Decisions by Discrimination: Differential Evidence for Medical Reasoning

Yunpeng Xiao[1], Yuzhang Xie[1], Junhao Huang[1], Carl Yang[1], Xiao Hu[2], Kai Shu[1]
[1]Department of Computer Science, Emory University,
[2]Nell Hodgson Woodruff School of Nursing, Emory University
{yunpeng.xiao, yuzhang.xie, junhao.huang,
j.carlyang, xiao.hu, kai.shu}@emory.edu

Abstract

Large language models (LLMs) are increasingly studied for clinical decision support, and medical question answering provides a testbed for evaluating this capability. In particular, medical multiple-choice questions (MCQs) reflect a core challenge in clinical decision-making: plausible options often share overlapping evidence, while only a few decisive factors distinguish the correct option from incorrect ones. However, existing retrieval-based and reasoning-oriented methods mainly focus on supplying broadly relevant knowledge or eliciting reasoning paths, rather than identifying the compact differential evidence that separates confusing alternatives. In this work, we present a retrieval-grounded differential evidence framework for medical MCQs. Our method extracts option-specific pieces of differential evidence and organizes them into concise evidence prompts for answer selection. This design better aligns external knowledge with differential reasoning while reducing irrelevant context. Experiments show consistent accuracy gains across multiple datasets and target LLMs. Ablation studies further show that our framework outperforms both directly using retrieved evidence chunks and closed-source LLM-generated general hints, and that its improvements are not solely explained by answer-proximal guidance, but are consistent with differential evidence serving as a compact aid for reasoning over competing alternatives. Our findings also offer insights for medical reasoning and other decision-making settings with multiple possible choices.

1 Introduction

Large language models (LLMs) are increasingly explored as clinical assistants, powering systems for triage, diagnosis support, and patient education (Singhal et al., 2023, 2025; Wu et al., 2024). However, even medically tuned LLMs remain prone to unsafe reasoning behaviors. A common challenge in clinical reasoning is that many candidate diagnoses or treatment decisions share a large core of overlapping facts, while differing only in a small number of critical discriminatory factors (Bowen, 2006; Cook and Décary, 2020). If a model fails to identify and reason over these few distinguishing facts, then even access to abundant relevant knowledge may still lead to misdiagnosis or other incorrect clinical decisions.

In medical multiple-choice questions (medical MCQs), this challenge is especially well exposed, making them a natural benchmark for studying differential clinical reasoning. In real clinical practice, competing diagnoses or treatment options often share substantial overlapping evidence, with only a few decisive factors determining the correct choice. Medical MCQs capture this structure naturally: several answer options may each align with part of the vignette, but only one is supported by the discriminative evidence that resolves the case. This makes the benchmark useful for evaluating whether models can identify clinically meaningful differences among plausible hypotheses, rather than relying on broadly relevant but non-discriminative facts.

However, existing methods are still not well aligned with this requirement. Retrieval-augmented generation (RAG) has become a common strategy for medical QA, supplementing LLMs with external knowledge from structured resources such as knowledge graphs and unstructured corpora such as textbooks. Current medical RAG pipelines are primarily designed to retrieve generally relevant evidence, not the small set of discriminative facts that separates plausible answer options. As a result, they often return a broad supportive context while failing to surface the evidence needed to resolve near-miss alternatives. Figure 1 presents a concrete example in which two clinically similar diseases share multiple symptoms, while only a small number of distinctive symptoms determine the final diagnosis.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5589–5603
November 6-10, 2026 ©2026 Association for Computational Linguistics

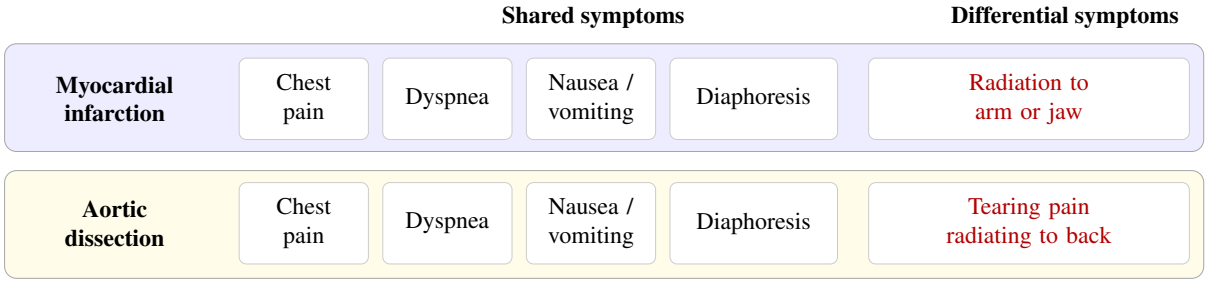


Figure 1: Two similar diseases may share most symptoms, while only a few discriminative symptoms determine the final diagnosis. Red text highlights the differential evidence.

We therefore propose a retrieval-grounded differential evidence framework for medical MCQs. Instead of appending broad supporting context, our method extracts anchor cues, retrieves cue–option co-occurrence passages from textbooks, and distills them into compact option-specific evidence for answer selection. This design reduces retrieved context length by about 85% while improving accuracy across benchmarks.

Our main contributions are:

- We formulate differential evidence as an evidence organization problem, emphasizing option-discriminative evidence rather than broadly relevant knowledge.
- We propose a four-stage framework that extracts anchor cues, retrieves cue–option co-occurrence evidence, and distills it into compact option-specific evidence for answer selection.
- Experiments on three medical MCQ benchmarks show consistent gains across multiple LLMs. Ablation studies with raw retrieved text, general hints, and evidence-only settings suggest that the gains come from differential evidence rather than more text or general answer cues.

2 Related Work

2.1 Medical QA and Retrieval-Augmented Generation (RAG)

Large-scale medical QA benchmarks, such as MedQA (Jin et al., 2021), MedXpertQA (Zuo et al., 2025), and MedBullets (Chen et al., 2025), have become important testbeds for evaluating fine-grained clinical reasoning in LLMs. In particular, multiple-choice medical QA provides a constrained decision-making setting in which several candidate options may all appear clinically plausible at first glance. This makes such benchmarks especially useful for assessing whether models can assist in real-world clinical decision making (Chen et al., 2025; Wei et al., 2026; Wang et al., 2025b).

Many medical QA tasks, especially MedQA-style problems, require reasoning beyond the information explicitly stated in the question stem, thus models often need to draw on background medical knowledge, including disease mechanisms, diagnostic criteria, treatment indications, and contraindications (Liévin et al., 2024; Lucas et al., 2024; Bhasuran et al., 2025). Since this knowledge is broad, evolving, and difficult to capture reliably through parametric memory alone, retrieval-augmented generation (RAG) over external medical knowledge bases has emerged as a natural solution (Yang et al., 2025b). By retrieving evidence from curated resources such as clinical guidelines, biomedical ontologies, textbooks, and knowledge graphs, medical RAG systems can ground answer selection in explicit evidence rather than latent memorization (Amugongo et al., 2025; Ke et al., 2025; Song et al., 2025). Recent work has explored this direction through methods such as MIRAGE (Xiong et al., 2024a), K-Comp (Cho and Lee, 2025), MKRAG (Shi et al., 2025), MedRAG (Zhao et al., 2025), Brief-Context (Zhang et al., 2025a), RAG^2(Sohn et al., 2025), KGARevion (Su et al., 2024), and Self-BioRAG (Jeong et al., 2024). In addition to traditional text-based RAG, recent studies have explored KG-guided retrieval and reasoning frameworks, such as Dr.Knows (Gao et al., 2025), MedIKAL (Jia et al., 2025), KERAP (Xie et al., 2025), KG4Diagnosis (Zuo et al., 2024), and KGARevion (Su et al., 2024), which leverage knowledge graphs to facilitate answer generation, diagnostic reasoning, and evidence retrieval.

Although retrieved external evidence can im-

prove factual grounding, conventional retrieval generally prioritizes overall semantic relevance to the question, often returning broad background knowledge that may support multiple candidate options simultaneously (Amugongo et al., 2025; Liu et al., 2025c; Xiong et al., 2024b). However, for medical multiple-choice reasoning, the key challenge is often not general relevance but discriminative utility: the retrieved evidence should help the model determine why one option is more appropriate than other clinically plausible alternatives under the specific case context (Chen et al., 2023; Singh and Shrivastava, 2025; Alhazmi et al., 2024). These observations make differential evidence especially well-suited for medical QA.

2.2 Option Discrimination and Evidence Selection in Multiple-Choice Reasoning

Existing work has increasingly suggested that the main challenge in multiple-choice reasoning is not merely identifying generally relevant knowledge, but distinguishing among candidate options that may all appear partially plausible. Early work on multi-choice reading comprehension emphasized joint interactions among the passage, question, and candidate answers, rather than question-only matching (Wang et al., 2018). Recent studies further show that LLMs are sensitive to candidate-option structure: some methods model elimination over options (Ma and Du, 2023), while others find that predictions can vary under answer-order permutations, indicating unstable discrimination among alternatives (Pezeshkpour and Hruschka, 2024). Related analyses also show that benchmarks often contain several plausible candidates, and that models may select the least incorrect option rather than one that is distinctly correct (Wang et al., 2025a; Palta et al., 2024). This issue is especially salient in medical MCQs, where options often share overlapping clinical facts and differ only in a few decisive mechanisms, contraindications, or management roles. In parallel, retrieval-based methods have incorporated answer options into retrieval or knowledge selection, such as question-aware knowledge graph prompting and answer-conditioned support retrieval (Zhang et al., 2025b; Liu et al., 2025b). However, these methods mainly improve option-aware knowledge access, while compact differential evidence for option-level comparison remains underexplored. Unlike prior option-aware retrieval methods that primarily improve access to option-related knowledge, our focus is on organizing knowledge into differential evidence that clearly distinguishing between the various options.

3 Methods

In this section, we will introduce the proposed differential evidence reasoning method for clinical decision-making, especially in multiple-choice questions. Figure 2 shows the framework of our proposed method, which consists of four stages: anchor cue extraction, cue-option co-occurrence retrieval, differential evidence generation, and evidence-based inference. The first three stages form the differential evidence construction pipeline, while the fourth stage uses the generated evidence for final answer prediction.

3.1 Definition of Differential Evidence

Given a multiple-choice question with candidate options

$$O = \{o_1, o_2, \ldots, o_m\},$$

we define differential evidence for an option o_i as external evidence that supports o_i more strongly than competing options under the current case context. In other words, differential evidence is not merely relevant knowledge, but knowledge with discriminative utility for separating clinically plausible alternatives.

Formally, if $s(e, o)$ denotes the degree to which an evidence item e supports option o, then ideal differential evidence for o_i would satisfy

$$s(e, o_i) > s(e, o_k), \quad \forall k \neq i,$$

or more generally provide substantially stronger support for o_i than for the strongest competing option.

In practice, explicit support scoring is difficult because raw textbook passages are often long, redundant, and only indirectly aligned with option-level distinctions. We therefore approximate differential evidence through a retrieval-and-distillation pipeline. The following subsections describe the first three stages of the framework for differential evidence construction, while the fourth stage performs evidence-based inference.

3.2 Definition of Atomic Facts and Anchor Cues

In medical multiple-choice questions, one data sample typically consists of the parts: the background context C, the explicit question Q and the options

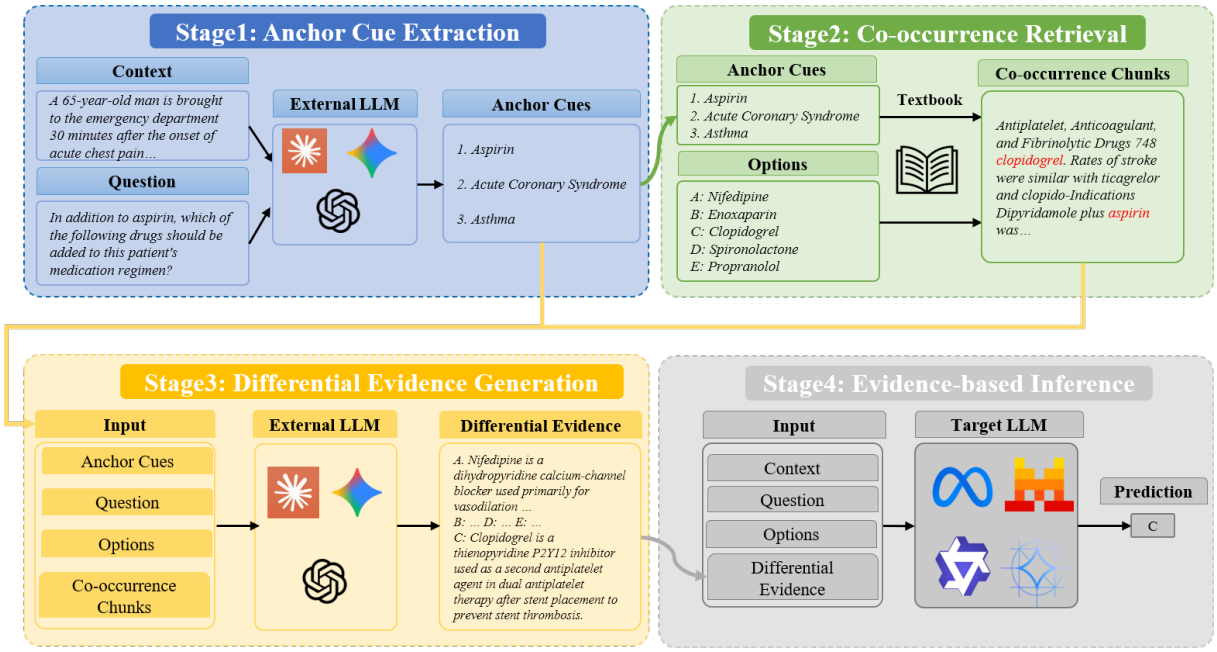


Figure 2: The framework of our proposed method.

O. Since the key clinical evidence is usually conveyed in the background description, we decompose C into a set of atomic facts, following the intuition of fine-grained factual decomposition in fact-checking (Min et al., 2023):

$$C = \{f_1, f_2, \ldots, f_n\},$$

where each f_i denotes a minimal clinical fact, such as a symptom, laboratory finding, medical history element, or treatment detail.

However, atomic decomposition alone is insufficient, as the fact set may still contain substantial redundancy and many facts that are only weakly relevant to the question Q and final decision. We therefore define anchor cue, which is to distill the original fact set into a compact set of highly salient cues. Formally, we select a subset of atomic facts,

$$A = \{a_1, a_2, \ldots, a_i\} \subseteq C, \quad i \ll n,$$

where A denotes the set of anchor cues.

This step can be understood as a general content distillation problem in text understanding: from a larger collection of fine-grained facts, we aim to retain only those that best preserve the core semantics and decision-critical information of the scenario. This step is not intended for full clinical reasoning or direct answer prediction, but to distill the scenario into a few salient anchor cues, similar to focused reading comprehension over clinical text. For example, the context mentions that *the patient developed tinnitus* while the question asks *Which medication caused this phenomenon*, then "tinnitus" serves as the anchor cue for this specific question.

3.3 Textbook Knowledge Retrieval through Anchor Cue–Option Co-occurrence

Similar to traditional RAG, we also need to leverage external knowledge to enhance the performance of the target model in medical scenarios. However, as noted in Section 2, current knowledge graphs are not yet sufficient to distinguish the subtle nuances between options and generate discriminative evidence; therefore, we identify chunks in textbooks that exhibit co-occurrence between anchor cues and the options as our preliminary source of external knowledge.

Specifically, by leveraging the powerful semantic understanding capabilities of external LLMs, we first input the context C and the question Q into these models to extract anchor cues

$$A = \{a_1, a_2, \ldots, a_i\},$$

and normalize these cues and generate their common variants (e.g., singular/plural forms, abbreviations, etc.). Similarly, we input the options O into the external LLMs, subjecting them to the same normalization process and generating their common variants

$$O_n = \{o_1, o_2, \ldots, o_j\}.$$

This strict input separation reduces the risk that external LLMs act as implicit solvers of the task. Instead, their role is limited to localized semantic normalization and variant generation, rather than answer inference over the full MCQ instance.

Next, we use the textbook provided in the MedQA dataset (Jin et al., 2021) to identify co-occurring chunks

$$p_{ij} = \text{Co-occurr}(a_i, o_j).$$

A chunk is considered co-occurring if it contains at least one normalized variant of an anchor cue and one normalized variant of an option concept. In this work, to control prompt length and reduce redundant evidence, we extract only one chunk for each anchor cue-option pair. Specifically, we extract a maximum of $|A| \times |O|$ paragraphs. More details are shown in appendix.

3.4 Textbook-Constrained Differential Evidence Generation

Compared with knowledge graphs, textbook knowledge provides substantially richer semantic content. However, it is often fragmented, redundant, and not explicitly organized around option-level distinctions, which makes it difficult to directly obtain discriminative evidence for specific answer choices. In addition, textbook retrieval may still miss relevant information due to limited coverage or retrieval failures. To address this, we use external LLMs to generate option-discriminative evidence while treating retrieved textbook knowledge as a constraint on the generation process. Formally, we generate the set of differential evidence $D = \{d_A, d_B, ...\}$ where d_A denotes the evidence specific to Option A, as follows:

$$D = ExternalLLM(A, Q, O, P).$$

Here, A is the set of anchor cues, O is the set of original options, $P = \bigcup p_{ij}$ is the set of retrieved external knowledge, and Q is the question. Each d_o is generated as a short, option-specific, condition-oriented statement capturing the conditions that most clearly distinguish option from competing alternatives. When retrieved evidence is missing for a given option, the LLM is allowed to provide limited supplementation from internal knowledge.

3.5 Differential Evidence Quality Control

Because external LLMs may have been pretrained on overlapping medical QA data, we introduce several safeguards to reduce answer leakage and bias. First, during anchor cue extraction and normalization, the external LLM only sees the context and question C, Q without the options O; during option extraction and normalization, it only sees the options O without the context and question C, Q; during differential evidence generation, it only sees the anchor cues, questions, options A, Q, O without context C. As a result, the external model never has access to the full MCQ instance at any stage. Second, the system prompt explicitly instructs the model to prioritize retrieved textbook evidence and only supplement missing details when retrieval is insufficient. Finally, for differential evidence generation with external LLMs, we use carefully designed prompts to encourage the production of option-discriminative evidence while preserving as much neutrality as possible.

3.6 Differential Evidence-Based Inference

Upon obtaining the differential evidence set, the target LLMs—typically distinct from the external LLM used in the preceding step—infer the answers to the medical MCQs using the following method:

$$Answer = TargetLLM(C, Q, O, D).$$

Here, C is the clinical context, Q is the question, O is the set of original options, and D is the generated differential evidence. The target LLM takes these components as input and produces the final answer.

4 Experiments

In this section, we will introduce the dataset used in the experiments, the LLMs employed to generate anchor cues and evidence (External LLMs), and the LLMs used for testing (Target LLMs). We then answer the following research questions:

- RQ1: Can differential evidence improve model performance?
- RQ2: Does differential evidence provide benefits beyond generic auxiliary evidence?
- RQ3: Does differential evidence genuinely support reasoning, or does it risk improving performance by providing answer-proximal guidance?

Model	MedQA (5-op)				MedBullets (5-op)				MedXpert-Text			
	Base	CoT	Diff.	CoT +Diff.	Base	CoT	Diff.	CoT +Diff.	Base	CoT	Diff.	CoT +Diff.
Llama-3-8B	52.40	65.02	65.12	**67.01**	44.16	49.35	50.00	**60.06**	11.31	14.73	14.53	**18.20**
Llama-3-70B	74.31	79.26	79.81	**81.70**	61.36	67.86	68.83	**68.83**	21.63	23.68	23.84	**23.88**
Gemma-3-27B-It	64.02	71.25	71.33	**75.81**	46.75	50.00	51.95	**53.57**	14.61	15.63	16.98	**19.27**
Qwen-3-32B	69.91	73.68	75.33	**78.95**	55.19	60.06	58.44	**64.61**	17.14	18.82	19.55	**21.47**
Mistral-3-8B	56.24	68.97	64.96	**74.16**	42.86	50.97	50.97	**59.09**	15.39	16.86	17.76	**19.35**
Average	63.38	71.64	71.31	**75.53**	50.06	55.65	56.04	**61.23**	16.02	17.94	18.53	**20.43**

Table 1: Performance comparison across models, datasets, and prompting settings. Base denotes direct answering without differential evidence, CoT denotes zero-shot chain-of-thought prompting, Diff. denotes augmentation with differential evidence, and CoT+Diff. combines chain-of-thought prompting with differential evidence. All values are reported in accuracy (%).

4.1 Datasets and Models

Datasets. We evaluate our method on MedQA (Jin et al., 2021), MedXpertQA-Text (Zuo et al., 2025), and MedBullets (Chen et al., 2025) because together they provide a strong testbed for scenario-grounded, option-discriminative medical reasoning. By contrast, some medical QA benchmarks, such as PubMedQA (Jin et al., 2019) and MedMCQA (Pal et al., 2022), generally feature shorter question stems and are more oriented toward knowledge-centric medical question answering. Therefore, they are less suitable for evaluating the scenario-grounded differential reasoning targeted in this work.

External LLMs. Motivated by RAG^2 (Sohn et al., 2025), we view the external LLM as a differential evidence generator. Its quality mainly affects the generated evidence, which then influences downstream reasoning. We tested several closed-source LLMs, including Claude-family models, and selected GPT-5 mini (Singh et al., 2026) as the evidence generator because it offered the best overall trade-off among evidence quality, instruction following, cost, and API stability. To improve reproducibility and assess dependence on closed-source models, we also tested open-source models. Among them, DeepSeek-V3.2 (Liu et al., 2025a) produced the most reliable differential evidence in our experiments. We therefore report an additional ablation using DeepSeek-V3.2 as the evidence generator.

Target LLMs. We evaluate our method on a series of open-source LLMs. These include Llama 3 (Grattafiori et al., 2024), Mistral (Mistral AI, 2025), Gemma (Gemma Team et al., 2025) and Qwen 3 (Yang et al., 2025a). All experiments are conducted through Amazon Bedrock under a unified inference setup, using deterministic decoding. More details are shown in appendix.

4.2 Performance of Differential Evidence (RQ1)

We use four settings for the experiments: **Baseline**, **CoT**, **Differential Evidence** and **CoT + Differential Evidence** . **Baseline** directly predicts the answer from the original context, question, and options. **CoT** uses the same input with an additional zero-shot chain-of-thought instruction. **Differential Evidence** further augments the input with our generated option-specific differential evidence. **CoT + Differential Evidence** setting simultaneously utilize CoT and differential evidence.

As shown in Table 1, the **CoT + Differential Evidence** setting achieves better performance than both **Baseline** and **CoT** across all five models. Notably, using **Differential Evidence** alone already achieves performance close to **CoT** in many settings, indicating that the generated evidence itself provides useful external knowledge and option-discriminative information.

We further examine whether the improvement is tied to a single proprietary generator by replacing GPT-5 mini with DeepSeek-V3.2 as the evidence generator. Table 2 shows that DeepSeek-generated evidence improves all five target models, increasing the average MedQA accuracy from 63.38% to 69.09%. While GPT-5 mini remains stronger on average, the consistent gains from DeepSeek-V3.2 indicate that the benefits of differential evidence are not exclusive to one generator, though generator quality still affects the magnitude of improvement.

Model	Base	+Diff. DeepSeek	+Diff. GPT-5 mini
Llama-3-8B	52.40	61.19	65.12
Llama-3-70B	74.31	79.96	79.81
Gemma-3-27B-It	64.02	68.97	71.33
Qwen-3-32B	69.91	73.29	75.33
Mistral-3-8B	56.24	62.06	64.96
Average	63.38	69.09	71.31

Table 2: Evidence generator ablation on MedQA (5-option). We compare differential evidence generated by GPT-5 mini and DeepSeek-V3.2 under the non-CoT setting.

4.3 Benefits Beyond Generic Retrieved Evidence and Generic Hints (RQ2)

To examine whether the gains of differential evidence go beyond simply adding more external text, we compare two CoT-based settings: **CoT + Text**, which directly appends the retrieved co-occurrence paragraphs to the input, and **CoT + Differential Evidence**, which instead provides the model with compact option-specific differential evidence distilled from those retrieved materials.

As shown in Table 3, directly using the retrieved co-occurrence text does not reliably improve performance. In many cases, **CoT + Text** even performs worse than the CoT baseline, especially on MedQA and MedBullets. By contrast, **CoT + Differential Evidence** consistently yields positive gains over CoT across all three datasets. This contrast suggests that the benefit does not come from merely exposing the model to more retrieved content, but from organizing external knowledge into a concise and discriminative form that is more aligned with multiple-choice decision making.

We further compare the output length of the two methods using word count. As shown in Table 4, differential evidence is substantially more compact than directly concatenated retrieved text across all three datasets. For example, on MedQA the average length decreases from 677.3 words for co-occurrence text to 108.5 words for differential evidence, with similar reductions on MedBullets and MedXpertQA-Text. This suggests that the gain of our framework comes not from adding more retrieved content, but from organizing external knowledge into a shorter and more discriminative form. Moreover, the substantial number of non-empty co-occurrence retrievals suggests that our method is supported by sufficient retrieved paragraphs, rather than depending purely on the external LLM to fabricate evidence.

These findings indicate that generic retrieved text can be too long, noisy, or weakly structured to effectively support reasoning, and may even distract the model from the decisive clues in the question. This observation is also consistent with prior studies reporting that directly concatenating retrieved medical text often provides limited benefit and can sometimes degrade performance due to irrelevant or diffuse context (Kim et al., 2025). In comparison, differential evidence appears to make better use of external knowledge by highlighting the most distinguishing facts for separating competing options, rather than requiring the model to infer these distinctions from raw retrieved paragraphs on its own.

To test whether the gains come merely from adding LLM-generated auxiliary medical text, we construct a general-hint baseline using the same GPT-5 mini generator. This baseline is conditioned on the same anchor cues, question, and answer options, but it does not use retrieved textbook co-occurrence chunks and is not instructed to identify option-discriminative conditions. Instead, it asks the generator to produce general option-level medical background notes.

As shown in Table 5, general hints provide only a modest average improvement over the base setting, increasing MedQA accuracy from 63.38% to 64.71%. In contrast, differential evidence improves the average accuracy to 71.31%. This gap suggests that the benefit of our method is not simply due to appending additional LLM-generated text, but comes from organized differential evidence.

4.4 Reasoning Support or Answer-Proximal Guidance? (RQ3)

We further investigate whether differential evidence improves performance by genuinely supporting reasoning, or by introducing answer-proximal guidance (e.g., the evidence includes a sentence like "The answer is ..."). To this end, we evaluate four input settings on three datasets: only options (Only O), only evidence (Only E), context + question + options (C+Q+O), and context + question + options + evidence (C+Q+O+E).

Among these settings, C+Q+O serves as the standard baseline, while C+Q+O+E tests whether differential evidence provides additional value beyond the original question input. The Only O and Only E settings are included to analyze how much predictive signal is contained in the options

Model	MedQA (5-op)				MedBullets (5-op)				MedXpert-Text			
	CoT +Text	Δ	CoT +Diff.	Δ	CoT +Text	Δ	CoT +Diff.	Δ	CoT +Text	Δ	CoT +Diff.	Δ
Llama-3-8B	58.84	-6.18	**67.01**	**+1.99**	42.53	-6.82	**60.06**	**+10.71**	14.78	+0.05	**18.20**	**+3.47**
Llama-3-70B	75.73	-3.53	**81.70**	**+2.44**	56.17	-11.69	**68.83**	**+0.97**	22.69	-0.99	**23.88**	**+0.20**
Gemma-3-27B-It	65.51	-5.74	**75.81**	**+4.56**	49.03	-0.97	**53.57**	**+3.57**	15.14	-0.49	**19.27**	**+3.64**
Qwen-3-32B	73.29	-0.39	**78.95**	**+5.27**	53.25	-6.81	**64.61**	**+4.55**	19.92	+1.10	**21.47**	**+2.65**
Mistral-3-8B	58.37	-10.60	**74.16**	**+5.19**	45.45	-5.52	**59.09**	**+8.12**	18.04	+1.18	**19.35**	**+2.49**
Average	66.35	-5.29	**75.53**	**+3.89**	49.29	-6.36	**61.23**	**+5.58**	18.11	+0.17	**20.43**	**+2.49**

Table 3: Comparison between generic retrieved text and differential evidence under the same chain-of-thought (CoT) reasoning framework. Δ over CoT denotes the absolute accuracy gain relative to the CoT baseline in Table 1.

Dataset	Diff. Mean	Diff. Median	Co-occ. Mean	Co-occ. Median	Non-empty
MedQA	108.5	108.0	677.3	571.0	1023/1273
MedBullets	112.2	112.5	734.6	621.5	264/308
MedXpert	207.1	207.0	1160.5	957.0	1829/2450

Table 4: Length comparison in words between co-occurrence text and differential evidence. Non-empty counts questions with at least one retrieved co-occurrence passage.

Model	Base	+General Hint	+Diff. Evidence
Llama-3-8B	52.40	53.57	65.12
Llama-3-70B	74.31	77.22	79.81
Gemma-3-27B-It	64.02	65.28	71.33
Qwen-3-32B	69.91	69.99	75.33
Mistral-3-8B	56.24	57.50	64.96
Average	63.38	64.71	71.31

Table 5: Comparison with GPT-5-mini-generated general hints on MedQA (5-option).

and the generated evidence, respectively. Specifically, Only O measures how much answer signal is already exposed by the candidate options themselves, while Only E tests whether the generated evidence alone is overly predictive of the correct answer. For clarity, we visualize the MedQA results in the main paper and report the full results on all three datasets in Appendix.

As shown in Figure 3, the evidence-only setting achieves clearly above-chance performance, indicating that differential evidence carries useful answer-relevant information. However, its performance remains substantially below the full-input setting. On average, Only E reaches 41.05% accuracy, whereas $C{+}Q{+}O{+}E$ reaches 71.31%, a gap of more than 30 percentage points. This suggests that differential evidence alone is not sufficient to solve the questions reliably in isolation.

These results suggest that the gains from differential evidence are not solely explained by direct answer cues. While the evidence itself is informative, it is not sufficient to fully solve the questions in isolation.

We emphasize that differential evidence is expected to be answer-relevant; otherwise it would not be useful evidence. The goal of this analysis is therefore not to show that evidence contains no predictive signal, but to test whether it collapses into an answer shortcut.

5 Discussion and Conclusion

We present a differential evidence framework for medical multiple-choice question answering. Experiments across three medical QA benchmarks and multiple target models show that differential evidence consistently improves direct answering and further complements chain-of-thought prompting. Comparisons with general hints and retrieved chunks suggest that the gains are not merely due to adding extra generated text, but depend on the quality and organization of the evidence. Leakage-oriented analyses further indicate that differential evidence is informative but does not function as a standalone answer channel. Although our experiments focus on medical MCQs, the underlying idea may apply more broadly to decision-making settings where candidate hypotheses, plans, or actions are explicitly provided or can be enumerated, and where a small number of discriminative facts determine the final choice.

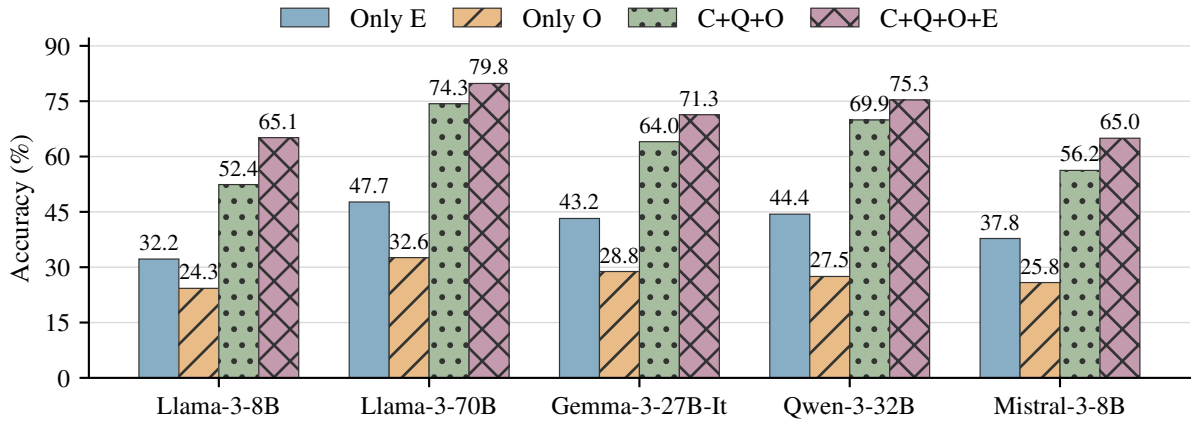


Figure 3: Comparison of four input settings on MedQA for RQ3: Only options (Only O), Only evidence (Only E), context+question+options (C+Q+O), and context+question+options+evidence (C+Q+O+E). Differential evidence alone provides useful signal, but yields the strongest gains when combined with the original question context and candidate options.

Limitations

Our study still has some limitations. First, our framework depends on the quality of the external evidence generator. Although GPT-5 mini and DeepSeek-V3.2 produce useful differential evidence in our experiments, weaker generators may produce overly generic, overly short, or answer-proximal statements, reducing downstream gains. In addition, while our prompts and leakage analyses aim to prevent the evidence from directly revealing the answer, subtle answer-proximal cues may still occur, particularly in management and contraindication questions. Finally, our experiments are limited to medical multiple-choice benchmarks with explicit candidate options, and extending the framework to open-ended or real-world decision settings will require additional validation.

Ethical Considerations

This work evaluates evidence-based answering on medical multiple-choice benchmarks. Benchmark accuracy does not establish clinical safety or effectiveness. Generated evidence, including supplementation from a model's internal knowledge, may be incomplete, inaccurate, or biased. These outputs should not be used as patient-specific medical advice or as a substitute for qualified clinical judgment; clinical use would require independent expert review and validation.

We used GPT-5 for limited drafting and language editing assistance during manuscript preparation. All technical content, experimental design, analysis, and final wording were reviewed and verified by the authors, who take full responsibility for the final paper.

References

Elaf Alhazmi, Quan Z Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 14437–14458.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. 2025. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine*, 8(1):166.

Judith L Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference*

of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3563–3599.

Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si, and Yin Zhang. 2023. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13207–13224.

Jeonghun Cho and Gary Lee. 2025. K-COMP: Retrieval-augmented medical domain question answering with knowledge-injected compressor. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6878–6901.

Chad E Cook and Simon Décary. 2020. Higher order thinking about differential diagnosis. *Brazilian journal of physical therapy*, 24(1):1–7.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In *31st International Conference on Computational Linguistics*, pages 9278–9298.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, et al. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.

Hyunjae Kim, Jiwoong Sohn, Aidan Gilson, Nicholas Cochran-Caggiano, Serina Applebaum, Heeju Jin, Seihee Park, Yujin Park, Jiyeong Park, Seoyoung Choi, et al. 2025. Rethinking retrieval-augmented generation for medicine: A large-scale, systematic expert evaluation and practical insights. *arXiv preprint arXiv:2511.06738*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Haochen Liu, Song Wang, Chen Chen, and Jundong Li. 2025b. Question-aware knowledge graph prompting for enhancing large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1388–1400.

Siru Liu, Allison B McCoy, and Adam Wright. 2025c. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Chenkai Ma and Xinya Du. 2023. Poe: Process of elimination for multiple choice reasoning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 4487–4496.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Mistral AI. 2025. Introducing Mistral 3.

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

Shramay Palta, Nishant Balepur, Peter A Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly problematic questions in multiple-choice benchmarks for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3451–3473.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Xiang Li, and Ninghao Liu. 2025. MKRAG: Medical knowledge retrieval augmented generation for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2024, pages 1011–1020.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2026. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.

Manish Singh and Manish Shrivastava. 2025. Options-aware dense retrieval for multiple-choice query answering. *arXiv preprint arXiv:2501.16111*.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753.

Jie Song, Zhichuan Xu, Mengqiao He, Jinhua Feng, and Bairong Shen. 2025. Graph retrieval augmented large language models for facial phenotype associated rare genetic disease. *NPJ digital medicine*, 8(1):543.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. KGARevion: an ai agent for knowledge-intensive biomedical qa. *arXiv preprint arXiv:2410.04660*.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025a. LLMs may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862.

Ruiyu Wang, Yuzhang Xie, Xiao Hu, Carl Yang, and Jiaying Lu. 2025b. BioMedJImpact: A comprehensive dataset and LLM pipeline for ai engagement and scientific impact analysis of biomedical journals. *arXiv preprint arXiv:2511.12821*.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 746–751.

Mingyang Wei, Dehai Min, Zewen Liu, Yuzhang Xie, Guanchen Wu, Carl Yang, Max SY Lau, Qi He, Lu Cheng, and Wei Jin. 2026. EpiQAL: Benchmarking large language models in epidemiological question answering for enhanced alignment and reasoning. *arXiv preprint arXiv:2601.03471*.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, et al. 2024. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646.

Yuzhang Xie, Hejie Cui, Ziyang Zhang, Jiaying Lu, Kai Shu, Fadi Nahab, Xiao Hu, and Carl Yang. 2025. KERAP: A knowledge-enhanced reasoning approach for accurate zero-shot diagnosis prediction using multi-agent llms. *arXiv preprint arXiv:2507.02773*.

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

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025b. Retrieval-augmented generation for generative artificial intelligence in health care. *Npj health systems*, 2(1):2.

Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, et al. 2025a. Leveraging long context in retrieval augmented language models for

medical question answering. *npj Digital Medicine*, 8(1):239.

Zeyu Zhang, Alessandro Moschitti, and Thuy Vu. 2025b. Retrieving support to rank answers in open-domain question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35086–35093.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. 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.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2024. KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. *arXiv preprint arXiv:2412.16833*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Additional Detailed Results for RQ3

We report the full RQ3 results on all three datasets in Appendix Table 6. We observe that on MedXpert-Text, the evidence-only setting is relatively strong and for some models is comparable to, or slightly better than, $C{+}Q{+}O$. A possible reason is that the overall baseline performance on this dataset is low, suggesting that the task is difficult for the tested models under the standard input. In such cases, even a small amount of compact external signal may lead to a noticeable gain. However, $C{+}Q{+}O{+}E$ still yields the strongest overall results, indicating that differential evidence is most effective when combined with the original context.

B Co-occurrence Retrieval Details

For external knowledge retrieval, we use the textbook corpus provided together with the MedQA dataset (Jin et al., 2021). Besides the question data, MedQA also releases reference textbooks that have already been segmented into chunks, and we directly perform co-occurrence retrieval over these pre-defined chunks. Since some chunks are overly long, we additionally cap the length of each retrieved chunk at 400 words. When multiple chunks satisfy the co-occurrence criterion for the same anchor cue-option pair, we keep only the first retrieved chunk.

We also explored several retrieval variants, including BM25-based retrieval strategies for identifying anchor-option co-occurrence paragraphs. In pilot experiments, these variants produced similar downstream trends. Since retrieval in our framework mainly serves as a grounding constraint for evidence distillation rather than as the final answer-selection mechanism, we use a simple and reproducible retrieval strategy in the main experiments.

C Implementation Details

We use a unified inference setup for all target models. For direct answer generation, we set *max new tokens* =16; for CoT-based settings, we set *max new tokens* =512 to allow sufficient space for intermediate reasoning. For decoding, we use *temperature* =0.0 and *top p*=1.0 across all experiments.

D Prompts

D.1 Anchor Cue Extraction and Normalization

```
SYSTEM
You are an extractor for USMLE/MedQA-style
question stems.

USER
You are an extractor for USMLE/MedQA-style
question stems.

Input:
- "Question": the full question stem

Task:
Identify:
1) `question`: the actual question being
asked, usually the final sentence
2) `entities`: up to 3 medical entities
extracted only from the context/background
part

Rules:
1) Use only the input stem.
2) Do not infer from answer choices.
3) Do not hallucinate image findings.
4) Do not output the context itself.
5) Extract entities only from the context,
not from the final question sentence.
6) Return at most 3 entities, choosing the
most answer-relevant ones.
7) Prefer concise, concrete, medically
meaningful entities.
8) Each entity must contain `text`,
`normalized`, and `variants`.
9) `variants` should contain up to 2 common
alternate forms only, generally includes
singular, plural, and abbreviation.
10) Deduplicate entities.

Example:
Input Question:
"A 25-year-old woman has fever, dysuria, and
urinary frequency for 2 days. Which of the
following is the most likely diagnosis?"

Output:
{
```

Model	MedQA (5-op)				MedBullets (5-op)				MedXpert-Text			
	Only O	Only E	Base	+Diff.	Only O	Only E	Base	+Diff.	Only O	Only E	Base	+Diff.
Llama-3-8B	24.27	32.21	52.40	**65.12**	23.05	35.71	44.16	**50.00**	13.14	14.61	11.31	**14.53**
Llama-3-70B	32.60	47.68	74.31	**79.81**	30.52	39.61	61.36	**68.83**	16.04	17.92	21.63	**23.84**
Gemma-3-27B-It	28.83	43.21	64.02	**71.33**	29.55	37.66	46.75	**51.95**	14.08	18.37	14.61	**16.98**
Qwen-3-32B	27.49	44.38	69.91	**75.33**	27.60	41.23	55.19	**58.44**	13.22	15.06	17.14	**19.55**
Mistral-3-8B	25.84	37.78	56.24	**64.96**	26.62	36.04	42.86	**50.97**	12.73	16.08	15.39	**17.76**
Average	27.81	41.05	63.38	**71.31**	27.47	38.05	50.06	**56.04**	13.84	16.41	16.02	**18.53**

Table 6: Full RQ3 results on all three datasets. Base denotes C+Q+O, and +Diff. denotes C+Q+O+E. We report accuracy (%) under four input settings: only options (Only O), only evidence (Only E), context+question+options (C+Q+O), and context+question+options+evidence (C+Q+O+E). Here, C+Q+O and C+Q+O+E correspond to the **Base** and **Diff.** settings in the main results table, respectively.

```
  "question": "Which of the following is the
  most likely diagnosis?",
  "entities": [
    {"text": "fever", "normalized": "fever",
    "variants": []},
    {"text": "dysuria", "normalized":
    "dysuria", "variants": []},
    {"text": "urinary frequency",
    "normalized": "urinary frequency",
    "variants": []}
  ]
}

Return structured JSON matching the schema
exactly.

[paste context and question]
```

D.2 Option Normalization

```
SYSTEM
You are an extractor for USMLE/MedQA-style
question stems.

USER
You will receive a JSON object with:
- "options": a dictionary of answer options,
such as {"A": "...", "B": "...", ...}

Your task is to normalize the medical
entities appearing in EACH option
separately.

GOAL:
For each option, extract the core medical
entity or entities explicitly present in
that option text,
and provide:
- text
- normalized
- variants

IMPORTANT RULES:
1) Use ONLY the option text provided.
2) Do NOT use the question stem.
3) Do NOT infer hidden meaning beyond what
is explicitly written in the option.
4) For a simple option such as a drug,
disease, or procedure name, usually extract
just one entity.
5) Prefer concise, text-grounded entity
spans.
6) Deduplicate entities within each option.
7) `normalized` should be a concise
canonical medical form (lower-case
preferred).
8) `variants` should contain up to 2 common
alternate names/forms only, such as:
   - abbreviation
   - singular/plural form
   - common short form
9) Do NOT include long explanations,
speculative synonyms, or unrelated concepts.
10) If an option contains no meaningful
medical entity, return an empty entity list
for that option.

Return output in exactly this shape:
{
  "options": [
    {
      "option_key": "A",
      "entities": [
        {
          "text": "...",
          "normalized": "...",
          "variants": ["...", "..."]
        }
      ]
    }
  ]
}

Do NOT return a dictionary keyed by option
labels.
Return a list under `options`.

Return structured JSON matching the provided
schema exactly.

[paste options]
```

D.3 Differential Evidence Generation

```
SYSTEM
You are a medical differential-evidence
generator for multiple-choice question
distillation.

USER
Input includes:
1. the original question,
2. anchor facts extracted from the clinical
background,
```

```
3. the answer options,
4. co-occurrence paragraphs between anchor
facts and options.
Task: For each option, generate one short
differential-evidence string that identifies
the option's most distinctive clinical
decision role in the setting of the
question.
Definition of differential evidence:
Differential evidence is a short,
option-specific clinical cue that places the
option on the most relevant decision axis
for this question, such as treatment role,
timing, mechanism, contraindication,
adverse-effect relevance, management
category, or typical use setting. It should
help distinguish the option conceptually
from competing options, but it must not read
like an answer explanation or a statement
about whether the option fits the case.
Guidelines:
- Use the original question only to infer
the general decision setting and the likely
clinical management axis.
- Do not restate or reconstruct the full
case.
- Use co-occurrence paragraphs as grounding
constraints when available.
- If no co-occurrence paragraph is
available, use internal medical knowledge
conservatively and keep it consistent with
the question setting.
- Prefer the option's most distinctive
decision role over generic drug summaries,
broad mechanism descriptions, or
nonessential side effects.
- For medications, prioritize therapeutic
role, management timing, treatment category,
or clinically important risk over general
pharmacology unless pharmacology is itself
the key distinguishing cue.
- Keep the evidence concise and focused.
- Write one sentence whenever possible.
Restrictions:
- Do not solve the question.
- Do not say or imply which option is
correct, incorrect, preferred, excluded, or
should be selected.
- Do not rank options.
- Do not copy the source text verbatim.
- Do not output reasoning steps, bullet
points, or extra commentary.
- Do not explain why the option matches or
fails to match the current case.
- Do not directly justify the statement
using the patient, vignette, described
findings, or anchor facts.
- Do not distinguish the option by saying
what it is not used for, not part of, or
less appropriate than.
- Avoid negative contrast constructions such
as: "rather than ...", "instead of ...",
"not part of ...", "not used for ...", "less
appropriate", "ruled out".
- Express the evidence as a positive
description of the option's own clinical
decision role.
- Avoid generic textbook descriptions that
are not useful for distinguishing the option
on the question's decision axis.
Output format:
Return only a JSON object mapping each
option label to its differential-evidence
string.
Each string should be a standalone concise
sentence.
[paste question, anchor cues, options and
co-occurrence paragraphs]
```

D.4 Answering with Differential Evidence

```
SYSTEM
You are taking a medical multiple-choice
exam.

USER
Use the question, options, and differential
evidence to choose the single best answer.

Differential evidence refers to the most
distinguishing evidence for each option,
that is, the evidence most useful for
separating that option from the competing
options.

Use the differential evidence as a set of
discriminative clues, and then look back to
the question context to identify the facts
most consistent with one option over the
others.

Do not treat the evidence as a final
conclusion by itself; use it to guide
context-grounded reasoning.

Reply with only one capital letter: A, B, C,
D, E, F, G, H, I, J.

Question: [paste context and question]

Options: [paste options]

Retrieved Evidence: [paste differential
evidences]

Answer:
```

D.5 CoT Answering with Retrieved Co-occurrence Text

```
SYSTEM
You are taking a medical multiple-choice
exam.

USER
Use the question, options, and retrieved
evidence to reason briefly and choose the
single best answer.

The retrieved evidence below consists of raw
co-occurrence text associated with the
question and candidate options. Use it as
supporting context, but do not assume it is
conclusive by itself.

Compare the options briefly, ground your
reasoning in the question context, and
choose the single best answer.
Keep the reasoning short (2-4 sentences). Do
not list option letters in the reasoning.
```

```
Output format:

Reasoning: <2-4 short sentences>

Final Answer: one capital letter among A, B,
C, D, E, F, G, H, I, J.

Question: [paste context and question]

Options: [paste options]

Retrieved Evidence: [paste differential
evidences]

Reasoning:
```