

ToToFilter: Topic- and Topology-Aware Noise Filtering Framework over Medical KGs for Evidence-Based Medical Q&A

Yufeng Wang^{1,2,3}, Weijie Chen¹, Chao Chen⁴, Xingfeng Liu¹,
Yizhe Yang¹, Zhihao Chen⁷, Qiming He², Yubo Chen⁵,
Kang Liu⁶, Xuefeng Wang^{2,3}, Xiangwen Liao^{1,2,*}, Jingfeng Liu^{2,3,*}

¹College of Computer and Data Science, Fuzhou University

²Interdisciplinary Institute for Medical Engineering, Fuzhou University

³NHC Key Laboratory of Cancer and Metabolism, Clinical Oncology School of
Fujian Medical University, Fujian Cancer Hospital

⁴School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)

⁵School of Information Engineering, Minzu University of China

⁶The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences

⁷Department of Electronic Information Science, Fujian Jiangxia University

*Correspondence: liaoxw@fzu.edu.cn; drjingfeng@fjmu.edu.cn

Abstract

Evidence-Based Medical Question Answering (EB-MQA) aims to provide accurate and verifiable responses to support clinical decisions. To retrieve evidence from general medical knowledge graphs (KGs), existing knowledge augmentation methods use semantic similarity or dynamic sampling for entity matching and agentic expansion or subgraph scoring for path retrieval. These approaches can introduce entity-level noise from irrelevant or generic entities and path-level noise from redundant or spurious reasoning chains. We propose ToToFilter, a train-free topic- and topology-aware noise filtering framework with Topic-Aware Entity Localization for query-relevant entity anchoring and Topology-Aware Path Filtering for path selection based on semantic consistency, structural compactness, and information specificity. Across four MQA benchmarks, ToToFilter achieves average accuracy gains of 2.30 and 1.22 percentage points over DALK on Llama3.1-8B and Qwen2.5-7B, with maximum gains of 2.60 and 2.78 points. On PrimeKGQA, it improves answer recall and triple recall by up to 8.81 and 19.70 percentage points. Further analyses verify the effectiveness of the proposed two-level filtering framework.

1 Introduction

Medical Question Answering (MQA) aims to answer clinical queries accurately (Singhal et al., 2023; Wang and Zhang, 2024). Evidence-based MQA (EB-MQA) follows Evidence-Based Medicine (Sackett et al., 1996) by providing supporting evidence for its answers (Gupta et al.,

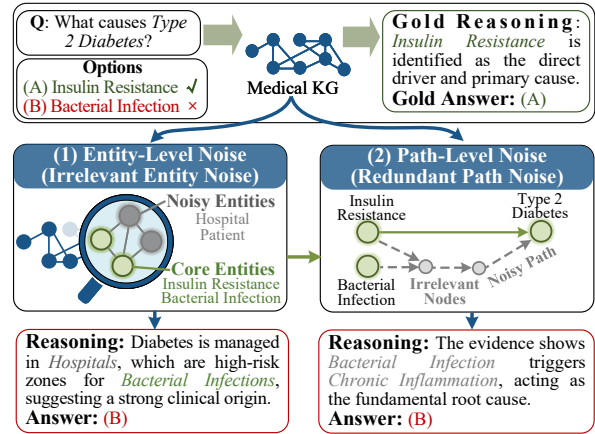


Figure 1: Noise challenges in general medical KG retrieval. (1) Entity-level noise: retrieval may favor high-frequency generic nodes over core entities, yielding irrelevant context. (2) Path-level noise: retrieval can introduce redundant paths that obscure direct causal links and trigger spurious pathogenic inferences.

2025), which is important for clinical decision support (Tun et al., 2025; Newton et al., 2024).

Current EB-MQA methods leverage large language models (LLMs) (Singhal et al., 2025; Hiro-sawa et al., 2024) with retrieval-augmented generation (RAG) to curb hallucinations (Xiong et al., 2024; Zhu et al., 2025; Pal et al., 2023). Beyond un-structured corpora, recent work retrieves evidence from large-scale general medical knowledge graphs (KGs), which provide broad-coverage structured relations (Chandak et al., 2023; Zhao et al., 2025). To denoise KG retrieval, existing methods mainly operate at two levels: entity localization and path reasoning. (1) At the **entity**-level, approaches use

generative matching and dynamic sampling (Jiang et al., 2025; Li et al., 2024), or structure-aware re-ranking/subgraph construction (Yang et al., 2024b; Matsumoto et al., 2024). (2) At the **path-level**, approaches construct reasoning chains through agentic expansion or path scoring (Wen et al., 2024; Xu et al., 2024), and improve path quality through iterative verification (Ning et al., 2025).

However, existing methods still frequently fail to localize topic-relevant core entities or introduce logically redundant paths (Jiang et al., 2023), which can overshadow valid evidence and impair reasoning reliability. This gives rise to two main forms of retrieval noise. As illustrated in Figure 1, **(1) Entity-Level Noise** arises when, for a question about the etiology of “Type 2 Diabetes”, retrieval drifts toward high-frequency generic nodes such as “Hospital” and “Patient” instead of the diagnostically central entity “Insulin Resistance”; therefore, the first challenging problem is how to *precisely anchor query-aligned core entities in massive KGs*. **(2) Path-Level Noise** arises when, even after the correct anchor is identified, a direct path to “Type 2 Diabetes” is overshadowed by a longer and less informative route through “Bacterial Infection”; thus, the second challenging problem is how to *filter candidate paths while balancing efficiency and precision to retain a high-quality subset of knowledge paths*.

In response, we propose a **Topic- and Topology-aware Noise Filtering** framework (ToToFilter) to reduce retrieval noise and improve reliability. It consists of two components: (1) At the entity-level, the proposed Topic-Aware Entity Localization (TEL) ranks and filters candidate mentions by their diagnostic relevance to the question, and then aligns the retained core concepts to KG nodes, thereby accurately locating relevant evidence nodes within the global graph. (2) At the path-level, Topology-Aware Path Filtering (TPF) recalls multi-hop paths among the localized KG nodes, scores them by semantic relevance to the question, hop-based compactness, and intermediate-node specificity, and then uses an LLM reranker to select the most useful evidence chains for answer prediction.

Across four MQA benchmarks, ToToFilter improves average accuracy over DALK by 2.30 and 1.22 percentage points on Llama3.1-8B and Qwen2.5-7B, respectively. On PrimeKGQA, it improves answer recall and triple recall by up to 8.81 and 19.70 points, respectively. Further analyses verify the effectiveness of TEL and TPF in reduc-

ing entity- and path-level retrieval noise. Overall, our contributions are as follows:

- (1) We characterize two recurrent sources of noise in medical KG retrieval: entity-level noise and path-level noise.
- (2) We propose ToToFilter, a train-free two-level filtering framework that addresses both sources through TEL and TPF.
- (3) Experiments on five benchmark datasets show that ToToFilter improves accuracy and retrieval quality across backbones, while analyses verify its ability to suppress noisy nodes and redundant paths.

2 Related Work

2.1 Knowledge Augmentation Methods in EB-MQA

Current EB-MQA methods benefit from the rapid advancement of LLMs (Singhal et al., 2025; Hiro-sawa et al., 2024), yet their inherent hallucinations constrain clinical applications (Zhu et al., 2025; Pal et al., 2023). To mitigate this, RAG is introduced to provide domain knowledge (Xiong et al., 2024). Some studies retrieve evidence from unstructured corpora (Sohn et al., 2025; Xiong et al., 2025b), but may introduce lengthy contexts and cause key information dispersion (Liu et al., 2024).

To structure knowledge more effectively, some studies build specialized KGs as curated tuples (Matsumoto et al., 2024; Yang et al., 2024b; Lin et al., 2024), but annotation is costly and transfer across medical subtopics is limited (Zhang et al., 2025; Abu-Salih et al., 2023). Recent work therefore leverages general-purpose medical KGs (Chandak et al., 2023; Zhao et al., 2025; Du et al., 2025), whose broad coverage comes with substantial noise, motivating stronger denoising methods.

2.2 Entity- and Path-Level Denoising in KG

Entity-level denoising filters topic-relevant entities via biomedical entity linking (Lin et al., 2024) or query-driven subgraph construction (Zhang et al., 2022). Path-level denoising addresses redundant or invalid reasoning chains via flow-based pruning (Chen et al., 2026), spurious-path mitigation (Hu et al., 2022), self-reflection loops (Jeong et al., 2024), process supervision (Xiong et al., 2025a), and multi-source evidence fusion (Liang et al., 2025). However, existing methods struggle to balance topological constraints with semantic scoring

efficiency, and high-quality path filtering on general medical KGs often incurs substantial computational overhead. In response, ToToFilter mitigates entity- and path-level noise via two-level filtering.

3 Preliminaries

3.1 Task Formalization

In this work, the goal is to answer a medical question correctly with high-quality supporting evidence retrieved from medical KGs. Formally, given a clinical query Q and a set of n candidate answer options $O = \{o_1, o_2, \dots, o_n\}$, a framework $\mathcal{F}$ predicts an answer $\hat{o} \in O$ together with a retrieved reasoning path set $\hat{\mathcal{T}}$:

$$(\hat{o}, \hat{\mathcal{T}}) = \mathcal{F}(Q, O), \quad (1)$$

where $\hat{o}$ is expected to match the ground-truth answer $o^* \in O$, and paths in $\hat{\mathcal{T}}$ serve as supporting evidence for answer generation. We instantiate $\mathcal{F}$ as backbone LLMs augmented with medical KGs.

3.2 Revisiting Failure Modes in Existing KG-Augmented Pipelines

We first probe the retrieval-side failure modes in existing KG-augmented EB-MQA systems built upon general-purpose medical KGs through a targeted diagnostic study. Specifically, we find that these methods are limited by: (1) *entity-level noise*, where topic-irrelevant mentions bias retrieval toward high-frequency generic nodes; and (2) *path-level noise*, where many multi-hop connections yield indirect or low-specificity chains that dilute discriminative evidence and mislead reasoning.

We adopt a controlled “patch-in” protocol for three KG-augmented baselines, *BasicKG*, *MindMap* (Wen et al., 2024), and *DALK* (Li et al., 2024), on MedQA and MMLU. For each baseline, we add two lightweight denoising operations and measure accuracy: (1) **Baseline + TEL**: we retain highly topic-relevant entities and anchor them to KG nodes. (2) **Baseline + TEL + TPF**: we further select paths that are more direct and specific to suppress redundant or hub-mediated chains. All other settings are unchanged, so performance changes can be attributed to *what evidence is retrieved*.

Figure 2 exhibits a consistent pattern across baselines and datasets. First, retaining highly relevant entities (+TEL) reliably improves accuracy (orange bars). Intuitively, when topic-irrelevant entities are admitted into the anchor set, retrieval tends to over-explore generic hubs and broadly

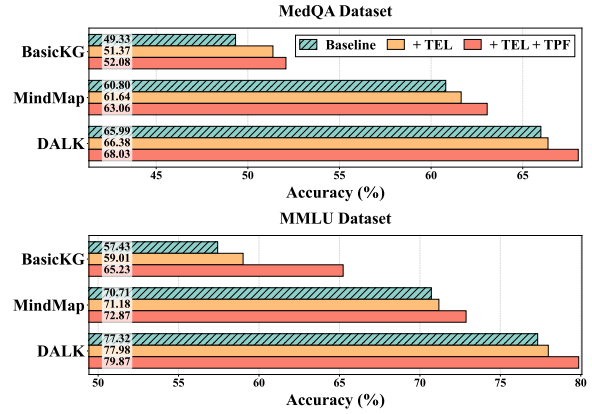


Figure 2: **Controlled patch-in diagnostic study on MedQA and MMLU.** We probe entity-level and path-level retrieval noise by progressively patching in TEL and TPF into three KG-augmented baselines (BasicKG, MindMap, DALK).

connected concepts, producing evidence that is plausible but weakly diagnostic for selecting the correct option. Second, choosing paths with direct associations (+TPF) produces additional improvements (red bars), validating our claim that *redundant or spurious paths persist even after better anchoring*. General-purpose medical KGs frequently admit many alternative multi-hop routes between the same core entities; a non-trivial portion of these routes are indirect, mediated by broad clinical concepts rather than clinically actionable evidence. Such paths easily mislead LLM-based reasoning: long or hub-heavy chains can appear coherent in natural language, while diluting option-discriminative evidence. Therefore, path selection must go beyond semantic plausibility and account for *question-specific topological cues* that correlate with specificity and evidential strength.

4 ToToFilter Framework

4.1 Overview

The preliminary results in Section 3.2 suggest that the limitations of existing KG-augmented methods could be due to *retrieval-side noise arising at both the entity and path levels*. This motivates our design of ToToFilter as a two-level framework that addresses entity-level noise and path-level noise. As shown in Figure 3, ToToFilter consists of two modules: (1) **Topic-Aware Entity Localization**, which extracts query-relevant core entities and anchors them to KG nodes; and (2) **Topology-Aware Path Filtering**, which retrieves and selects informative reasoning paths among the anchored entities.

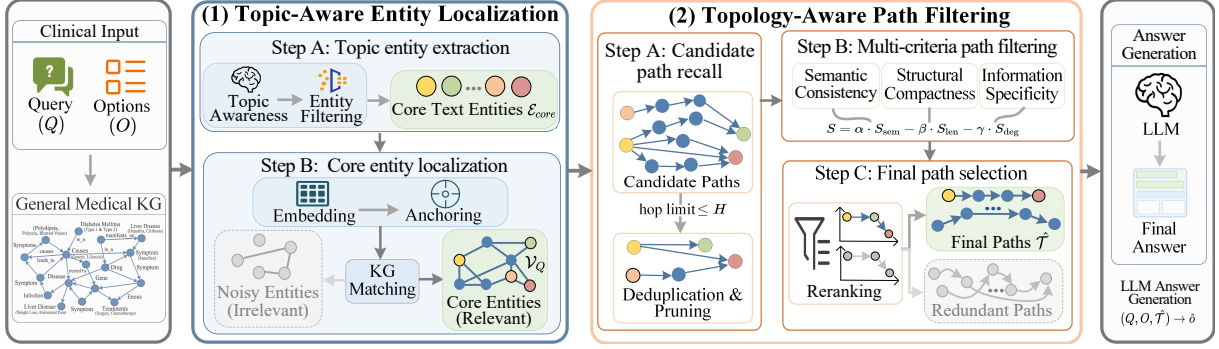


Figure 3: **Overview of the ToToFilter Framework.** This framework employs a two-level filtering strategy with two components: (1) Topic-aware Entity Localization (TEL), which performs entity-level extraction, filtering, and anchoring from the clinical query; and (2) Topology-aware Path Filtering (TPF), which performs path-level scoring and selection to retain informative reasoning paths for answer generation.

4.2 TEL: Topic-Aware Entity Localization

Before entity localization, query screening classifies questions as BIOMEDICAL or OTHER. Questions classified as OTHER are answered by the same backbone without graph paths, while BIOMEDICAL questions proceed through TEL and TPF.

For biomedical questions, KG retrieval starts by mapping entities from the unstructured query to KG nodes as anchors for retrieval. Since inaccurate or overly generic anchors may introduce entity-level noise by steering retrieval toward high-degree hubs, as discussed in Section 3.2, TEL identifies diagnostically important entities and aligns them with KG nodes in two steps.

Step A: Topic entity extraction. A language model first extracts candidate biomedical entities from the question and answer options, after which TEL filters and ranks them according to their clinical roles in the reasoning process, including core anchors in the question stem, discriminative findings, and answer-choice target candidates. The selected mentions are parsed, deduplicated, and truncated to at most five entities, yielding the core entity set $\mathcal{E}_{\text{core}} = \{e_i\}_{i=1}^{|\mathcal{E}_{\text{core}}|}$. The corresponding extraction and selection prompts are provided in Tables 19 and 20.

Step B: Core entity localization. Medical concepts often appear in diverse surface forms (Fu et al., 2023; Fan et al., 2025), whereas KG nodes are typically normalized and may lack aliases (Kim et al., 2025). To link free-text entities to normalized KG nodes for path retrieval in Section 4.3, we adopt a vector-similarity-based soft anchoring strategy. Given a KG $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, each core entity

$e_i \in \mathcal{E}_{\text{core}}$ is mapped to its nearest KG node:

$$\hat{v}_i = \arg \max_{v \in \mathcal{V}} \cos(\text{Enc}(e_i), \text{Enc}(v)), \quad (2)$$

where $\cos(\cdot)$ denotes cosine similarity. We retain only high-confidence matches to form the anchored node set:

$$\mathcal{V}_Q = \left\{ \hat{v}_i \mid e_i \in \mathcal{E}_{\text{core}}, \cos(\text{Enc}(e_i), \text{Enc}(\hat{v}_i)) > \tau \right\}. \quad (3)$$

where τ is a confidence threshold.

4.3 TPF: Topology-Aware Path Filtering

Even with precise entity anchoring, retrieval over dense medical KGs may still return paths that are semantically related to the query but topologically non-diagnostic, such as overly long chains or paths mediated by generic hubs (Figure 1(2)). Our preliminary study further shows that filtering paths over the core entity subgraph improves evidence quality (Figure 2). Therefore, this module aims to identify a compact set of query-relevant and structurally informative reasoning paths, ultimately producing the final path set $\hat{\mathcal{T}}$ as high-quality evidence for answer generation.

Step A: Candidate path recall. Given the anchored node set $\mathcal{V}_Q$ and a maximum hop limit H , we first recall candidate shortest paths between anchored node pairs:

$$\mathcal{T}_{\text{cand}} = \bigcup_{\{u,v\} \subseteq \mathcal{V}_Q} \text{ASP}_{\leq H}(u, v; \mathcal{G}), \quad (4)$$

where $\text{ASP}_{\leq H}(u, v; \mathcal{G})$ denotes all shortest paths from u to v in $\mathcal{G}$ with at most H hops. To control computational cost, $\mathcal{T}_{\text{cand}}$ is deduplicated and truncated to at most N_{max} paths.

Step B: Multi-criteria path filtering. Each candidate path $p \in \mathcal{T}_{\text{cand}}$ with ℓ hops, is represented as a node sequence $(v_0, v_1, \dots, v_\ell)$. A high-quality path p is expected to have three complementary properties: semantic relevance, structural compactness, and information specificity.

(1) *Semantic relevance.* A useful reasoning path should remain aligned with the question; otherwise, even a valid KG path may introduce question-irrelevant evidence (Yu et al., 2025, 2024). We measure semantic alignment between a path and the query by

$$S_{\text{sem}} = \cos(\text{Enc}(\text{Txt}(p)), \text{Enc}(Q)), \quad (5)$$

where $\text{Txt}(\cdot)$ converts the path to natural language.

(2) *Structural compactness.* A reasoning path should also remain compact, since each additional hop introduces extra entities and relations, making the path more likely to accumulate noise and become less direct (Jiang et al., 2024). We therefore use hop count as the compactness score, i.e., $S_{\text{len}} = \ell$, which is negatively weighted in the final score to penalize overly long paths.

(3) *Information specificity.* Informative paths should avoid generic hub nodes. In dense KGs, high-degree nodes often connect to many concepts and therefore carry weak discriminative value for the current question (Sun et al., 2024). Accordingly, we penalize paths that pass through high-degree intermediate nodes:

$$S_{\text{deg}} = \max_{v \in \{v_1, \dots, v_{\ell-1}\} \setminus \mathcal{V}_Q} \log(1 + \deg(v)), \quad (6)$$

where $\deg(v)$ is the degree of node v . Anchored query nodes in $\mathcal{V}_Q$ are excluded from this penalty because they serve as core anchors. The maximum operator characterizes the strongest hub effect along a path, reflecting that a single highly connected intermediate node can make the whole path less specific to the current question. If no penalizable intermediate node exists, we set $S_{\text{deg}} = 0$.

Candidate paths are then sorted by the coarse ranking score S :

$$S = \alpha \cdot S_{\text{sem}} - \beta \cdot S_{\text{len}} - \gamma \cdot S_{\text{deg}}, \quad (7)$$

where α , β , and γ are hyperparameter weights. Only the top N_c paths are retained as the coarse candidate set $\mathcal{T}_{\text{cr}}$.

Step C: Final path selection. Although the coarse filtering removes many irrelevant paths, it

Algorithm 1: ToToFilter Inference

Input: Clinical query Q ; candidate options O ; knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$

Output: Predicted answer $\hat{o}$

```

1  $\hat{\mathcal{T}} \leftarrow \emptyset$ ;
2 if  $Q$  is classified as BIOMEDICAL then
3   Extract and filter  $\mathcal{E}_{\text{core}}$  by diagnostic
   relevance (Sec. 4.2);
4   Anchor entities to obtain  $\mathcal{V}_Q$  by
   Eq. (2)–(3);
5   if  $|\mathcal{V}_Q| \geq 2$  then
6     Construct  $\mathcal{T}_{\text{cand}}$  by Eq. (4);
7     if  $|\mathcal{T}_{\text{cand}}| > 0$  then
8       Coarse-filter paths by
       Eq. (5)–(7) and keep top- $N_c$  as
        $\mathcal{T}_{\text{cr}}$ ;
9       Obtain  $\hat{\mathcal{T}}$  by Eq. (8);
10      if  $|\hat{\mathcal{T}}| = 0$  then
11         $\hat{\mathcal{T}} \leftarrow \text{top-}N_f$  from  $\mathcal{T}_{\text{cr}}$ ;
12 Generate  $\hat{o}$  from  $Q, O, \hat{\mathcal{T}}$  by Eq. (9);
13 return  $\hat{o}$ ;
```

may still retain weakly informative ones. We therefore apply a final LLM-based reranking step:

$$\hat{\mathcal{T}} = \text{LLM}(\mathcal{T}_{\text{cr}}, Q, N_f; \mathcal{P}_{\text{rerank}}), \quad (8)$$

where the LLM selects up to N_f question-relevant and option-discriminative paths from $\mathcal{T}_{\text{cr}}$ under the reranking instruction $\mathcal{P}_{\text{rerank}}$. If no valid reranked result is returned, we fall back to the top- N_f paths from the coarse filtering.

4.4 Answer Generation

Given the input question Q , candidate options O , and the evidence $\hat{\mathcal{T}}$, the framework predicts the final answer by

$$\hat{o} = \text{LLM}(Q, O, \hat{\mathcal{T}}; \mathcal{P}_{\text{gen}}), \quad (9)$$

where $\mathcal{P}_{\text{gen}}$ is the answer generation instruction.

4.5 Inference and Complexity Analysis

Algorithm 1 summarizes the inference pipeline of ToToFilter. Let $K_H(u, v)$ denote the number of shortest paths retrieved for an anchored node pair $\{u, v\}$ within H hops before truncation. Its overall time complexity is $\mathcal{O}(|\mathcal{E}_{\text{core}}| \log |\mathcal{V}| + |\mathcal{V}_Q|^2(|\mathcal{V}| + |\mathcal{R}|) + \sum_{\{u, v\} \subseteq \mathcal{V}_Q} H K_H(u, v) + |\mathcal{T}_{\text{cand}}| \log |\mathcal{T}_{\text{cand}}|)$ (see Appendix A for details).

In practice, the retrieval cost depends mainly on the number of anchored node pairs and retrieved paths. TEL limits the number of anchored nodes, while TPF filters retrieved paths before LLM reranking and answer generation. The empirical efficiency is further evaluated by the Pareto analysis in Section 5.7.

5 Experiments and Analysis

5.1 Experimental Setup

To evaluate ToToFilter on MQA tasks, we assess **accuracy** on four benchmarks: MedQA (Jin et al., 2021), MMLU (Hendrycks et al., 2020), PubMedQA (Jin et al., 2019), and MedExQA (Kim et al., 2024). For retrieval quality, we report **Answer Recall@k** and **Triple Recall@k** at $k \in \{5, 8, 10\}$ on PrimeKGQA (Yan et al., 2024).

ToToFilter is compared with representative baselines spanning direct KG retrieval (BasicKG), graph-augmented reasoning and prompting with GoG (Xu et al., 2024) and MindMap (Wen et al., 2024), subgraph or hypothesis-guided evidence retrieval with SubgraphRAG (Li et al., 2025) and HyKGE (Jiang et al., 2025), and dynamic KG retrieval with DALK (Li et al., 2024). Experiments are conducted with two open-source LLMs, Qwen2.5-7B (Yang et al., 2024a) and Llama3.1-8B (Grattafiori et al., 2024). Prompt templates and hyperparameters are fixed before evaluation and shared across both backbones and all four datasets. Query screening is applied in all main experiments, and accuracy is computed over the full test set, including samples without retrieved KG paths. For paired comparisons, we report exact McNemar tests and paired-bootstrap 95% confidence intervals on sample-aligned predictions.

The appendix provides more details for datasets (B.1), baselines (B.2), metrics (B.3), implementation settings (B.4), extra results (C), and prompt templates (D).

5.2 Main Result

Table 1 reports the full-test results. Relative to DALK, ToToFilter achieves average accuracy gains of 2.30 percentage points on Llama3.1-8B and 1.22 points on Qwen2.5-7B; the largest single-dataset gains are 2.60 and 2.78 points, respectively. Per-dataset paired tests identify a significant improvement on Qwen2.5-7B/MedQA, with a 2.75-point gain, a 95% confidence interval of $[0.39, 5.11]$, and $p = 0.026$. The remaining seven comparisons have

Table 1: Accuracy (%) on four MQA benchmarks. Avg. denotes average accuracy. The **best** and **second-best** results are highlighted.

Method	MedQA	MMLU	Pub MedQA	Med ExQA	Avg
Llama3.1-8B					
Vanilla LLM	60.21	70.81	74.00	70.85	68.97
BasicKG	49.33	57.43	74.40	70.11	62.82
GoG	55.93	64.70	68.40	55.53	61.14
MindMap	60.80	70.71	74.60	70.11	69.06
SubgraphRAG	59.54	73.55	73.40	72.23	69.68
HyKGE	59.35	77.52	74.40	72.28	70.89
DALK	65.99	77.32	76.20	74.79	73.58
ToToFilter	68.03	79.87	78.80	76.81	75.88
Qwen2.5-7B					
Vanilla LLM	51.14	65.33	68.80	73.94	64.80
BasicKG	54.99	74.92	71.20	73.40	68.63
MindMap	51.14	67.86	71.60	74.89	66.37
GoG	57.74	74.29	67.60	72.66	68.07
SubgraphRAG	59.39	74.39	72.23	73.40	69.85
HyKGE	56.95	77.66	69.20	76.06	69.97
DALK	60.30	78.71	73.80	75.64	72.11
ToToFilter	63.08	79.35	74.60	76.28	73.33

confidence intervals including zero. Appendix C.1 reports the complete results.

Several direct KG retrieval baselines fall below the vanilla LLM on individual datasets, showing that unfiltered graph evidence can interfere with answer prediction. ToToFilter obtains the highest point estimate in all eight backbone–dataset settings by filtering entity and path noise before answer generation.

5.3 Evidence Retrieval Quality Analysis

To assess retrieval quality after noise filtering, we evaluate competitive baselines on PrimeKGQA using two recall-based metrics: answer recall, which measures whether retrieved evidence covers the gold answer entity, and triple recall, which measures overlap between retrieved triples and the gold reasoning evidence. As shown in Table 2, ToToFilter consistently performs best on both metrics across both backbones. On Llama3.1-8B, it surpasses the second-best method by 8.81 points in answer recall and 19.70 points in triple recall; on Qwen2.5-7B, the gains are 5.92 and 18.65 points, respectively. This shows that ToToFilter improves both answer-oriented evidence localization and path-level noise suppression.

Regarding baseline performance, representative methods reveal a clear trade-off between answer-oriented retrieval and path filtering. DALK performs relatively better on triple recall but lags on answer recall, while HyKGE and SubgraphRAG achieve stronger answer recall yet remain lim-

Table 2: Retrieval quality comparison on PrimeKGQA dataset (all results in %). The **best** and **second-best** results in each backbone block are highlighted in bold and shaded respectively.

Backbone	Method	Answer Recall $\uparrow$				Triple Recall $\uparrow$			
		@5	@8	@10	Avg	@5	@8	@10	Avg
Llama3.1-8B	GoG	23.33	27.42	29.08	26.61	4.38	6.21	6.93	5.84
	MindMap	26.08	31.25	33.58	30.30	1.04	1.24	1.29	1.19
	SubgraphRAG	33.17	34.50	34.58	34.08	8.75	10.07	10.11	9.64
	HyKGE	29.08	35.58	37.25	33.97	8.75	10.72	11.36	10.28
	DALK	24.42	29.08	32.67	28.72	10.42	12.38	13.68	12.16
	ToToFilter	42.42	43.00	43.25	42.89	31.39	31.99	32.19	31.86
Qwen2.5-7B	GoG	25.50	31.58	33.25	30.11	6.61	9.15	10.17	8.64
	MindMap	33.08	39.17	42.50	38.25	0.57	0.82	0.85	0.75
	SubgraphRAG	39.33	41.75	41.92	41.00	12.71	14.33	14.39	13.81
	HyKGE	25.58	29.83	31.42	28.94	9.68	11.00	11.51	10.73
	DALK	35.67	40.17	43.58	39.81	19.18	21.89	23.60	21.56
	ToToFilter	46.67	47.00	47.08	46.92	39.81	40.33	40.49	40.21

Table 3: Impact of TPF scoring components on performance and graph topology. Best results are highlighted in **bold**.

ID	Sem	Len	Deg	Acc $\uparrow$	Avg Hops $\downarrow$	Avg Deg $\downarrow$
MedQA						
0				61.35	2.93	2380.39
1	✓			61.84	2.86	2142.04
2		✓		61.32	2.67	1138.59
3			✓	61.43	2.73	1018.21
4	✓	✓		62.97	2.53	1183.57
5	✓		✓	62.69	2.59	1105.76
6		✓	✓	62.57	2.53	1034.76
7	✓	✓	✓	63.08	2.53	1058.64
MMLU						
0				78.40	2.75	1883.49
1	✓			78.81	2.75	1785.41
2		✓		78.45	2.63	1091.02
3			✓	78.75	2.73	1040.25
4	✓	✓		78.97	2.52	959.86
5	✓		✓	79.16	2.58	1020.76
6		✓	✓	79.24	2.54	1089.54
7	✓	✓	✓	79.35	2.52	965.09

ited in suppressing path-level noise. In contrast, ToToFilter improves both metrics simultaneously through TEL and TPF. Using the top-10 retrieved paths, ToToFilter also achieves consistently higher gold-triple precision and hit rate than DALK and MindMap (Appendix C.2).

5.4 Ablation Study

To validate the contribution of each core component in ToToFilter, we compare the full model against the vanilla LLM baseline and two ablation variants: (1) **w/o TEL** replaces diagnostic filtering and soft matching with naive truncation to the top-5 entities and threshold-free matching; (2) **w/o TPF** directly selects the top-10 recalled paths without

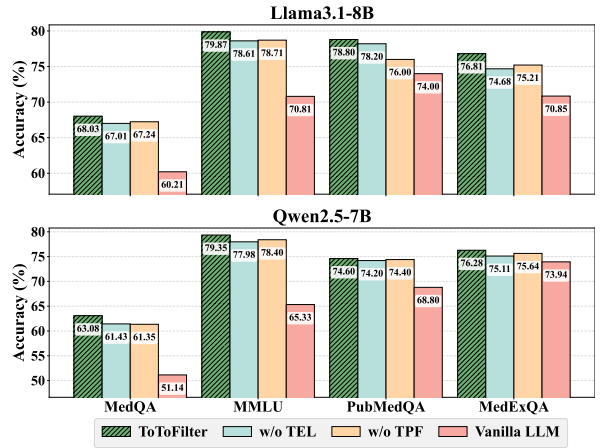


Figure 4: Ablation study of ToToFilter on Llama3.1-8B and Qwen2.5-7B. We compare the full model with two ablated variants and the vanilla LLM across four medical QA benchmarks.

multi-dimensional scoring or reranking.

As shown in Figure 4, removing either module reduces accuracy on both backbones. Removing TEL produces the larger decline, supporting core-entity filtering and soft matching, while removing TPF confirms the contribution of path scoring and reranking. A further decomposition shows that combining coarse scoring with LLM reranking achieves the highest macro-average accuracy on both backbones. Detailed results are provided in Appendix C.3.

5.5 TPF Scoring Mechanism Analysis

We analyze TPF by ablating three scoring terms in Eq. (7): semantic relevance (Sem), structural compactness (Len), and information specificity (Deg). For eight configurations, Table 3 reports accuracy, average hops, and average degree, where lower

Table 4: **Case study on MedQA.** Red marks hallucinated entities or spurious paths in baselines; green marks clinically relevant evidence identified by ToToFilter.

Question: 36F smoker, eyelid retraction, proptosis, extraocular muscle enlargement with tendon sparing, low TSH/high T4, TSI(+). Which intervention would most likely transiently worsen her eye symptoms? Options: (A) External orbital radiation (B) Selenium (C) Systemic corticosteroids (D) Radioactive iodine			
Method	Captured Entities	Reasoning Path	Answer & Rationale
MindMap	Smoke, CXCL8 , Selenium, IgG4 hepatopathy	Smoke → CXCL8 → inflammation → IgG4 hepatopathy	C ✗ Misframes the case as general inflammation and argues steroids worsen metabolic instability.
Subgraph-RAG	Palpitations , Heat intolerance , Arrhythmia	Palpitations → arrhythmia; Heat intolerance → temperature dysregulation	C ✗ Focuses on thyrotoxic burden and attributes worsening to steroid side effects.
DALK	Thyrotoxicosis, External orbital radiation , Extraocular muscle, Proptosis	External orbital radiation → extraocular muscle → local edema	A ✗ Treats radiation as an acute irritant causing transient worsening.
ToToFilter	Eyelid retraction , Proptosis , Smoking , TSI (+) , Radioactive iodine	(Orbitopathy + TSI) → Graves' ophthalmopathy ; Smoking → higher risk; Radioactive iodine → antigen release → flare	D ✓ Correctly identifies Graves' ophthalmopathy and links radioactive iodine to transient autoimmune flare.

hops and degree indicate higher retrieval quality.

Among the single-term settings (1–3), Sem (1) contributes most to accuracy, whereas Len (2) and Deg (3) mainly reduce hops and node degree. Pairwise combinations improve the balance between answer accuracy and path structure, and the full setting (7) attains the highest accuracy on both datasets. On MedQA, average degree decreases from 2380.39 to 1058.64 and average hops from 2.93 to 2.53. Additional parameter sensitivity analyses for the entity threshold, hop limit, path budgets, and TPF scoring weights are reported in Appendix C.4.

5.6 Retrieval Control Analysis

To better characterize the retrieval process, we conduct control analyses on option access, KG routing, query screening, and inference budget. Removing answer options from retrieval reduces macro-average accuracy by 1.50 and 0.84 percentage points on Llama3.1-8B and Qwen2.5-7B, respectively. KG paths are retrieved for 30.60% and 28.60% of samples, while disabling query screening lowers accuracy to 74.03 and 72.37. On Qwen2.5-7B, ToToFilter also uses 46.9% and 38.2% fewer tokens than DALK and MindMap under a shared 1,024-token generation cap.

We then evaluate whether the retained paths contribute to answer generation. Path perturbations change 15.97–16.71% of predictions and reduce macro-average accuracy by 0.84–1.19 percentage points. Automatic evidence-faithfulness evaluation on Qwen2.5-7B/MedQA further shows higher support scores for ToToFilter than DALK

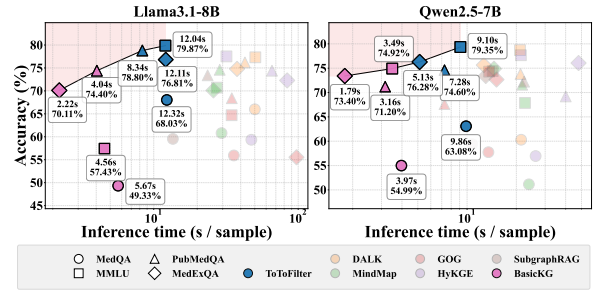


Figure 5: **Pareto analysis (accuracy vs. inference time).** Points denote method–benchmark results (color: method; marker: benchmark; x-axis: inference time in log). The black curve is the Pareto frontier; shading marks the preferred upper-left region.

and MindMap among samples with retrieved evidence. These results show that the retrieved paths have a measurable effect on final answer selection. Detailed settings and results are provided in Appendix C.5.

5.7 Pareto Efficiency and Accuracy Analysis

Figure 5 reports a Pareto analysis (Edelman et al., 2023) to examine the trade-off between accuracy and inference time. The Pareto frontier marks the optimal boundary where no method achieves both higher accuracy and lower latency.

ToToFilter lies on or near the Pareto frontier for both backbones. Across nearly all benchmarks, our method occupies the high-accuracy, moderate-latency segment, outperforming strong baselines such as DALK and HyKGE, which either require much longer inference times for similar accuracy or fall short in performance. MindMap and GoG

are dominated in most tasks. Notably, BasicKG achieves the lowest latency but at a substantial accuracy cost, whereas ToToFilter shifts the frontier upward to higher accuracy with only moderate inference time increase. The raw inference-time data used in this analysis are provided in Appendix C.6.

5.8 Case Study

Table 4 compares how different methods select entities and reasoning paths on a representative MedQA case. MindMap anchors on peripheral nodes such as “CXCL8” and “IgG4-related hepatopathy”, while SubgraphRAG drifts toward generic thyrotoxic signs. Both miss the orbital evidence central to the question. DALK retrieves partially relevant entities but confuses the treatment option “external orbital radiation” with a clinical finding. These entity errors further lead to spurious reasoning paths and incorrect answers. In contrast, ToToFilter preserves the decisive evidence, including “eyelid retraction, proptosis, smoking history, TSI (+),” and “radioactive iodine”, and forms a coherent reasoning chain centered on Graves’ ophthalmopathy, leading to the correct decision. Additional case studies are provided in Appendix C.7.

6 Conclusion

In this paper, we propose ToToFilter, a framework for mitigating interference from irrelevant entities and redundant reasoning paths in KG-based MQA. Extensive experiments on multiple MQA benchmarks show that ToToFilter achieves higher average accuracy than strong baselines across both LLM backbones. **Future work** will focus on developing query-adaptive weighting strategies, learning richer joint semantic-structural representations, and extending the framework to heterogeneous knowledge sources and broader clinical tasks.

Limitations

While ToToFilter improves both answer accuracy and retrieval quality, several limitations should be noted. First, many medical questions depend on temporality, severity, comorbidity, and patient-specific context, whereas general medical KGs mainly encode static relational facts. This mismatch limits the ability of retrieved paths to fully support clinical reasoning. Second, large medical KGs naturally contain many plausible multi-hop connections, making it difficult to completely avoid weakly diagnostic paths during retrieval.

Ethics Statement

This work uses public benchmarks and knowledge resources and does not collect or process personally identifiable information. The method is intended for research on evidence-grounded MQA and requires qualified clinician oversight for any clinical use. Llama3.1-8B and Qwen2.5-7B are used for academic research under their applicable licenses. AI assistants were used only for language polishing; all assisted text was reviewed and revised by the authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476060, No. 82472845), the National Key Research and Development Program of China (Grant No. 2023YFB3106504), and the Fujian Provincial Health Commission Science and Technology Project (No. 2022ZD01005).

References

- Bilal Abu-Salih, Muhammad Al-Qurishi, Mohammed Alweshah, Mohammad Al-Smadi, Reem Alfayez, and Heba Saadeh. 2023. [Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities](#). *Journal of Big Data*, 10(1):81.
- Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. [Building a knowledge graph to enable precision medicine](#). *Scientific Data*, 10(1):67.
- Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2026. [PathRAG: Pruning graph-based retrieval augmented generation with relational paths](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30183–30191.
- Enjun Du, Siyi Liu, and Yongqi Zhang. 2025. [Mixture of length and pruning experts for knowledge graphs reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 432–453, Suzhou, China. Association for Computational Linguistics.
- Benjamin L Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. 2023. [Pareto frontiers in deep feature learning: Data, compute, width, and luck](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 48021–48034.
- Yongqi Fan, Kui Xue, Zelin Li, Xiaofan Zhang, and Tong Ruan. 2025. [An LLM-based framework for biomedical terminology normalization in social media via multi-agent collaboration](#). In *Proceedings of*

- the 31st International Conference on Computational Linguistics, pages 10712–10726, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zihao Fu, Yixuan Su, Zaiqiao Meng, and Nigel Collier. 2023. [Biomedical named entity recognition via dictionary-based synonym generalization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14621–14635, Singapore. Association for Computational Linguistics.
- 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*.
- Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. [A dataset of medical questions paired with automatically generated answers and evidence-supported references](#). *Scientific Data*, 12(1):1035.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *CoRR*, abs/2009.03300.
- Takanobu Hirosawa, Yukinori Harada, Kazuya Mizuta, Tetsu Sakamoto, Kazuki Tokumasu, and Taro Shimizu. 2024. [Evaluating chatgpt-4’s accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians: Experimental study for diagnostic cases](#). *JMIR Formative Research*, 8:e59267.
- Nan Hu, Sheng Bi, Guilin Qi, Meng Wang, Yuncheng Hua, and Shirong Shen. 2022. [Improving core path reasoning for the weakly supervised knowledge base question answering](#). In *Database Systems for Advanced Applications*, pages 162–170. Springer.
- 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.
- Boran Jiang, Yuqi Wang, Yi Luo, Dawei He, Peng Cheng, and Liangcai Gao. 2024. [Reasoning on efficient knowledge paths: Knowledge graph guides large language model for domain question answering](#). In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 142–149. IEEE.
- Chunyang Jiang, Tianchen Zhu, Haoyi Zhou, Chang Liu, Ting Deng, Chunming Hu, and Jianxin Li. 2023. [Path spuriousness-aware reinforcement learning for multi-hop knowledge graph reasoning](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3181–3192, Dubrovnik, Croatia. Association for Computational Linguistics.
- Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. [HyKGE: A hypothesis knowledge graph enhanced RAG framework for accurate and reliable medical LLMs responses](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11856, Vienna, Austria. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [Pubmedqa: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. 2025. [Learning from negative samples in biomedical generative entity linking](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10714–10730, Vienna, Austria. Association for Computational Linguistics.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. [MedExQA: Medical question answering benchmark with multiple explanations](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.
- Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, Huan Liu, Li Shen, and Tianlong Chen. 2024. [DALK: Dynamic co-augmentation of LLMs and KG to answer alzheimer’s disease questions with scientific literature](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2187–2205, Miami, Florida, USA. Association for Computational Linguistics.
- Mufei Li, Siqi Miao, and Pan Li. 2025. [Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation](#). In *The Thirteenth International Conference on Learning Representations*.
- Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. 2025. [RGAR: Recurrence generation-augmented retrieval for factual-aware medical question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4006–4033, Suzhou, China. Association for Computational Linguistics.

- Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. [Biomedical entity linking as multiple choice question answering](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2390–2396, Torino, Italia. ELRA and ICCL.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. [Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models](#). *Bioinformatics*, 40(6):btac353.
- Nicki Newton, Adeola Bamgboje-Ayodele, Rowena Forsyth, Amina Tariq, and Melissa T Baysari. 2024. [How are clinicians’ acceptance and use of clinical decision support systems evaluated over time? a systematic review](#). *Studies in Health Technology and Informatics*, 310:259–263.
- Yingpeng Ning, Yuanyuan Sun, Ling Luo, Yanhua Wang, Yuchen Pan, and Hongfei Lin. 2025. [MedTrust-RAG: Evidence verification and trust alignment for biomedical question answering](#). In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1791–1794. IEEE.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. [Med-HALT: Medical domain hallucination test for large language models](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. [Evidence based medicine: what it is and what it isn’t](#). *BMJ*, 312(7023):71–72.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 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, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 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, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xun Sun, Xi Xiao, Zhaoguo Wang, Guangwu Hu, Xuhui Jiang, Bin Zhang, and Hao Li. 2024. [Understanding the influence of extremely high-degree nodes on graph anomaly detection](#). In *Pattern Recognition*, pages 19–34. Springer Nature Switzerland.
- Hein Minn Tun, Hanif Abdul Rahman, Lin Naing, and Owais Ahmed Malik. 2025. [Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review](#). *Journal of Medical Internet Research*, 27:e69678.
- Dandan Wang and Shiqing Zhang. 2024. [Large language models in medical and healthcare fields: applications, advances, and challenges](#). *Artificial Intelligence Review*, 57(11):299.
- Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. [MindMap: Knowledge graph prompting sparks graph of thoughts in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025a. [RAG-Gym: Optimizing reasoning and search agents with process supervision](#). *arXiv preprint arXiv:2502.13957*.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025b. [Improving retrieval-augmented generation in medicine with iterative follow-up questions](#). In *Pacific Symposium on Biocomputing*, volume 30, pages 199–214. World Scientific Publishing Company.

- Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. [Generate-on-graph: Treat LLM as both agent and KG for incomplete knowledge graph question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18410–18430, Miami, Florida, USA. Association for Computational Linguistics.
- Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. 2024. [Bridging the gap: Generating a comprehensive biomedical knowledge graph question answering dataset](#). In *ECAI 2024*, pages 1198–1205. IOS Press.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024a. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024b. [KG-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 155–166, Bangkok, Thailand. Association for Computational Linguistics.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. [Evaluation of retrieval-augmented generation: A survey](#). In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.
- Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. [Rankrag: Unifying context ranking with retrieval-augmented generation in llms](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 121156–121184. Curran Associates, Inc.
- Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022. [Subgraph retrieval enhanced model for multi-hop knowledge base question answering](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5773–5784, Dublin, Ireland. Association for Computational Linguistics.
- Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, Jian Wang, Donghu Sun, Henry Chung, Jun Zhou, Eric Zhou, Ben Lee, Peili Zhang, Xing Qiu, Tingting Zhao, and Jinfeng Zhang. 2025. [A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research](#). *Nat Mach Intell*, 7(4):602–614.
- 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 Web Conference 2025*, pages 4442–4457. Association for Computing Machinery.
- Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. [Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

A Complexity Analysis

In Algorithm 1, entity filtering and anchoring contribute $\mathcal{O}(|\mathcal{E}_{\text{core}}| \log |\mathcal{V}|)$. For the anchored node set $\mathcal{V}_Q$, all unordered node pairs are queried for path retrieval. For each pair $\{u, v\} \subseteq \mathcal{V}_Q$, Neo4j `allShortestPaths` performs graph traversal with cost $\mathcal{O}(|\mathcal{V}| + |\mathcal{R}|)$ and returns $K_H(u, v)$ shortest paths whose length does not exceed H . Returning these paths therefore contributes an additional $\mathcal{O}(HK_H(u, v))$ cost. Over all anchor pairs, the path-retrieval cost is

$$\mathcal{O} \left(|\mathcal{V}_Q|^2 (|\mathcal{V}| + |\mathcal{R}|) + \sum_{\{u,v\} \subseteq \mathcal{V}_Q} HK_H(u, v) \right). \quad (10)$$

After retrieval, the returned paths are merged, deduplicated, and truncated to at most N_{max} candidates. Sorting the retained candidate paths by the coarse score contributes $\mathcal{O}(|\mathcal{T}_{\text{cand}}| \log |\mathcal{T}_{\text{cand}}|)$. The overall time complexity is therefore

$$\begin{aligned} &\mathcal{O} \left(|\mathcal{E}_{\text{core}}| \log |\mathcal{V}| \right. \\ &\quad + |\mathcal{V}_Q|^2 (|\mathcal{V}| + |\mathcal{R}|) \\ &\quad + \sum_{\{u,v\} \subseteq \mathcal{V}_Q} HK_H(u, v) \\ &\quad \left. + |\mathcal{T}_{\text{cand}}| \log |\mathcal{T}_{\text{cand}}| \right). \end{aligned} \quad (11)$$

Here, $K_H(u, v)$ denotes the number of shortest paths returned for an anchor pair before truncation and depends on the graph structure, hop limit, and node degree.

The implementation queries all unordered anchor pairs sequentially. After all pair queries are completed, the returned paths are merged in query order, and duplicate paths retain their first occurrence. The merged set is then truncated to at most $N_{\text{max}} = 500$ paths. The database queries do not

use $N_{\max}$ for early stopping. Coarse ranking subsequently retains at most $N_c = 50$ paths, and LLM reranking selects at most $N_f = 10$ paths.

Excluding storage of the KG and LLM parameters, storing the returned paths before truncation requires $\mathcal{O}\left(H \sum_{\{u,v\} \subseteq \mathcal{V}_Q} K_H(u,v)\right)$ space. After truncation, the retained candidate paths require at most $\mathcal{O}(N_{\max}H)$ space.

B Additional Experimental Details

B.1 Dataset Details

We evaluate ToToFilter from two complementary perspectives: medical question answering (MQA) performance and retrieval quality. Unless otherwise noted, all results reported in this work are obtained on the public test split only.

For MQA evaluation, we use four benchmarks. **MedQA** (Jin et al., 2021) is a large-scale medical question answering benchmark derived from USMLE-style examination questions, which evaluates broad medical reasoning and diagnostic knowledge (*English split: train/dev/test = 10,178/1,272/1,273*). **MMLU** (Hendrycks et al., 2020) is evaluated on the medical-domain subsets used in our benchmark protocol, covering professional medicine, clinical knowledge, anatomy, college medicine, and related biomedical topics (*949 test questions*). **PubMedQA** (Jin et al., 2019) requires answering biomedical research questions with yes/no/maybe labels based on scientific abstracts, emphasizing evidence-grounded reasoning over literature content (*labeled split: train/validation/test = 450/50/500*). **MedExQA** (Kim et al., 2024) is designed for expert-aligned medical explanation and answer generation, placing greater emphasis on explanation faithfulness in addition to final prediction accuracy (*development/test = 25/940, 965 questions in total*).

For retrieval evaluation, we use **PrimeKGQA** (Yan et al., 2024), a benchmark constructed over the PrimeKG medical knowledge graph. Unlike standard MQA benchmarks, PrimeKGQA provides executable query supervision and supporting graph evidence, making it suitable for evaluating the quality of entity localization and path retrieval (*83,999 question-answer pairs in total, with public train/validation/test splits*). We use this benchmark to assess whether the retrieved subgraphs cover the gold answer entities and supporting triples required for downstream reasoning.

B.2 Baseline Details

We compare ToToFilter against representative baselines spanning KG-based and retrieval-augmented reasoning frameworks.

BasicKG is a simple KG retrieval baseline constructed to reflect a standard graph-augmented QA pipeline. It first extracts entities from the input question using an LLM, aligns them to KG nodes through embedding similarity, retrieves shortest paths between anchored entities, and then textualizes the retrieved evidence together with the candidate options for final answer generation. This baseline does not include topic-aware entity filtering, topology-aware path filtering, or reranking.

GoG (Xu et al., 2024) treats the LLM as both a reasoning agent and a knowledge source under a Thinking-Searching-Generating framework. It alternates between reasoning and external search, allowing the model to iteratively expand evidence during inference.

MindMap (Wen et al., 2024) builds a KG-based prompting framework that organizes retrieved knowledge into structured reasoning paths in the form of mind maps. By combining implicit parametric knowledge with external graph evidence, it aims to improve reasoning transparency and interpretability.

SubgraphRAG (Li et al., 2025) retrieves task-relevant subgraphs through learned graph scoring. It uses MLP-based parallel triple scoring to identify relevant graph fragments, which are then supplied to the language model as external evidence.

HyKGE (Jiang et al., 2025) extends KG exploration with hypothesis generation and fragment-level reranking. It explicitly models intermediate reasoning hypotheses and uses them to improve graph evidence selection.

DALK (Li et al., 2024) adopts a two-level retrieval framework over dynamic knowledge graphs. It combines staged retrieval with self-aware knowledge acquisition to improve evidence quality under complex reasoning settings.

B.3 Metric Details

We evaluate ToToFilter on both diagnostic accuracy and retrieval quality.

Diagnostic Accuracy. For MQA, we report **Accuracy**, defined as the proportion of questions for which the predicted answer exactly matches the gold label. We also report **Average Accuracy**, computed as the arithmetic mean of accuracy scores

across all four MQA datasets, to summarize overall model performance across benchmarks with different formats and difficulty levels.

Retrieval Quality. Following (Yan et al., 2024), we use **Answer Recall@k** and **Triple Recall@k**. **Answer Recall@k** measures whether the gold answer entities are covered by the retrieved subgraph. Specifically, after collecting the top- k retrieved triples, we extract all unique entities appearing in these triples and compute the proportion of gold answer entities covered by this entity set. **Triple Recall@k** measures the proportion of gold supporting triples recovered among the top- k retrieved triples, reflecting how well the retrieval module captures relevant relational evidence. We report both metrics at $k \in \{5, 8, 10\}$ to evaluate retrieval quality under different evidence budgets.

Path Precision. For the top- K retrieved paths, we expand the paths into triples and deduplicate the resulting triple set. Path Precision@ K is the proportion of these triples that match gold supporting triples. Hit@ K is the proportion of samples for which the top- K paths contain at least one gold supporting triple.

Paired Statistical Evaluation. For sample-aligned predictions, we compute paired-bootstrap 95% confidence intervals for the accuracy difference and apply the exact McNemar test to per-question correctness.

B.4 Implementation Details

All methods are implemented in PyTorch 2.6.0 and all experiments are conducted on two NVIDIA A800 GPUs. The backbone LLMs are deployed with vLLM through an OpenAI-compatible API interface. Unless otherwise specified, all methods use the same backbone-specific inference settings. Retrieved graph evidence is converted into textual context and concatenated with the question, candidate options, and task instruction for answer generation.

Details for ToToFilter. For entity alignment, we use Sentence-BERT (Reimers and Gurevych, 2019) with an MPNet backbone as $\text{Enc}(\cdot)$ and set $\tau = 0.92$. The maximum hop limit is $H = 4$, the candidate budget is $N_{\max} = 500$, and $(\alpha, \beta, \gamma) = (8.0, 0.5, 0.3)$. Coarse ranking retains at most $N_c = 50$ paths, and reranking selects at most $N_f = 10$ paths.

Table 5: Paired statistical comparison with DALK.

Backbone	Dataset	N	Δ Acc.	95% CI	McNemar p
Qwen2.5-7B	MedQA	1272	+2.75	[0.39, 5.11]	0.026
	MMLU	949	+0.63	[-1.69, 2.85]	0.651
	PubMedQA	500	+0.80	[-2.00, 3.80]	0.683
	MedExQA	940	+0.64	[-1.28, 2.55]	0.586
Llama3.1-8B	MedQA	1273	+2.04	[-0.47, 4.56]	0.131
	MMLU	948	+2.53	[-0.84, 6.01]	0.168
	PubMedQA	500	+2.60	[0.00, 5.20]	0.066
	MedExQA	940	+2.02	[-0.32, 4.26]	0.096

Table 6: Path precision and hit rate on PrimeKGQA (%).

Backbone	Method	P@10	Hit@10
Llama3.1-8B	DALK	3.29	24.56
	MindMap	0.36	3.46
	ToToFilter	27.20	63.99
Qwen2.5-7B	DALK	6.21	38.50
	MindMap	0.24	2.00
	ToToFilter	48.25	77.71

C Additional Experimental Results

C.1 Paired Statistical Evaluation

Predictions are aligned by sample ID before comparison. Table 5 reports ToToFilter relative to DALK, with confidence intervals expressed in accuracy percentage points.

Qwen2.5-7B/MedQA is the only comparison with $p < 0.05$ and a confidence interval above zero.

C.2 Path Precision on PrimeKGQA

Table 6 reports path precision and hit rate using the top-10 retrieved paths. ToToFilter achieves higher gold-triple precision and hit rates than DALK and MindMap on both backbones.

C.3 TPF Stage Decomposition

Coarse-only retains Eq. (7) and removes Step C. Rerank-only disables all three coarse-scoring terms and applies the LLM reranker to the same recalled path set. TEL, path recall, and answer generation remain fixed.

As shown in Table 7, the full pipeline achieves the highest macro-average accuracy on both backbones. Across the eight backbone-dataset settings, it performs best in seven, with coarse-only exceeding the full pipeline by 0.75 points on Qwen2.5-7B/PubMedQA. Among KG-success samples, coarse ranking reduces 112.98–241.61 recalled paths to 27.13–35.71, and LLM reranking retains 6.32–7.72 final paths.

Table 7: Macro-average accuracy and retained paths for the two TPF stages.

Backbone	Coarse-only		Rerank-only		Full
	Acc.	Paths	Acc.	Paths	
Llama3.1-8B	74.55	11.74	74.01	2.64	75.88
Qwen2.5-7B	71.84	10.22	72.37	2.04	73.33

Table 8: Macro-average accuracy under TEL-threshold and hop-limit variations. Values follow the listed setting order.

Parameter	Values	Llama3.1-8B	Qwen2.5-7B
TEL threshold τ	0.8 / 0.92 / 1.0	73.38 / 75.88 / 68.97	71.33 / 73.33 / 64.80
Hop limit H	4 / 6 / 8	75.88 / 72.81 / 73.66	73.33 / 71.36 / 70.06

C.4 Sensitivity Analysis of TEL and TPF Hyperparameters

As shown in Table 8, $\tau = 0.92$ and $H = 4$ achieve the highest macro-average accuracy on both backbones. Increasing the entity threshold to 1.0 or expanding the hop limit to 6 or 8 consistently reduces accuracy, indicating that broader entity or path coverage introduces additional retrieval noise.

On Qwen2.5-7B/MedQA, we further vary one TPF scoring weight at a time. The accuracy ranges for α , β , and γ are 60.80–63.08, 60.72–63.08, and 60.09–63.08, respectively, with the default value attaining the highest accuracy in each scan. Path-budget scans cover $N_c \in \{10, 20, 30, 40, 50, 60, 80, 100\}$ and $N_f \in \{3, 5, 8, 10, 12, 15\}$. The default $N_c = 50$ and $N_f = 10$ are optimal or near-optimal in most settings and primarily bound the evidence context.

C.5 Retrieval Controls and Evidence Dependence

This subsection reports detailed results for option access, KG routing and query screening, inference budget, path perturbation, and evidence faithfulness.

Option access. In the Q-only setting, answer options are excluded from TEL entity extraction and re-ranking, KG anchoring, TPF scoring, and LLM path re-ranking, and are provided only for final multiple-choice prediction. Unsuccessful outputs are counted as incorrect.

As shown in Table 9, Q-only retrieval lowers macro-average accuracy by 1.50 and 0.84 percentage points on Llama3.1-8B and Qwen2.5-7B, respectively. Qwen2.5-7B/MMLU is the only setting that improves, from 79.35 to 80.49. When answer

Table 9: Accuracy (%) under the default and Q-only retrieval settings.

Backbone	Setting	MedQA	MMLU	PubMedQA	MedExQA	Avg.
Llama3.1-8B	Full	68.03	79.87	78.80	76.81	75.88
	Q-only	66.90	78.50	77.15	74.97	74.38
Qwen2.5-7B	Full	63.08	79.35	74.60	76.28	73.33
	Q-only	61.51	80.49	72.89	75.08	72.49

Table 10: Per-dataset KG routing ratios (%).

Backbone	Dataset	N	Success	Skipped_OOD	no_entity	no_path
Llama3.1-8B	MedQA	1273	45.01	1.57	32.21	21.21
	PubMedQA	500	21.40	3.80	70.60	4.20
	MMLU	949	29.40	9.69	45.21	15.70
	MedExQA	940	17.13	21.91	49.04	11.91
Qwen2.5-7B	MedQA	1273	41.87	1.73	36.37	20.03
	PubMedQA	500	20.00	6.80	70.00	3.20
	MMLU	949	28.13	11.59	46.58	13.70
	MedExQA	940	15.74	27.87	47.87	8.51

options are also removed from final answer generation, macro-average F1/EM are 29.95/17.56 and 18.68/7.62 for Llama3.1-8B and Qwen2.5-7B, respectively, providing a preliminary evaluation of open-text transfer on the same benchmarks.

KG routing and query screening. Each sample is assigned to one of four mutually exclusive routing states. *Success* denotes successful retrieval of KG path evidence; *Skipped_OOD* denotes a question classified as OTHER; *no_entity* denotes fewer than two high-confidence anchors; and *no_path* denotes sufficient anchors but no usable graph path.

As shown in Table 10, the dataset-size-weighted Success/Skipped_OOD/no_entity/no_path ratios are 30.60/9.20/45.10/15.10 for Llama3.1-8B and 28.60/11.70/46.60/13.10 for Qwen2.5-7B. In both backbones, no_entity accounts for the largest proportion of KG bypass. Disabling query screening reduces macro-average accuracy from 75.88 to 74.03 and from 73.33 to 72.37, respectively. The screening step takes 0.168 and 0.103 seconds per item.

Inference budget. We measure actual API usage over the 3,662 Qwen2.5-7B test samples across the four MQA datasets under a shared 1,024-token generation cap.

As shown in Table 11, ToToFilter uses 46.9% fewer tokens than DALK and 38.2% fewer than MindMap. Its average number of LLM calls is lower than DALK and close to MindMap.

Path perturbation. To evaluate the contribution of retrieved paths to answer prediction, we perturb the final evidence set by removing the three highest-ranked paths, replacing paths with those

Table 11: Average API usage per sample with Qwen2.5-7B.

Method	Tokens/item	Calls/item
Vanilla LLM	288.8	1.00
MindMap	5,618.8	4.94
DALK	6,543.4	6.87
ToToFilter	3,472.7	5.50

Table 12: Macro-average accuracy and prediction changes under path perturbations.

Setting	Acc. (%)	Δ Acc. (pp)	Flip Rate (%)
Original paths	74.60	–	–
Remove top-3	73.42	-1.19	15.97
Random replacement	73.63	-0.97	16.71
Low-score only	73.41	-1.19	16.03
No evidence	73.77	-0.84	16.41

from other samples, retaining only low-ranked paths, or removing all KG evidence. The results are macro-averaged across both backbones and all four datasets.

As shown in Table 12, all four perturbations reduce macro-average accuracy by 0.84–1.19 percentage points and change 15.97–16.71% of predictions, showing that the retrieved paths affect final answer selection.

Automatic evidence faithfulness. We further evaluate whether the retrieved evidence supports final predictions on Qwen2.5-7B/MedQA. An automatic judge assigns a support score of 2 for direct support, 1 for partial but insufficient support, and 0 for absent, irrelevant, or unsupported evidence. The Evidence column reports the proportion of samples with retrieved evidence. *Faith. all* is computed over all samples, whereas *Faith. E*, *Strong E*, and *Correct+Faith. E* are computed over samples with retrieved evidence.

Among samples with retrieved evidence, ToToFilter obtains higher support scores than DALK and MindMap. Its all-sample score is lower than DALK because samples without KG evidence receive zero, while the absolute scores remain low for all three methods.

C.6 Details for Pareto Efficiency and Accuracy Analysis

Tables 14 and 15 report the inference times used in Figure 5.

Across both backbones, ToToFilter is faster than DALK, MindMap, SubgraphRAG, and HyKGE on all four datasets, while BasicKG remains the fastest method.

Table 13: Automatic evidence-faithfulness evaluation on Qwen2.5-7B/MedQA (%).

Method	Evidence	Faith. all	Faith. E	Strong E	Correct+Faith. E
DALK	100.00	3.42	3.42	2.59	2.44
MindMap	100.00	1.89	1.89	1.65	1.57
ToToFilter	43.52	2.91	6.68	5.23	4.69

Table 14: Inference time (s/iteration), Llama3.1-8B.

Method	MedQA	MMLU	PubMedQA	MedExQA
BasicKG	5.67	4.56	4.04	2.22
GoG	35.85	34.61	34.70	97.27
MindMap	29.48	27.33	28.64	25.90
SubgraphRAG	13.57	15.11	23.61	14.10
HyKGE	47.39	31.87	65.85	83.89
DALK	50.08	50.50	42.78	37.73
ToToFilter	12.32	12.04	8.34	12.11

C.7 Detailed Case Studies

As a supplement to Section 5.8, we present two detailed case studies through a comprehensive comparison of the retrieval results, reasoning paths, and predicted answers of the baselines and ToToFilter.

To compare how different KG-enhanced reasoning methods retrieve evidence and construct answer paths, we report a thyroid-related clinical case in Table 16. Although several baseline methods capture partially relevant clinical entities, they fail to establish the critical connection between Graves’ ophthalmopathy, smoking history, and the transient worsening effect of radioactive iodine therapy. In contrast, ToToFilter retrieves more targeted evidence and produces a reasoning chain that better aligns with the core diagnostic and therapeutic cues in the question.

To compare the retrieved entities, reasoning paths, and predicted answers produced by different methods, we analyze a gynecological emergency involving ectopic pregnancy and prior untreated pelvic infection in Table 17. The comparison shows that baseline methods tend to overemphasize direct but overly generic pathogen–disease associations, whereas ToToFilter more accurately captures the latent clinical progression from untreated, indolent pelvic inflammatory disease to tubal scarring and ectopic pregnancy.

D Prompt Templates

This section provides the prompts used in ToToFilter. For TEL, the entity extraction and re-ranking prompts are shown in Tables 19 and 20. Candidate entities are extracted from the question and options, re-ranked by their clinical roles (core anchors, key discriminators, target candidates, and context), and

Table 15: Inference time (s/iteration), Qwen2.5-7B.

Method	MedQA	MMLU	PubMedQA	MedExQA
BasicKG	3.97	3.49	3.16	1.79
GoG	13.50	13.64	7.28	15.13
MindMap	23.86	22.59	21.70	14.61
SubgraphRAG	10.74	14.33	22.29	13.01
HyKGE	26.23	21.31	40.06	47.99
DALK	21.40	21.05	21.21	12.66
ToToFilter	9.86	9.10	7.28	5.13

then parsed, deduplicated, and truncated to form $\mathcal{E}_{\text{core}}$.

For TPF, the path re-ranking prompt $\mathcal{P}_{\text{rerank}}$ is provided in Table 21. After path selection, $\mathcal{P}_{\text{verb}}$ (Table 22) converts the selected KG paths into textual evidence:

$$C = \text{Txt}(\hat{\mathcal{T}}; \mathcal{P}_{\text{verb}}). \quad (12)$$

If $\hat{\mathcal{T}} = \emptyset$, we set $C = \emptyset$. The answer generation prompt $\mathcal{P}_{\text{gen}}$ is provided in Table 23.

The query-screening prompt used before KG retrieval is provided in Table 18.

Table 16: Detailed comparison of reasoning paths for Case Study 1.

Item	Content
Question	A 36-year-old woman comes to the clinic because of tearing and a foreign body sensation in her eyes bilaterally, which has gradually worsened over the last several weeks. She also notes having occasional palpitations, nervousness, sweating, and heat intolerance. Her past medical history is unremarkable. She reports a 20-pack-year smoking history and is currently a daily smoker. Physical examination shows an anxious, trembling woman. She has eyelid retraction bilaterally, with an inability to fully close her eyes. Her extraocular motility is limited on upgaze. There is no thyromegaly, and no thyroid nodules are noted. Laboratory studies reveal a thyroid-stimulating hormone level of $0.1 \mu\text{U/mL}$ and total T4 of $42 \mu\text{g/dL}$. Thyroid-stimulating immunoglobulin is positive. CT scan of the orbits shows proptosis and marked enlargement of the extraocular muscle with sparing of the tendons. Which of the following would most likely transiently worsen this patient's eye symptoms?
Options	A. External orbital radiation B. Selenium supplementation C. Systemic corticosteroids D. Treatment with radioactive iodine
Method	Retrieval and Reasoning Process
MindMap	Captured Entities: Smoke, CXCL8 (Interleukin-8), IgG4-related hepatopathy, Fever, Selenium, Piroxicam, Multiple endocrine neoplasia. Path: Smoke → CXCL8 → Inflammation → IgG4-related hepatopathy → Systemic drug interaction → Selenium → Piroxicam. Reasoning: The clinical picture is dominated by markers of systemic inflammation (CXCL8, Fever) and complex pathology potentially involving IgG4-related disease and hepatic abnormalities. In the context of such unstable systemic inflammation and potential liver involvement, the administration of high-dose systemic corticosteroids is associated with a risk of immune suppression and metabolic disruption, which could aggravate the underlying inflammatory dysregulation suggested by the extracted path, rather than addressing the specific thyroid pathology. Predicted Answer: C (Incorrect)
SubgraphRAG	Captured Entities: Palpitations, Arrhythmia, Eyelid retraction, Heat intolerance, Abnormality of temperature regulation. Path: Palpitations → parent-child → Arrhythmia; Heat intolerance → parent-child → Abnormality of temperature regulation. Reasoning: The patient's primary complaints of palpitations (arrhythmia) and heat intolerance (thermoregulation abnormality) indicate a state of heightened sympathetic activity and metabolic instability. Systemic corticosteroids are well-documented to induce side effects such as fluid retention, hypertension, and tachycardia, which would directly exacerbate these specific cardiovascular and thermoregulatory symptoms. Without specific knowledge of thyroid-eye interactions, corticosteroids are identified as the agent most likely to compound the existing systemic burden. Predicted Answer: C (Incorrect)
DALK	Captured Entities: Thyrotoxicosis, External orbital radiation, Extraocular muscle, Proptosis. Path: External orbital radiation → target → Extraocular muscle → effect → Local inflammation/Edema. Reasoning: The patient exhibits marked enlargement of the extraocular muscles and proptosis. External orbital radiation involves the direct application of high-energy rays to the orbital tissues. In the acute phase, radiation therapy inevitably induces local tissue inflammation and edema in the already enlarged muscles before any therapeutic fibrosis occurs. This direct physical trauma to the orbit is interpreted as the primary cause for transient worsening of ocular symptoms compared to systemic pharmacological interventions. Predicted Answer: A (Incorrect)
ToToFilter	Captured Entities: Eyelid retraction, Proptosis, Smoking history (20-pack-year), Thyroid-stimulating immunoglobulin (Positive), Radioactive iodine, Systemic corticosteroids. Path: Patient symptoms + TSI positive → Graves' ophthalmopathy; Graves' ophthalmopathy + Smoking → High risk of progression; Radioactive iodine → Immunogenic surge → Exacerbation of orbitopathy. Reasoning: The patient presents with active Graves' ophthalmopathy characterized by proptosis and eyelid retraction, superimposed on a significant smoking history. While radioactive iodine is a standard treatment for hyperthyroidism, it causes a release of thyroid antigens that can trigger an autoimmune response, leading to a transient but significant worsening of ocular symptoms, particularly in smokers. In contrast, systemic corticosteroids are indicated prophylactically to prevent this exacerbation. Therefore, the intervention most likely to worsen the condition is radioactive iodine. Predicted Answer: D (Correct)

Table 17: Detailed comparison of reasoning paths for Case Study 2.

Item	Content
Question	A 24-year-old married woman presents to the emergency department with severe abdominal pain since last night. She also complains of scant vaginal bleeding. She says she visited a physician last year who said she had a pelvic infection, but she was never treated because of insurance issues. She also says her period has been delayed this month. She is afebrile. The pulse is 124/min and the blood pressure is 100/70 mm Hg. On examination, her abdomen is distended and tender. A pregnancy test was positive. A complication of infection with which of the following organisms most likely led to this patient's condition?
Options	A. Candida albicans B. Neisseria gonorrhoeae C. Chlamydia trachomatis D. Haemophilus influenzae
Method	Retrieval and Reasoning Process
MindMap	Captured Entities: Weak pulse, Acute gonococcal cervicitis, Sinusitis, Recurrent Haemophilus influenzae infections. Path: Acute cervicitis → disease disease → Acute gonococcal cervicitis → indication → Antibiotics. Reasoning: The patient's history of pelvic infection is mapped via the reasoning path to "Acute cervicitis" and specifically "Acute gonococcal cervicitis." The path retrieves explicit links between acute cervical inflammation and Neisseria gonorrhoeae. Given the severity of the abdominal pain described, the logic aligns with the clinical presentation of gonorrhea, which is known for causing more acute and intense symptomatic PID compared to other pathogens. The retrieval emphasizes the direct classification of cervicitis/PID with gonorrhea, leading to its selection as the causative agent. Predicted Answer: B (Incorrect)
SubgraphRAG	Captured Entities: Abdominal pain, IFNA2 (Interferon Alpha-2), Hepatitis. Path: Abdominal pain → associated with → IFNA2 → associated with → Inflammation/Hepatitis. Reasoning: The specific retrieval of IFNA2 and hepatitis in the context of severe abdominal pain points toward an aggressive infectious etiology. In a gynecological setting, this constellation strongly implies Fitz-Hugh–Curtis syndrome, a classic complication of virulent gonococcal infection involving liver capsule inflammation. Consequently, the decision is driven by the alignment between retrieved inflammatory markers, the severity of local symptoms, and the pathogen's high-pathogenicity profile, favoring Option B over the typically subclinical course of Chlamydia. Predicted Answer: B (Incorrect)
DALK	Captured Entities: Abdominal pain, Pelvic inflammatory disease, Vaginal disease, Neisseria gonorrhoeae. Path: Female reproductive system disease → Pelvic inflammatory disease → disease disease → Neisseria gonorrhoeae. Reasoning: Neisseria gonorrhoeae is recognized as the prototypical causative agent of acute pelvic inflammatory disease (PID), traditionally associated with intense localized inflammation and rapid symptomatic onset. The reasoning is heavily dictated by the direct disease-disease relationship between PID and Neisseria found in the knowledge graph, positioning this organism as the primary diagnostic anchor. By prioritizing this canonical association, the logic aligns the patient's acute abdominal distress with the high virulence and acute pathogenicity profile of Neisseria. Consequently, the decision is driven by the strongest retrieved etiological link, failing to account for the indolent, asymptomatic nature of the untreated history, which would clinically favor the subclinical presentation of Chlamydia. Predicted Answer: B (Incorrect)
ToToFilter	Captured Entities: Abdominal pain (Severe), Vaginal bleeding, Pregnancy test (+), Pelvic infection (Untreated), Pulse (124/min), Blood pressure (100/70 mm Hg). Path: Patient symptoms (Pain + Bleeding + Positive Pregnancy) → Ectopic pregnancy; Untreated pelvic infection → Tubal scarring → Ectopic pregnancy; Chlamydia trachomatis → Indolent/Asymptomatic PID → High risk of untreated scarring. Reasoning: The patient presents with the classic triad of ectopic pregnancy: abdominal pain, amenorrhea, and vaginal bleeding, confirmed by a positive pregnancy test. The vital signs suggest a ruptured ectopic pregnancy requiring emergency care. The critical historical factor is the untreated pelvic infection. Chlamydia trachomatis is the most common cause of "silent" or indolent PID, which frequently goes untreated due to mild symptoms, leading to significant tubal scarring that predisposes the patient to ectopic pregnancy. In contrast, Neisseria typically causes acute, symptomatic infection that is more likely to prompt treatment. Predicted Answer: C (Correct)

Table 18: Query-screening prompt ($\mathcal{P}_{\text{dom}}$) used before KG retrieval.

Variable	Prompt Content
$\mathcal{P}_{\text{dom}}$	Task: Classify the domain of the following medical exam question. Question: “{question_text}” Categories: 1. BIOMEDICAL: Questions about Diseases, Drugs, Anatomy, Physiology, Genetics, Clinical Diagnosis, Pathology, Treatment. (Requires Knowledge Graph) 2. OTHER: Questions about Medical Physics, IT Systems (PACS, DICOM), Hospital Management, Medical Ethics, Professionalism, Statistics, Legal issues. (Knowledge Graph is irrelevant) Instructions: - If the question mentions specific proteins, genes, or clinical symptoms, choose BIOMEDICAL. - If the question mentions computer systems, protocols, or ethical dilemmas, choose OTHER. Output ONLY one word: “BIOMEDICAL” or “OTHER”.

Table 19: Prompt for Entity Extraction (Step A in the TEL module).

Variable	Prompt Content
–	Extract all the biomedicine-related entity from the following question and choices, output each entity in a single line with a serial number (1., 2., ...). Question: {question_text} The extracted entities are:

Table 20: Prompt for Entity Re-ranking (Step A in the TEL module).

Variable	Prompt Content
–	You are an expert medical reasoning assistant. I will provide you with a raw list of entities extracted from a question. Your task is to re-rank and select the top {max_entities} entities based on their importance for constructing a reasoning path to answer the question. The User’s Question: “{question_text}” The Raw List of Entities: — {raw_entity_text} — Ranking Criteria (Order of Importance): 1. Core Anchors (Highest Priority): The central subject in the question text, such as the drug administered, known disease history, exposure, intervention, or a specific gene mutation. These are the start nodes of reasoning. 2. Key Discriminators: Specific symptoms, lab findings, mechanisms, or physiological changes mentioned in the text that help distinguish the correct answer. 3. Target Candidates: Key entities found in the answer choices, such as potential side effects, diagnoses, treatments, or mechanisms. These are the end nodes of reasoning. 4. Context: General anatomical terms or vague symptoms, which receive the lowest priority unless needed to support a medical relation. OUTPUT FORMAT INSTRUCTIONS (STRICTLY FOLLOW): 1. Output ONLY the entity names. 2. NO markdown formatting. 3. NO explanations, definitions, or labels. 4. NO introductory text. 5. Output each entity on a new line. Example of CORRECT Output: Risperidone Agranulocytosis Schizophrenia Gynecomastia Weight loss Your Ranked List:

Table 21: Prompt for Path LLM Re-ranking ($\mathcal{P}_{\text{rerank}}$), corresponding to Eq. (8).

Variable	Prompt Content
$\mathcal{P}_{\text{rerank}}$	You are a world-class expert in medical knowledge graphs. You must precisely select the top {MAX_FINAL_PATHS} paths from the list below that are most critical for answering this multiple-choice question. The selected paths must be directly relevant to the specific options and capable of distinguishing between correct and incorrect answers. Question: {question_text} Path List: {raw Path List} Output ONLY the numbers of the most relevant paths, separated by commas. If none are relevant or if they are duplicates, output NONE. Absolutely NO explanations, NO filler text, and NO sentences like “none of them”—just output the numbers or NONE.

Table 22: Prompt for Path Textualization ($\mathcal{P}_{\text{verb}}$), corresponding to Eq. (12).

Variable	Prompt Content
$\mathcal{P}_{\text{verb}}$	You are a PrimeKG expert. PrimeKG is an UNDIRECTED graph. For every triple, you MUST output the CORRECT clinical meaning according to these OFFICIAL PrimeKG rules: 1. contraindication: ALWAYS “Patients with [disease/phenotype] should NOT use [drug]” → No matter the order, the second entity is the drug if one is a drug 2. drug_effect: ALWAYS “[drug] may cause [effect/phenotype]” 3. indication: ALWAYS “[drug] is approved to treat [disease]” 4. off-label use: ALWAYS “[drug] is used off-label for [disease]” 5. disease_phenotype_positive/negative: → If one side is disease → “[disease] typically [presents with / does not present with] [phenotype]” 6. All other relations: use neutral “is associated with” Examples (order doesn’t matter): 1. Acetaminophen - drug_effect - Liver Failure → Acetaminophen may cause Liver Failure (if overdosed). 2. Penicillin - contraindication - Hypersensitivity → Patients with Hypersensitivity should NOT use Penicillin. 3. Metformin - indication - Diabetes Mellitus → Metformin is approved to treat Diabetes Mellitus. Now convert these triples: {Path} Output format: Evidence 1: [correct sentence] Evidence 2: [correct sentence] ... Only output the sentences, no extra text.

Table 23: Prompt for Answer Generation ($\mathcal{P}_{\text{gen}}$), corresponding to Eq. (9).

Variable	Prompt Content
$\mathcal{P}_{\text{gen}}(\text{CoT})$	You are an excellent AI assistant to answering the following question. Question: {question_text} You have some medical knowledge information in the following: {Natural Language Paths} Answer: Let’s think step by step:
$\mathcal{P}_{\text{gen}}(\text{Final})$	You are an excellent AI assistant to answering the following question. Question: {question_text} You have some medical knowledge information in the following: {Natural Language Paths} Answer: Let’s think step by step: {cot} The final answer (output the letter option) is: