test-bed for *minimal sufficient* evidence selection: any system that returns the full candidate context wins recall trivially but pays in precision and tokens. Second, the PubMedQA label distribution is moderately skewed toward `yes` (552 / 1,000); PubMedQA accuracy is therefore reported as a coarse signal and COVID-QA Rouge-L is the more sensitive answer-level metric on this benchmark (cf. main paper, Section 5.1).

G Prompts and Worked Examples

For full reproducibility, this appendix lists the three principal LLM prompts used by MEGA : (i) the *subquery decomposition* prompt that rewrites a clinical question into a set of focused retrieval queries (§G.1); (ii) the *answer-generation* prompt for CRC-EVIDENCEQA / PINGCE50, the bilingual high-difficulty clinical benchmark with cited-sentence outputs (§G.2); and (iii) the *answer-generation* prompt for MED-EVIDENCEQA / OPENDATA, the English open-domain benchmark (PubMedQA + COVID-QA) with sentence-index ground truth (§G.3). All prompts are presented verbatim in English; the original Chinese templates in our codebase are semantically identical and are referenced by file name and line number above. We deliberately keep the wording short and instruction-style so that behaviour is dominated by the retrieved evidence and the IGS-selected subset rather than by prompt engineering.

G.1 Subquery Decomposition Prompt

The subquery rewriter turns one clinical question into three retrieval-friendly variants that cover *different aspects* of the original question. This expands hybrid-retrieval recall before the IGS scorer prunes the candidate pool back to a compact sufficient subset.

Prompt: Subquery Decomposition

```
System:
You are a professional medical information-retrieval expert, skilled at deciding
whether a passage provides positive or negative evidence for a given clinical
question and at extracting the relevant content.
User:
You are a professional medical information-retrieval expert.  Rewrite the
following medical question into 3 related, more specific sub-questions that
broaden retrieval coverage and help fetch the relevant clinical-guideline
passages.
Original question:  {question}
Each sub-question must be:
  1.  Specific and unambiguous, easy to retrieve;
  2.  Cover a different aspect of the original question;
  3.  Use precise medical terminology.
```

```
Example.
Q: "Is there evidence-based support for combining mitoxantrone in the second-line
treatment of BRAF V600E-mutant patients?"
Rewrite to 3 sub-questions:
  1.  What second-line treatment options are available for BRAF V600E-mutant
patients?
  2.  Under what conditions is mitoxantrone typically used?
  3.  What drugs are available for treating BRAF V600E-mutant disease?
Output the sub-question list directly, one per line, with no numbering or extra
formatting.
```

Worked Example: CRC-High

Input question. *"For a cT3N1 mid-low rectal-cancer patient with dMMR/MSI-H status, what neoadjuvant strategy is recommended, and is adjuvant chemotherapy still required after surgery?"*

Generated sub-queries.

1. Recommended neoadjuvant chemoradiotherapy regimens for cT3N1 mid-low rectal cancer.
2. Role of immune checkpoint inhibitors in the neoadjuvant setting for dMMR/MSI-H rectal cancer.
3. Indications for adjuvant chemotherapy in stage III rectal cancer after R0 resection.

Worked Example: Med-EvidenceQA, PubMedQA-style

Input question. *"Does perioperative dexmedetomidine reduce postoperative delirium in elderly patients undergoing non-cardiac surgery?"*

Generated sub-queries.

1. Randomised trials of dexmedetomidine versus placebo for postoperative delirium in elderly patients.
2. Mechanism by which dexmedetomidine modulates postoperative neuro-inflammation.
3. Dosing protocols for perioperative dexmedetomidine infusion in non-cardiac surgery.

G.2 Answer-Generation Prompt for CRC-EvidenceQA

CRC-EVIDENCEQA questions are bilingual, expert-curated, and demand verbatim citation of the original guideline. The prompt therefore (i) scopes the model to colorectal-cancer expertise, (ii) enforces evidence-grounded reasoning, and (iii) asks for inline citation indices that mirror the snippet IDs delivered by the IGS+FPTAS selector.

Prompt: CRC-EvidenceQA Answer Generation

```
System (with retrieved context):
You are an expert in colorectal-cancer diagnosis and treatment.  Answer the
question strictly based on the provided guideline content.  Keep the answer
concise and professional.
System (no retrieved context, fallback):
You are an expert in colorectal-cancer diagnosis and treatment.  Please answer
the question briefly.
User:
## Relevant guideline content
{context}
## Clinical question
{question}
Please answer strictly based on the guideline content above; quote the original
wording whenever possible.  If the guideline is insufficient to answer the
question, state this explicitly.
```

Worked Example: CRC-EvidenceQA

Question. *"A 58-year-old male is diagnosed with cT3N1, dMMR, KRAS/NRAS/BRAF wild-type mid-rectal cancer. What neoadjuvant therapy should be offered?"*

Selected Evidence Returned by IGS+FPTAS

[1] (CSCO 2025, §Rectal-Cancer Neoadjuvant Therapy) "For mid-low rectal cancer with cT3-4 or N+, long-course CRT (50.4 Gy / 28 fx with concurrent capecitabine) is the standard recommendation."

[2] (NCCN 2025, §MSI-H Subgroup) "For locally advanced dMMR/MSI-H rectal cancer, neoadjuvant immune checkpoint blockade (e.g. PD-1 monotherapy) yields high pCR rates and may obviate radiotherapy."

Generated Answer

For this cT3N1 dMMR mid-rectal cancer patient, the guideline-recommended neoadjuvant strategy is PD-1-based immune checkpoint blockade as first-line, given the documented high pathological complete response rate in dMMR/MSI-H disease[2]. Long-course chemoradiotherapy (50.4 Gy with concurrent capecitabine) remains an alternative if immunotherapy is unavailable or contraindicated[1].

G.3 Answer-Generation Prompt for Med-EvidenceQA (OPENDATA)

MED-EVIDENCEQA pools PubMedQA and COVID-QA: each item ships with a sentence-tokenised passage and a gold set of *evidence sentence indices*. Our answer prompt is intentionally short and English-native, mirroring the open-domain biomedical-QA literature, and asks the model to cite passage indices so that the answer can be programmatically aligned with gold evidence indices.

Prompt: Med-EvidenceQA Answer Generation

```
System:
You are a medical expert.  Use the provided reference passages to answer the
question.  Be structured, cite passage indices that support each claim, and do
not fabricate information beyond the references.
User:
Question:  {question}
Reference passages:
  [1] {passage_1}
  [2] {passage_2}
  ...
  [k] {passage_k}
Provide a structured answer in English, citing passage indices in square
brackets.  If the passages do not contain enough information to answer, say so
explicitly.
```

Worked Example: PubMedQA Item

Question. *"Does maternal serum AFP at 16 weeks predict preterm birth?"*

Reference Passages after IGS+FPTAS Selection (k=3 of 11 candidates)

[1] "Among 8 432 singleton pregnancies, MoM-adjusted maternal serum AFP >2.0 at 15–18 weeks was associated with a 2.4-fold increased risk of spontaneous preterm birth (p<0.001)."

[2] "After adjustment for placental abruption and pre-eclampsia, the association weakened (OR 1.6, 95 % CI 1.2–2.1) but remained statistically significant."

[3] "Authors concluded that elevated mid-trimester AFP is an independent predictor and may be useful for risk stratification in combination with cervical-length screening."

Generated Answer

Yes. Elevated mid-trimester maternal serum AFP (MoM > 2.0 at 15–18 weeks) is associated with a 2.4-fold increase in spontaneous preterm birth in singleton pregnancies[1], and the association remains significant after adjusting for placental abruption and pre-eclampsia (OR 1.6, 95 % CI 1.2–2.1)[2]. The authors recommend combining AFP with cervical-length screening for risk stratification[3].

Worked Example: COVID-QA Item

Question. *"Which cytokine elevation correlates most strongly with COVID-19 ICU admission?"*

Reference Passages after IGS+FPTAS Selection (k=2)

[1] "Serum IL-6 levels were significantly higher in ICU patients (median 76.8 pg/mL) than in non-ICU patients (median 6.8 pg/mL, $p<0.0001$)."

[2] "In multivariate logistic regression, IL-6 was the strongest independent predictor of ICU transfer (OR 4.7, 95 % CI 2.6–8.5), outperforming TNF-α, IL-10 and CRP."

Generated Answer

IL-6 shows the strongest correlation with ICU admission: median serum IL-6 is roughly an order of magnitude higher in ICU than in non-ICU COVID-19 patients (76.8 vs 6.8 pg/mL, $p<0.0001$)[1], and IL-6 is the strongest independent predictor of ICU transfer in multivariate analysis (OR 4.7, 95 % CI 2.6–8.5), outperforming TNF-α, IL-10 and CRP[2].

G.4 LLM-as-Judge Scoring Prompt

For completeness, we also list the LLM-as-judge prompt used to score CRC-EVIDENCEQA answers along the four dimensions reported in Table 1 . The judge model is run with τ=0 and is held identical across all baselines and MEGA .

Prompt: LLM-as-Judge Scoring

```
System:
You are an objective and rigorous medical-QA evaluator; please score the
candidate answer independently and objectively.
User:
Score the candidate answer against the reference answer.
## Question
{question}
## Reference answer
{reference}
## Candidate answer
{answer}
## Scoring dimensions (integer 1-10 each)
  1.  Accuracy (factual consistency with the reference)
  2.  Completeness (coverage of all aspects of the question)
  3.  Citation quality (whether claims are properly supported)
  4.  Overall
Output strictly in the following format:
Accuracy:  X/10
Completeness:  X/10
Citation:  X/10
Overall:  X/10
Comment:  <one sentence>
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper studies compact sufficient evidence selection for medical evidence-grounded QA, proposes MEGA with subquery-based retrieval, hidden-state Information Gain Scoring, and an FPTAS evidence selector, and evaluates it on curated and public medical QA benchmarks. These claims match the problem formulation, methodology, theoretical analysis, and experimental results presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix B.3 details the limitations of our current work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for the compact sufficient evidence selection objective, including nonnegative evidence scores, token costs, and the concave coverage function. The NP-hardness result is proved by a polynomial-time reduction from 0/1 KNAPSACK, with the full proof

provided in Appendix C.1. For the FPTAS-style selector, the paper explicitly states that the $(1-\epsilon)$ approximation guarantee applies to the linearized modular knapsack surrogate under standard 0/1 knapsack assumptions, while the subsequent cleanup step is used to refine the selected subset under the original set-wise objective.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5.5 describes the benchmark construction, dataset splits, LLM backbones, retrieval and reranking models, frozen LLM used for Information Gain Scoring, evaluation metrics, and baseline methods. It also provides the prompts for subquery decomposition, answer generation, and LLM-as-Judge evaluation in the appendix, together with the MEGA inference algorithm and the FPTAS-style selector pseudocode. These details are sufficient to reproduce the main experimental comparisons and support the paper's main claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code, part of benchmark, prompts, and evaluation scripts in the Appendex G and supplementary materials. The paper also describes the dataset construction process, model backbones, retrieval and reranking settings, evaluation metrics, and algorithmic pseudocode in Appendix D.1. These materials are intended to allow faithful reproduction of the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the dataset construction and splits, the evaluated LLM backbones, retrieval and reranking models, frozen LLM used for Information Gain Scoring, evaluation metrics, and baseline settings. Since MEGA is a training-free framework, there is no gradient-based training optimizer. The key inference hyperparameters, including retrieval size, candidate pool size, metric weights, and FPTAS approximation factor, are described in the methodology and appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide the statistical significance of the experiments in Table 1 and Figure 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided our compute resources details in Appendix D.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper is designed for evidence-grounded medical QA and does not involve human-subject intervention, clinical deployment, or patient-identifiable data. The constructed benchmarks use curated guideline/literature snippets and public QA datasets, and the system is evaluated as a decision-support research prototype rather than as a replacement for clinicians. We also emphasize citation-grounded generation and evidence inspection to reduce unsupported medical claims.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the potential positive impact of improving medical QA reliability through compact, inspectable, and evidence-grounded evidence selection. It also acknowledges potential risks, including over-reliance on generated answers, incomplete or biased evidence retrieval, and misuse in settings without clinician oversight.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper does not release a new high-risk generative model. The proposed framework uses existing LLMs as frozen backbones and focuses on evidence selection for citation-grounded QA. For data release, we plan to release only reproducible benchmark components and derived annotations that comply with the licenses and terms of the underlying sources. Any medical guideline or literature content subject to redistribution restrictions will be released as metadata, references, or processing scripts rather than raw copyrighted text.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources for the public datasets, baseline methods, retrieval models, reranking models, and LLM backbones used in the experiments. We respect the terms of use of the underlying medical literature, guidelines, and model APIs. For any constructed benchmark data, release will follow the licenses and redistribution constraints of the original sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The paper proposed two evidence-grounded medical QA benchmarks, CRC-EvidenceQA and Med-EvidenceQA. We document their construction process, data sources, task formats, evidence annotations, dataset statistics, and evaluation protocols in the main paper and appendix F. We also provide prompts and worked examples to clarify how the assets are used in evaluation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not conduct crowdsourcing experiments or human-subject experiments. No participants were recruited, and no compensation was provided. The clinician-collaborative benchmark construction involved expert review of medical QA examples and evidence snippets, rather than a human-subject behavioral study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject intervention, patient recruitment, or collection of patient-identifiable data. Therefore, there are no study-participant risks or IRB requirements associated with human-subject research. The work uses public datasets, curated guideline/literature snippets, and de-identified or synthetic benchmark examples for evaluating evidence-grounded QA.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage is central to the proposed framework and is described in the paper. MEGA uses LLMs for subquery decomposition, hidden-state Information Gain Scoring with a frozen LLM, and citation-grounded answer generation. The paper specifies the evaluated answer-generation backbones, the frozen LLM used for evidence scoring, and the prompts used for subquery generation, answer generation, and LLM-as-Judge evaluation in the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.