Enhancing Clinical Multiple-Choice Questions Benchmarks with Knowledge Graph Guided Distractor Generation

Running Yang [*1] **Wenlong Deng** [*1 2] **Minghui Chen** [1] **Yuyin Zhou** [3] **Xiaoxiao Li** [1 2]

Abstract

Clinical tasks such as diagnosis and treatment require strong decision-making abilities, highlighting the importance of rigorous evaluation benchmarks to assess the reliability of large language models (LLMs). In this work, we introduce a knowledge-guided data augmentation framework that enhances the difficulty of clinical multiple-choice question (MCQ) datasets by generating distractors (*i.e.*, incorrect choices that are similar to the correct one and may confuse existing LLMs). Using our KG-based pipeline, the generated choices are both clinically plausible and deliberately misleading. Our approach involves multi-step, semantically informed walks on a medical knowledge graph to identify distractor paths—associations that are medically relevant but factually incorrect—which then guide the LLM in crafting more deceptive distractors. We apply the designed knowledge graph guided distractor generation (`KGGDG`) pipeline, to six widely used medical QA benchmarks and show that it consistently reduces the accuracy of state-of-the-art LLMs. These findings establish `KGGDG` as a powerful tool for enabling more robust and diagnostic evaluations of medical LLMs. Our code and experimentation dataset are released at `https://github.com/ryyrn/Knowledge_Graph_Guided_Distractor_Generation`.

1. Introduction

Large language models (LLMs) (OpenAI et al., 2024; Team et al., 2024; Liu et al., 2024; Qwen et al., 2024) have demonstrated impressive performance across various medical question-answering (QA) tasks, reaching near-expert accuracy and surpassing 80% precision on established clinical benchmarks such as MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). However, the saturation of these benchmarks limits their effectiveness in reliably evaluating LLMs under realistic diagnostic scenarios, which typically involve more complex and nuanced cases.

Recent efforts have focused on constructing more challenging benchmarks by leveraging real-world clinical questions (Chen et al., 2025; Xie et al., 2024). For instance, Chen et al. (2025) compile cases from the JAMA Network Clinical Challenge archive and curate USMLE Step 2 & 3-style questions from open-access tweets on X. Similarly, Xie et al. (2024) expand this effort by mining clinical questions from prestigious sources such as The Lancet and The New England Journal of Medicine. While these benchmarks advance the quality and authenticity of question design, the development of more sophisticated and clinically misleading distractors remains an open and underexplored area.

Early distractor generation methods rely on corpus-based techniques (Zesch & Melamud, 2014; Hill & Simha, 2016), leveraging syntactic patterns and similarity metrics. However, such approaches often produce distractors that are overly simplistic or semantically unrelated, especially in specialized domains like medicine (Alhazmi et al., 2024). To overcome these limitations, more recent approaches (Taslimipoor et al., 2024; Bitew et al., 2022) fine-tune pretrained models to generate harder distractors. While effective, these methods require additional annotated training data, which may not be readily available in the medical domain. Recent LLM-based prompting methods (Tran et al., 2023) eliminate the need for fine-tuning but often generate distractors based solely on implicit model knowledge, which can limit their effectiveness in generating challenging distractors.

Biomedical knowledge graphs (KGs) offer a promising solution to this challenge. KGs such as PrimeKG (Chandak et al., 2023) organize rich and structured relationships between medical entities. Recent work (Wu et al., 2025) demonstrates that incorporating knowledge graphs into reasoning pipelines can significantly enhance the reasoining quality and reliability, highlighting their potential to support more accurate and explainable medical inference.

[*]Equal contribution [1]University of British Columbia [2]Vector Institute [3]UC Santa Cruz. Correspondence to: Xiaoxiao Li <xiaoxiao.li@ece.ubc.ca>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

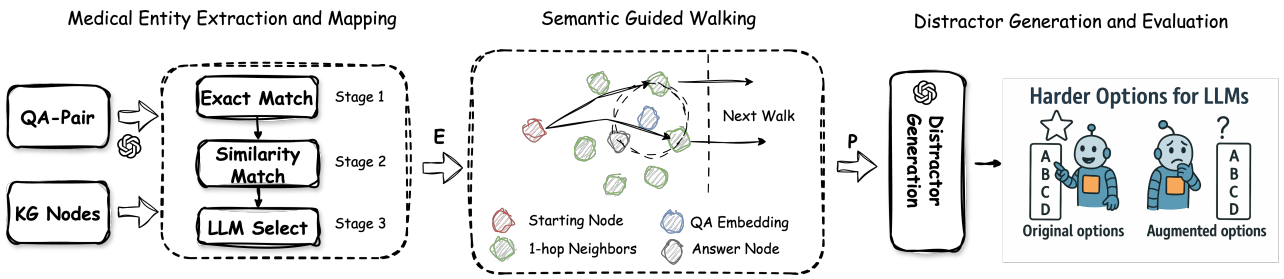

Figure 1. **Overview of Our Distractor Generation Pipeline.** We begin by extracting and mapping entities from each medical Q&A pair (see Section 3.1.1). Then, starting from the question nodes, we perform n-step semantic-guided (most similar to Q&A context) walks on the knowledge graph to reach nodes outside the answer set, generating distractor paths (see Section 3.1.2). These paths serve as factual but misleading cues for constructing challenging distractors. Finally, we use the obtained misleading paths to prompt an LLM to produce distractors, and evaluate their effectiveness on various LLMs (see Section 3.2).

In this paper, we propose a knowledge-guided distractor generation (KGGDG), designed to produce clinically plausible yet deliberately misleading distractors. Our method performs semantically guided walks over PrimeKG, starting from entities related to the question and avoiding those linked to the correct answer. These reasoning paths serve as structured and informative prompts for LLMs, enabling them to generate distractors that are both contextually relevant and diagnostically deceptive. We apply KGGDG to six widely used medical QA benchmarks and evaluate its impact across six state-of-the-art LLMs—including DeepSeek, Qwen, and Gemini. Our results demonstrate that KGGDG consistently reduces model accuracy, revealing weaknesses in clinical reasoning that are otherwise hidden by simpler distractor sets, and offering a more rigorous foundation for evaluating medical LLMs. Our contributions are as follows:

Semantic-Guided Reasoning Path Extraction: We introduce a semantic-guided walk algorithm over a biomedical knowledge graph to extract structured reasoning paths that inform distractor generation while explicitly avoiding the correct answer.

Knowledge-Guided Distractor Generation: We propose KGGDG, a novel framework that combines biomedical knowledge graphs with prompt-based LLM generation to produce clinically plausible but deliberately misleading distractors—enhancing the difficulty and diagnostic rigor of medical MCQ datasets.

Comprehensive Evaluation Across Benchmarks and Models: We apply KGGDG to six widely used medical QA benchmarks and evaluate its impact on six state-of-the-art LLMs, demonstrating that our method consistently reduces model accuracy and reveals gaps in clinical reasoning performance.

2. Related Work

Medical QA Benchmarks. Multiple-choice question (MCQ) benchmarks such as MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and the medical subset of MMLU (Hendrycks et al., 2021) are widely used to assess the clinical capabilities of LLMs. While these benchmarks have enabled significant progress, current LLMs have already demonstrated strong performance on many of them (Wu et al., 2025). To push the boundaries further, more difficult benchmarks such as MMMU-Pro (Wang et al., 2024) and the Human-Last Exam (Phan et al., 2025) have been proposed, introducing more challenging medical questions. However, the systematic design of difficult options remains underexplored in the medical domain.

KG and KG-Guided Data Augmentation. Biomedical knowledge graphs (KGs), such as Hetionet (Himmelstein & Baranzini, 2014), BioKG (Zhang et al., 2023), and PrimeKG (Chandak et al., 2023), structure biological and medical entities into relational networks that encode rich semantic relationships across diverse domains (Hogan et al., 2021; Cui et al., 2025; Lu et al., 2025). These KGs have been widely used for tasks such as concept grounding (Zhang et al., 2020), entity linking (Hogan et al., 2021), reasoning generation (Wu et al., 2025), and graph-based information retrieval (Peng et al., 2024). In the context of data augmentation, prior work has leveraged KGs for question generation by using predefined templates to verbalize structured queries or training decoders that copy node attributes directly from subgraphs (Chen et al., 2023). KGs have also been used to augment reasoning, such as guiding large language models (LLMs) with fact-based reasoning paths to produce coherent chains of thought (Wu et al., 2025). Despite these advances, the use of KGs for guiding distractor generation remains largely unexplored.

Distractor Generation Early distractor generation methods used corpus-based techniques (Zesch & Melamud, 2014; Hill & Simha, 2016), relying on syntactic patterns and similarity metrics. While simple, these often produced weak or irrelevant distractors, especially in domains like medicine (Alhazmi et al., 2024). Later approaches fine-tuned pretrained models to generate more challenging distractors (Taslimipoor et al., 2024; Bitew et al., 2022), but required annotated data, which is limited in specialized fields. Prompt-based methods (Tran et al., 2023) using LLMs eliminate the need for fine-tuning but depend on the model's inherent capabilities, which can limit their effectiveness in generating challenging distractors.

3. Method

In this section, we introduce our proposed pipeline, which utilizes KG to guide LLMs in generating more challenging answer options for clinical questions. Specifically, we replace the incorrect choices in the original QA benchmark datasets with more confusing wrong options by making them more similar to the correct answer, making it harder for the LLM to choose the right one. The method consists of two key components: (1) extraction of distract reasoning paths, and (2) path-guided distractor generation. We refer to knowledge graph-guided distractor generation piepline as KGGDG.

3.1. Extraction Distract Reasoning Paths from Knowledge Graph

In this section, we detail the process of retrieving distract reasoning paths from the knowledge graph G.

3.1.1. Medical Entity Extraction and Mapping

Given a Multi-Choice question answer pair $(Q, A, \mathbf{O}_{ori})$, we ulize the Language Model LLM to identify two sets of medical entities: one from the question text and one from the answer:

$$\mathcal{E}^Q = \{e_i^Q\}_{i\in[n]}, \quad \mathcal{E}^A = \{e_j^A\}_{j\in[m]},$$

where n and m denote the number of entities in Q and A, $\mathcal{E}^Q$ and $\mathcal{E}^A$ denote the entity sets extracted from the question and the answer, respectively. The union of these sets is denoted as $\mathcal{E} = \mathcal{E}^Q \cup \mathcal{E}^A$.

Then these entities are mapped to the corresponding nodes in the knowledge graph G through a three-step mapping process (as shown in Figure 1):

Stage 1 (Exact Match): For each entity e, the algorithm first checks whether there is an exact string match node on KG. If an exact match is found, the corresponding node is selected.

Stage 2 (Similarity Match): If an exact match is not found, we use a medical text embedding model ϕ (Balachandran, 2024) to encode each entity $e \in E$ and compute its similarity with the embeddings of node in the G and the top similarity score exceeds a predefined threshold τ (set to 0.85 in our case), the most similar entity is then selected.

$$\hat{e} = \arg\max_{s_k \in G} \cos(e, s_k), \text{ if } \cos(e, s_k) > \tau \tag{1}$$

Stage 3 (LLM-based Selection): If no suitable candidate is found in the first two stages, we prompt the LLM to analyze the question-answer context together with the entity name, and select the most relevant node from the top 10 most similar nodes identified in stage 2 (denoted as S).The selection prompt, P_{select}, is illustrated in the Appendix Appendix A.

$$\hat{e} = LLM\left(S, Q, A \mid P_{\text{select}}\right), \tag{2}$$

Finally, we derive mapped entity sets from the graph, denoted as $V^Q = \{v_i^Q\}_{i\in[n]}$ and $V^A = \{v^A\}_{j\in[m]}$, respectively.

3.1.2. Semantic Guided Distract Path Extraction

After mapping the entities, the next goal is to identify reasoning paths that are logically coherent yet lead to incorrect answers. To achieve the goal, we introduce an *n-step semantic-guided walk* over the KG, beginning at a node in the question set V^Q and terminating at a node outside the correct answer set V^A (as shown in Figure 1).

As shown in Algorithm 1, we begin by computing a guidance vector $\mathbf{z}_i$ by embedding the concatenated question and its correct answer:

$$\mathbf{z} = \phi(Q \| A),$$

where ϕ denotes the embedding function and $\mathbf{z} \in \mathbb{R}^d$. For each start node $v_i \in V^Q$, a beam search of depth n is then performed. At each step, we retrieve the 1-hop neighbors of the current node and *exclude* any that are part of the avoidance set V^A. The cosine similarity between each remaining neighbor's embedding and the guidance vector $\mathbf{z}_i$ is computed as:

$$\text{sim}(v') = \cos(\phi(v'), \mathbf{z}_i) = \frac{\phi(v') \cdot \mathbf{z}_i}{\|\phi(v')\| \cdot \|\mathbf{z}_i\|}.$$

We then select the top-k most semantically similar neighbors and extend the current path with each of them, forming the beam for the next step. This iterative expansion can produce up to k^n candidate paths per start node.

The final set $\mathbf{R}^Q$ consists of semantically plausible reasoning paths that deliberately avoid correct answers, thus serving as effective distractor paths.

Algorithm 1 Semantic Guided Distract Path Extraction

Input: Question Q, Answer A, Start nodes V^Q, Avoidance set V^A, Walk length n, Beam size k, Knowledge Graph G
Output: Set of reasoning paths $\mathbf{R}^Q$
Compute guidance vector: $\mathbf{z} = \phi(Q \| A)$
Initialize path set: $\mathbf{R}^Q \leftarrow []$
for each start node $v_i \in V^Q$ **do**
  Initialize beam: $\mathcal{B}_0 \leftarrow [[v_i]]$
  for $t = 1$ to n **do**
    Initialize new beam: $\mathcal{B}_t \leftarrow []$
    for each path $r \in \mathcal{B}_{t-1}$ **do**
      Let v_{t-1} be the last node in r
      Retrieve 1-hop neighbors: $\mathcal{N} = \mathcal{N}(v_{t-1}, G)$
      Exclude answer-related nodes: $\mathcal{N}' = \mathcal{N} \setminus \hat{\mathcal{E}}^A$
      Compute similarity scores:

$$\text{sim} = \{\cos(\phi(v'), \mathbf{z})\}_{v' \in \mathcal{N}'}$$

      Select top-k similar neighbors: $V^k \leftarrow$ Top-$k(\mathcal{N}', \text{sim})$
      for each $v' \in V^k$ **do**
        Extend path: $r' \leftarrow r \cup [v']$
        Add r' to $\mathcal{B}_t$
      end for
    end for
  end for
  Append all paths from $\mathcal{B}_n$ to $\mathbf{R}^Q$
end for
Return $\mathbf{R}^Q$

3.2. Path-Guided Distractor Generation

By leveraging the step-by-step distractor paths in $\mathbf{R}^Q$, we incorporate medically plausible yet incorrect knowledge into the distractor generation process.

Specifically, we prompt the *LLM* to analyze the given distractor reasoning paths and generate K distractor options—where K corresponds to the required number of distractors—that are both misleading and medically incorrect. As described in Appendix A, our prompt enforces three essential criteria for effective distractor generation: (1) each distractor must be strictly incorrect, (2) it must be highly misleading, and (3) it should incorporate the provided distractor reasoning path. This approach leads to the generation of more challenging distractors:

$$\mathbf{O} = LLM\left(Q, A, \mathbf{R}^Q \mid P_{\text{distract}}\right), \quad (3)$$

Here, $\mathbf{O}$ denotes the generated set of distractor options, and P_{distract} refers to our carefully designed prompting template. We then replace the original distractors $\mathbf{O}_{ori}$ with $\mathbf{O}$ to form a new multiple-choice question-answering instance, represented as $[Q, A, \mathbf{O}]$.

4. Experiment Setup

Choosing of LLM and KG. In our framework, we use Deepseek-V3 (Liu et al., 2024) as the LLM in the KGGDG pipeline, due to its strong performance in task completion and instruction following. For structured medical knowledge, we leverage PrimeKG (Chandak et al., 2023), which aggregates information from 20 high-quality biomedical resources to represent 17,080 diseases through 4,050,249 relationships across ten major biological scales: biological process, protein, disease, phenotype, anatomy, molecular function, drug, cellular component, pathway, and exposure.

Benchmark Dataset Our knowledge-guided augmentation pipeline is applied to six widely used medical multiple-choice question (MCQ) datasets: MedBullets (Chen et al., 2024), MedQA (USMLE)(Jin et al., 2021), Lacent(Xie et al., 2024), MedMCQA (validation set)(Pal et al., 2022), MedXpert(Zuo et al., 2025), and NEJM (Xie et al., 2024) QA. Each dataset comprises MCQs with one correct answer and several distractors. For every question, we preserve the original question text and correct answer, while replacing the distractors with those generated by our approach—resulting in a more challenging benchmark.

Evaluation Setup. We evaluate six large language models in different sizes across all datasets: DeepSeek V3 (Liu et al., 2024), Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct (Qwen et al., 2024), Gemini-1.5 Pro (Team et al., 2024), and Gemini-2.5 Flash. Each model is assessed under three different settings: (1) using the original distractors, (2) using distractors directly generated by an LLM, and (3) using distractors generated through our KGGDG. For each question, the model is prompted to select the most likely correct answer from the provided options. Accuracy is calculated as the percentage of correctly answered questions, averaged over three runs for all models.

5. Results

In this section, we present the main results, and additional ablation studies are provided in Appendix A..

KGGDG Introduces Greater Challenge: As shown in Table 1, replacing the original distractors with KGGDG leads to a noticeable drop in model accuracy. For example, DeepSeek V3's performance declines from 67.02% to 56.92% on average—a 10-point drop. Similar patterns are observed across the Qwen and Gemini model families, demonstrating that our KGGDG is consistently more challenging and provide a more rigorous evaluation of LLM medical QA capabilities.

Effectiveness of KG: To evaluate the impact of incorporating a knowledge graph (KG), we compare our approach against a baseline where the LLM directly generates dis-

Base model + Method	MedBullets↓	MedQA↓	Lacent↓	MedMCQA↓	MedXpert↓	NEJM↓	Avg. ↓
DeepSeek V3							
Options (Original)	70.21 (0.35)	85.04 (0.45)	72.70 (1.51)	76.88 (0.21)	**21.73** (0.57)	75.57 (0.46)	67.02
Options (Aug by LLM Directly)	**56.62** (1.04)	**65.18** (0.33)	64.68 (0.94)	68.64 (0.62)	37.59 (0.42)	61.43 (0.76)	59.02
Options (KGGDG)	59.59 (0.91)	70.03 (0.31)	**56.33** (0.75)	**66.15** (0.16)	30.09 (0.63)	**59.34** (0.30)	**56.92**
Qwen2.5-7B-Ins							
Options (Original)	46.24 (1.78)	60.12 (0.85)	56.74 (1.00)	56.54 (0.37)	**11.38** (0.80)	55.50 (1.32)	47.75
Options (Aug by LLM Directly)	36.87 (0.20)	45.95 (0.96)	46.48 (1.17)	51.20 (0.88)	24.19 (0.48)	43.34 (1.33)	41.34
Options (KGGDG)	**35.50** (0.52)	**44.87** (0.57)	**40.36** (1.58)	**45.24** (0.45)	14.36 (0.37)	**38.28** (0.82)	**36.43**
Qwen2.5-32B-Ins							
Options (Original)	59.02 (0.40)	71.41 (0.47)	63.85 (1.00)	61.84 (0.72)	**14.57** (0.20)	65.44 (0.63)	56.02
Options (Aug by LLM Directly)	51.26 (1.05)	**55.30** (0.00)	56.00 (0.38)	58.28 (0.53)	30.99 (0.74)	54.10 (1.05)	50.99
Options (KGGDG)	**46.69** (0.71)	57.15 (0.25)	**46.24** (2.02)	**54.10** (0.21)	21.21 (0.46)	**48.63** (1.48)	**45.67**
Qwen2.5-72B-Ins							
Options (Original)	67.47 (0.91)	77.58 (0.45)	66.59 (2.87)	69.81 (0.32)	**15.51** (0.33)	70.10 (0.20)	61.18
Options (Aug by LLM Directly)	**55.02** (1.38)	**58.99** (2.09)	60.46 (0.80)	63.57 (0.37)	31.84 (1.48)	55.09 (0.10)	54.16
Options (KGGDG)	55.14 (0.00)	62.60 (0.66)	**49.71** (1.14)	**58.79** (1.10)	21.06 (0.54)	**50.03** (0.20)	**49.56**
Gemini-1.5-Pro							
Options (Original)	71.12 (0.52)	82.38 (0.35)	66.42 (1.25)	74.49 (0.49)	**19.70** (0.18)	70.10 (0.71)	64.04
Options (Aug by LLM Directly)	**55.36** (0.52)	**63.51** (0.09)	58.81 (0.25)	66.67 (0.49)	35.20 (0.48)	55.32 (0.80)	55.81
Options (KGGDG)	58.90 (0.69)	67.64 (0.45)	**51.28** (1.15)	**61.41** (0.58)	28.54 (0.66)	**53.87** (1.05)	**53.61**
Gemini-2.5-Flash							
Options (Original)	76.60 (1.75)	88.55 (0.40)	74.20 (0.43)	78.67 (0.66)	**29.95** (0.61)	81.21 (1.24)	71.53
Options (Aug by LLM Directly)	65.07 (0.59)	**69.82** (0.37)	65.42 (0.51)	72.29 (0.48)	45.92 (0.13)	66.96 (0.56)	64.25
Options (KGGDG)	**64.96** (1.38)	74.38 (0.35)	**57.90** (1.37)	**67.37** (0.28)	36.77 (0.55)	**65.56** (0.79)	**61.16**

Table 1. Accuracy results (in %), averaged over 3 independent runs and rounded, across six datasets. We compare three settings: (1) original options, (2) augmented options generated directly by LLM, and (3) augmented options generated by our KGGDG pipeline. Rows highlighted in blue correspond to our knowledge-guided distractor generation method. Values in parentheses denote the sample standard deviation across the three runs.

tractors without KG guidance. The results, presented in Table 1, show that while LLMs alone can produce relatively challenging distractors, our KG-guided distractor generation pipeline, KGGDG, yields even more difficult ones. Specifically, we observe an average accuracy drop of around 5% for the Qwen model series and over 2% for both the Deepseek-V3 and Gemini model series. These results highlight the effectiveness of KG integration in increasing distractor difficulty.

KGGDG on hardest dataset: As shown in Table 1, for the most challenging dataset, MedXpert, KGGDG—which incorporates KG guidance into the generation process—consistently increases question difficulty compared to LLM-only augmentation, as indicated by an average accuracy drop of around 10%. However, we also observe that both LLM-based augmentation and KGGDG struggle to make these already difficult questions significantly harder than the original versions. This limitation may stem from the LLMs' insufficient understanding of complex medical content, but KG guidance helps mitigate this issue to some extent.

Model Robustness to the difficulty: As shown in Table 1, Gemini-2.5-Flash demonstrates the greatest robustness to difficulty augmentation, exhibiting the smallest performance drop—10% in absolute terms and a relative drop of 12% calculated as the decrease divided by its original accuracy. In contrast, Qwen2.5-7B-Inst shows the largest relative drop, with a 20% reduction relative to its original performance.

6. Conclusion

In this work, we present a knowledge-guided augmentation framework for enhancing clinical multiple-choice question datasets through the generation of challenging distractors. By leveraging biomedical knowledge graphs and semantic-guided walks, we extract structured reasoning paths that serve as misleading but clinically plausible cues. These paths inform prompt-based generation of distractors by LLMs, resulting in augmented questions that retain clinical coherence while significantly increasing task difficulty. Our empirical evaluation across six medical QA benchmarks and multiple LLMs demonstrates that the proposed augmentation pipeline consistently lowers model accuracy—revealing weaknesses in clinical reasoning that are obscured by existing benchmarks. As a result, our KGGDG delivers a more robust and diagnostic assessment, significantly enhancing the evaluation of clinical LLM reliability.

7. Limitations and Future Directions

While KGGDG successfully raises the difficulty level of medical QA benchmarks, its effectiveness depends on the quality and comprehensiveness of the underlying KG and the LLM used. Beyond generating the QA datasets for better LLM benchmarking, our future work includes two main directions: (1) exploring the integration of KGGDG into supervised fine-tuning and reinforcement learning frameworks to enhance LLM performance on medical QA tasks; and (2) analyzing the differences in LLM reasoning patterns when presented with original (simple) QA benchmark data versus our generated challenging distractor options, with the goal of further improving LLM reasoning ability in medical QA.

Acknowledgment

This project is supported by the CIFAR AI Chair Award, the Canada Research Chair Fellowship, NSERC, CIHR, Gemini Academic Program and IITP.

References

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., and Alhazmi, A. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *arXiv preprint arXiv:2402.01512*, 2024.

Balachandran, A. Medembed: Medical-focused embedding models, 2024. URL https://github.com/abhinand5/MedEmbed.

Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., and Demeester, T. Learning to reuse distractors to support multiple-choice question generation in education. *IEEE Transactions on Learning Technologies*, 17:375–390, 2022.

Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*, 2023. doi: https://doi.org/10.1038/s41597-023-01960-3. URL https://www.nature.com/articles/s41597-023-01960-3.

Chen, H., Fang, Z., Singla, Y., and Dredze, M. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024. URL https://arxiv.org/pdf/2402.18060.

Chen, H., Fang, Z., Singla, Y., and Dredze, M. 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)*, pp. 3563–3599, 2025.

Chen, Y., Wu, L., and Zaki, M. J. Toward subgraph-guided knowledge graph question generation with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Cui, H., Lu, J., Xu, R., Wang, S., Ma, W., Yu, Y., Yu, S., Kan, X., Ling, C., Zhao, L., Qin, Z. S., Ho, J. C., Fu, T., Ma, J., Huai, M., Wang, F., and Yang, C. A review on knowledge graphs for healthcare: Resources, applications, and promises, 2025. URL https://arxiv.org/abs/2306.04802.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Hill, J. and Simha, R. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 23–30, 2016.

Himmelstein, D. S. and Baranzini, S. E. Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *bioRxiv*, 2014. doi: 10.1101/011569. URL https://www.biorxiv.org/content/early/2014/12/11/011569.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37, July 2021. ISSN 1557-7341. doi: 10.1145/3447772. URL http://dx.doi.org/10.1145/3447772.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Lu, Y., Goi, S. Y., Zhao, X., and Wang, J. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications, 2025. URL https://arxiv.org/abs/2501.11632.

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

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

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., and Tang, S. Graph retrieval-augmented generation: A survey, 2024. URL https://arxiv.org/abs/2408.08921.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2024.

Taslimipoor, S., Benedetto, L., Felice, M., and Buttery, P. Distractor generation using generative and discriminative capabilities of transformer-based models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5052–5063, 2024.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Tran, A., Angelikas, K., Rama, E., Okechukwu, C., Smith, D. H., and MacNeil, S. Generating multiple choice questions for computing courses using large language models. In *2023 IEEE Frontiers in Education Conference (FIE)*, pp. 1–8. IEEE, 2023.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Wei, S.-L., Wu, C.-K., Huang, H.-H., and Chen, H.-H. Unveiling selection biases: Exploring order and token sensitivity in large language models. *arXiv preprint arXiv:2406.03009*, 2024.

Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.-I., et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., and Zhou, Y. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024.

Zesch, T. and Melamud, O. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 143–148, 2014.

Zhang, H., Liu, Z., Xiong, C., and Liu, Z. Grounded conversation generation as guided traverses in commonsense knowledge graphs, 2020. URL https://arxiv.org/abs/1911.02707.

Zhang, Y., Pan, F., Sui, X., Yu, K., Li, K., Tian, S., Erdengasileng, A., Han, Q., Wang, W., Wang, J., Wang, J., Sun, D., Chung, H., Zhou, J., Zhou, E., Lee, B., Qiu, X., Zhao, T., and Zhang, J. Biokg: a comprehensive, high-quality biomedical knowledge graph for ai-powered, data-driven biomedical research. *bioRxiv*, 2023. doi: 10.1101/2023.10.13.562216. URL https://www.biorxiv.org/content/early/2023/10/17/2023.10.13.562216.

Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., and Zhou, B. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A. Appendix

Prompt: QA Entity Extraction

```
QA_Extract_prompt = """
You are a helpful, pattern-following medical assistant.

Given both a clinical question and its correct answer, precisely extract all entities from each text
separately.

### Output Format
Strictly follow the JSON structure below.
The type of each entity MUST strictly belong to one of:
1.  gene/protein
2.  drug
3.  effect/phenotype
4.  disease
5.  biological_process
6.  molecular_function
7.  cellular_component
8.  exposure
9.  pathway
10.  anatomy


{
  "Question Entity": [
    {"id": "1", "type": "some_type", "name": "entity_name"},
    {"id": "2", "type": "some_type", "name": "entity_name"}
  ],
  "Answer Entity": [
    {"id": "1", "type": "some_type", "name": "entity_name"},
    {"id": "2", "type": "some_type", "name": "entity_name"}
  ]
}

### Example
Question:
A 72-year-old man presents to his primary care physician for a general checkup.  The patient works as a
farmer and has no concerns about his health.  He has a medical history of hypertension and obesity.  His
current medications include lisinopril and metoprolol.  His temperature is 99.5°F (37.5°C), blood pressure
is 177/108 mmHg, pulse is 90/min, respirations are 17/min, and oxygen saturation is 98% on room air.
Physical exam is notable for a murmur after S2 over the left sternal border.  The patient demonstrates a
stable gait and 5/5 strength in his upper and lower extremities.  Which of the following is another possible
finding in this patient?

Answer:
Femoral artery murmur

Output:


{
  "Question Entity": [
    {"id": "1", "type": "disease", "name": "hypertension"},
    {"id": "2", "type": "disease", "name": "obesity"},
    {"id": "3", "type": "drug", "name": "lisinopril"},
    {"id": "4", "type": "drug", "name": "metoprolol"},
    {"id": "5", "type": "effect/phenotype", "name": "murmur after S2"},
    {"id": "6", "type": "anatomy", "name": "left sternal border"},
    {"id": "7", "type": "anatomy", "name": "upper extremities"},
    {"id": "8", "type": "anatomy", "name": "lower extremities"}
  ],
  "Answer Entity": [
    {"id": "1", "type": "anatomy", "name": "Femoral artery"},
    {"id": "2", "type": "effect/phenotype", "name": "murmur"}
  ]
}

### Input
question:
{question}

answer:
{answer}

output:
"""
```

Prompt: Fallback Entity-To-Node Selection

```
Fallback_Selection_prompt = """ You are a helpful, pattern-following medical assistant.
Given a medical question and its answer, a query entity which is extracted from the question, and a list of
similar entities.
Select ONE most correlated entity from the list of similar entities based on the question and query entity.
SELECTED ENTITY MUST BE IN THE SIMILAR ENTITIES LIST. DO NOT invent or create any entity outside of the
given list.
IF there is not suitable entity in the similar entities, directly return the NONE.

### Output Format
Strictly follow the JSON structure below:


    {
        "selected_entity": {
            "name": "selected_entity_name",
            "id": a int number, the index of the selected entity in the similar entities list, from 0 to N-1,
            "reason": "reason for choosing this entity"
        }
    }


if there is no suitable entity, return:


    {
        "selected_entity": {
            "name": "NONE",
            "id": "NONE",
            "reason": "reason for not choosing any entity,
                       you need to explain why the entities in the similar entities list are not suitable"
        }
    }


### Input:
question: {question}
answer: {answer}
query entity: {query_entity}
similar entities: {similar_entities}

output:
"""
```

Prompt: Misleading Distractor Generation

```
misleading_distractor_prompt = """
You are a medical domain expert helping to design clinically challenging multiple-choice questions.

Your task is to generate 3 distractors for a given clinical question.  These distractors must be medically
incorrect options, but plausible enough to confuse even experienced clinicians.

You will be provided:
- A clinical question
- Its correct answer
- A set of reasoning paths derived from a biomedical knowledge graph

These reasoning paths may offer relevant associations (e.g., symptoms, treatments, conditions) that can
inspire clinically misleading distractors | but you are not required to use them directly.

---

IMPORTANT INSTRUCTIONS:
- This is a distractor generation task | not a selection task.
- Use the reasoning paths to inform and inspire distractors if they are helpful.
- If the paths are not helpful, rely entirely on your own medical knowledge to generate challenging
distractors.
- Every distractor must be clearly incorrect for the question but highly plausible (e.g., shares symptoms,
affects similar populations, is a common misconception, or is mechanistically related).
- Do not include any option that could be interpreted as correct or partially correct.

---

Input:
Question:  {input_question}
Correct Answer:  {correct_answer}
Reasoning Paths (if any):
{reasoning_paths}

---

Output Format:
Return exactly 3 distractors in strict JSON format, with a justification for why each one is misleading.


{
  "Distractors": ["Distractor1", "Distractor2", "Distractor3"],
  "Justifications": {
    "Distractor1": "Explain why this is a misleading but incorrect answer (e.g., symptom overlap,
                    treatment confusion, common misdiagnosis).",
    "Distractor2": "...",
    "Distractor3": "..."
  }
}

"""
```

Effect of Option Shuffling

Since LLMs can exhibit option bias (Wei et al., 2024), tending to favor certain choices as the correct answer, we further conduct an ablation study in this section to examine the effect of shuffling the answer options. Tables 2 and 3 show that shuffling answer choices has a negligible impact on accuracy tested with two models. For DeepSeek V3, the largest absolute change is only 0.32 percentage points (pp) with the original options, and just 0.14 pp when using our KGGDG. Qwen 2.5-7B-Inst shifts by at most 0.82 pp in the LLM-augmented setting and only 0.57 pp with KGGDG. Because every absolute difference is below 1 pp, answer-position randomisation does not meaningfully affect model performance. Moreover, KGGDG delivers the smallest shifts across both models, Showing its augmented data the most shuffle-invariant.

Base model + Method	MedBullets↓	MedQA↓	Lacent↓	MedMCQA↓	MedXpert↓	NEJM↓	Avg.↓
DeepSeek V3							
Options (Original)	68.61 (2.33)	84.80 (0.53)	73.53 (0.38)	76.56 (0.49)	**21.39** (0.67)	75.33 (0.27)	66.70
Options (Aug by LLM Directly)	**56.16** (0.69)	**65.00** (0.37)	66.42 (0.29)	67.84 (0.16)	37.49 (0.09)	62.42 (0.56)	59.22
Options (KGGDG)	59.25 (0.35)	69.68 (0.58)	**57.15** (1.37)	**65.54** (0.42)	30.83 (0.31)	**59.92** (0.27)	**57.06**
Qwen2.5-7B-Ins							
Options (Original)	46.58 (1.03)	59.39 (0.34)	56.33 (3.26)	56.12 (1.34)	**10.70** (0.37)	53.52 (0.56)	47.11
Options (Aug by LLM Directly)	38.70 (2.67)	**46.21** (1.15)	49.63 (0.86)	52.65 (0.99)	24.02 (0.23)	41.77 (1.11)	42.16
Options (KGGDG)	**36.99** (2.47)	46.27 (1.15)	**40.45** (1.63)	**45.52** (1.09)	14.34 (0.42)	**38.45** (1.58)	**37.00**

Table 2. ***Unshuffled*** **option order.** Accuracy results (in %), averaged over 3 independent runs and rounded, across six datasets. We compare three settings: (1) original options, (2) augmented options generated directly by LLM, and (3) augmented options generated by our KGGDG pipeline. ***Unshuffled*** means the position of the correct answer in the augmented options is preserved to match its original index in the raw dataset. Rows highlighted in blue correspond to our knowledge-guided distractor generation method. Values in parentheses denote the sample standard deviation across the three runs.

Base model + Method	Unshuffled↓	Shuffled↓	$\lvert\Delta\rvert$ ↓
DeepSeek V3			
Options (Original)	66.70	67.02	0.32
Options (Aug by LLM Directly)	59.22	59.02	0.20
Options (KGGDG)	57.06	56.92	0.14
Qwen2.5-7B-Ins			
Options (Original)	47.11	47.75	0.64
Options (Aug by LLM Directly)	42.16	41.34	0.82
Options (KGGDG)	37.00	36.43	0.57

Table 3. Aggregate accuracy (%, averaged across six datasets) before and after randomizing the answer-choice order. $\lvert\Delta\rvert$ is the *absolute* difference between the shuffled and unshuffled scores. All $\lvert\Delta\rvert$ are below 1 percentage point which indicate no systematic benefit or harm from shuffling. Rows highlighted in blue correspond to our knowledge-guided distractor generation pipeline. **The main paper uses *shuffled* KGGDG-augmented options as the default evaluation benchmark**.