MEC: Medical Evidence Capsules for Retrieval-Augmented Generation in Medical Multimodal Question Answering

Zhenghua Xu[1], Xinwei Dai[1], Bo Wang[2⋆], and Tian Tian[2]

[1] School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[2] School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
wangbo@hebut.edu.cn

Abstract. In medical multimodal question answering (QA), retrieval-augmented generation (RAG) serves as a core paradigm for improving model accuracy and mitigating hallucinations, yet existing approaches typically rely on a single source of evidence and thus struggle to support reliable and consistent reasoning. In practice, medical questions often require the integration of multiple evidence carriers, including narrative medical texts such as textbooks and clinical guidelines, structured medical knowledge systems, experiential priors implicitly encoded in expert models, and multimodal clinical examinations such as medical images and their associated reports. These evidence sources differ substantially in representation and granularity, making it difficult to accurately bind conclusions to traceable evidence snippets during retrieval, thereby increasing the risk of hallucinated answers. We propose Medical Evidence Capsules (MEC), an evidence organization mechanism for medical multimodal QA. MEC reformulates evidence from heterogeneous sources into compact capsule structures, enabling unified utilization of multiple evidence sources. We further construct a medical multimodal retrieval corpus that integrates both textual and imaging evidence, and evaluate MEC on five medical QA benchmarks. Experimental results demonstrate that MEC consistently yields performance improvements across both text-only QA and visual question answering (VQA) benchmarks. The resources of our work are available at https://github.com/fruitbitlab/MEC.

Keywords: Medical Question Answering · Retrieval-Augmented Generation · Large Language Models · Vision-Language Models

1 Introduction

Large language models (LLMs) and Large vision models (LVMs) have achieved significant progress in medical question answering (QA) [22, 19], and retrieval-augmented generation (RAG) has been widely adopted to improve accuracy

⋆ Corresponding author.

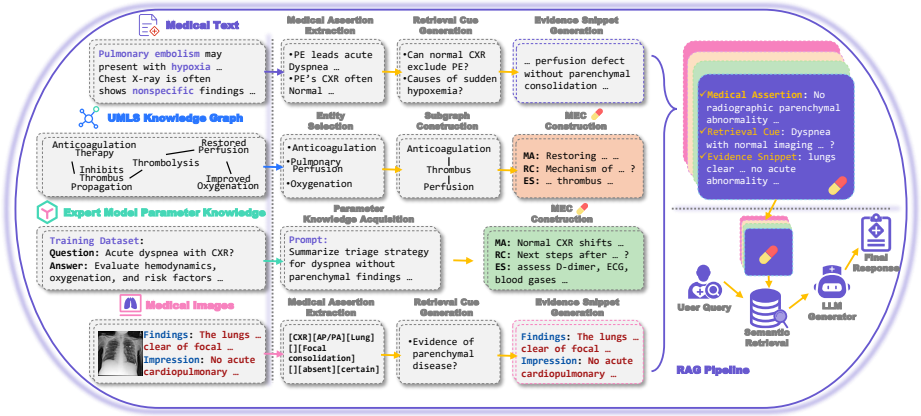

Fig. 1. Overview of Medical Evidence Capsules. Heterogeneous evidence sources are normalized into capsules to support unified retrieval and constrained generation.

and mitigate hallucinations in high-risk scenarios [26]. However, in medical QA, the core challenge lies in whether retrieved evidence can be reliably used. If conclusions cannot form stable correspondences with explicit evidence, even linguistically plausible outputs may introduce hallucinations due to mismatched evidence granularity or insufficient alignment [25].

In medical multimodal QA, RAG introduces external medical knowledge to constrain the generation process, yet existing methods handle multi-source evidence in a relatively coarse manner. Most approaches retrieve textual, structured, and imaging information separately and directly concatenate them as input [23], assuming that the model can perform knowledge selection and alignment on its own. However, this often leads to mismatched evidence granularity [21, 24] and incomparable similarity signals, making it difficult to stably bind generated conclusions to citable evidence snippets, thereby weakening answer consistency and traceability [18].

The key bottleneck of generative models in the medical domain is not the lack of information, but how to establish controllable usage relationships among multi-source evidence [15]. Medical questions often involve textual resources, structured knowledge, parameterized experience from models, and imaging materials, which differ substantially in granularity and representation [9]. Existing approaches typically concatenate these sources as independent contexts [17] without explicitly constraining how evidence should be used [21], forcing the model to rely on implicit alignment during generation and thereby increasing the risk of hallucinations.

Based on these issues, we refocus the problem from "how much evidence to retrieve" to "how retrieved evidence is used," and propose Medical Evidence Capsules (MEC), an evidence organization mechanism for medical multimodal QA. MEC restructures information from heterogeneous sources into compact evidence capsules, each explicitly comprising a medical assertion, a retrieval cue,

and a citable evidence snippet. With this structure, textual resources, parameterized experience from models, and imaging evidence can be retrieved and invoked collaboratively under a unified interface, so that retrieval returns structured evidence components that can directly support reasoning rather than loosely concatenated contexts, thereby improving answer consistency and traceability.

Building upon this framework, we construct a large-scale medical multimodal retrieval corpus that integrates medical texts, structured knowledge, parameterized experience from models, as well as imaging data and corresponding reports. We evaluate MEC on text-only QA and visual QA benchmarks, where it consistently demonstrates stable improvements across different settings. In summary, our contributions are as follows: (1) we propose MEC, which explicitly constrains evidence usage through a compact capsule structure, providing a unified retrievable evidence representation for medical multimodal QA; (2) we build a retrieval corpus containing over 550,000 medical assertions, offering reusable resources and evaluation foundations for medical multimodal QA.

2 Methods

2.1 Overview of Medical Evidence Capsule

In this section, we introduce the overall design of Medical Evidence Capsules (MEC). MEC is an evidence organization and invocation structure for medical multimodal QA, designed to support reliable utilization of multi-source evidence in RAG. With MEC, the retrieval stage returns structured evidence capsules rather than loosely concatenated contexts, providing clearer evidence constraints for the generation stage.

2.2 Evidence from Medical Texts

Medical textbooks, clinical guidelines, and review articles provide important knowledge sources for medical QA, yet their long-form narratives are difficult to directly leverage in RAG. To improve the usability of textual evidence, we build assertion-centric textual evidence capsules, enabling textual knowledge to be reliably retrieved and directly used for reasoning.

We first extract semantically coherent text segments from the textbook corpus, forming the collection $\mathcal{T} = \{t_i\}$. For each segment, we employ a LLM to distill its core medical assertion [12]:

$$a_i = \mathcal{F}_{\text{assert}}(t_i). \tag{1}$$

Medical assertions compactly capture the key facts or judgments expressed in the text, reducing semantic ambiguity during generation. Based on each assertion, we further construct a paired evidence snippet and retrieval cue, which provide necessary context and characterize typical invocation patterns:

$$(e_i, c_i) = \mathcal{F}_{\text{support}}(t_i, a_i). \tag{2}$$

Finally, each textual evidence item is organized as a capsule consisting of the assertion a_i, the retrieval cue c_i, and the evidence snippet e_i. This design ensures that retrieval returns structured evidence components rather than lengthy raw contexts, providing clearer evidence constraints for generation.

2.3 Evidence from Structured Resources

Structured medical knowledge provides explicit entity definitions and relational constraints, serving as an important evidence source beyond narrative texts. We build graph-based evidence capsules from the UMLS [2], transforming medical entities and their relations into retrievable and citable structured evidence.

We first select entities with unique concept identifiers and well-defined semantic types from UMLS to form an entity set $\mathcal{V}$. To avoid isolated nodes and overly redundant hub entities, we choose evidence anchors by jointly considering structural importance and semantic specificity. For an entity $u \in \mathcal{V}$, its structural importance is defined as:

$$c(u) = \alpha \log(1 + \deg(u)) + (1 - \alpha)\,\mathrm{PR}(u). \tag{3}$$

where $\deg(u)$ denotes node degree, $\mathrm{PR}(u)$ is the PageRank score, and α balances local connectivity and global centrality. Meanwhile, we quantify medical specificity using the semantic type $s(u)$ via the information content:

$$i(u) = -\log \frac{|\{v \in \mathcal{V} : s(v) = s(u)\}|}{|\mathcal{V}|}. \tag{4}$$

which prioritizes more discriminative entities in the medical knowledge system. Around each selected anchor, we construct a scale-controlled local subgraph $\mathcal{G}_u$ and employ a LLM to convert the structured relations into an medical evidence capsule:

$$(a, e, c) = \mathcal{F}_{\mathrm{graph}}(\mathcal{G}_u). \tag{5}$$

where a is the medical assertion, e is the evidence snippet describing the relations, and c is the retrieval cue. This process organizes graph knowledge into MEC, enabling structured evidence to participate in multi-source reasoning within a unified framework.

2.4 Evidence from Model Parameters

In addition to explicit texts and structured resources, expert-level language models also encode substantial medical experiential priors in their parameters [11]. To avoid the uncontrolled inference that may arise from directly invoking such implicit knowledge, we convert parametric knowledge into retrievable and constrained MEC evidence capsules.

Specifically, given QA pairs (q, a) from the training split of datasets, we generate explanatory evidence under explicit answer-alignment constraints:

$$e = \mathcal{F}_{\mathrm{param}}(q, a). \tag{6}$$

Table 1. Comparison of different RAG methods, retrieval return fields, and LLM scales on Text QA. Bold values indicate the best performance across models and datasets.

Retrieval Evidence	Method	MecQA			Med-FlashCards			HfMedQA		
		Qw-3B	Qw-7B	Lm-8B	Qw-3B	Qw-7B	Lm-8B	Qw-3B	Qw-7B	Lm-8B
—	Wo-RAG	0.2550	0.2653	0.2384	0.1470	0.1473	0.1854	0.1285	0.1274	0.1770
Original Text	SR-RAG	0.3682	0.3899	0.3212	0.3310	0.3499	0.3381	0.2834	0.3015	0.2654
	FLARE	0.3416	0.3508	0.3356	0.2839	0.3007	0.3224	0.2370	0.2418	0.2344
	DRAGIN	0.3578	0.3588	0.3122	0.3144	0.3135	0.3122	0.3092	0.3014	0.2709
Evidence Snippet	SR-RAG	0.3905	0.3989	0.3330	0.3348	0.3435	0.3312	0.2845	0.3067	0.2676
	FLARE	0.3492	0.3526	0.3372	0.2827	0.2982	0.3255	0.2371	0.2424	0.2340
	DRAGIN	0.3598	0.3607	0.3374	0.3154	0.3118	0.3149	0.3125	0.3003	0.2720
MEC (Ours)	SR-RAG	**0.3982**	**0.4489**	0.3580	**0.4034**	**0.4100**	**0.3952**	0.3113	0.3493	0.2800
	FLARE	0.3517	0.3653	0.3531	0.3164	0.3252	0.3568	0.2516	0.2756	0.2570
	DRAGIN	0.3684	0.3895	**0.3593**	0.3434	0.3522	0.3380	**0.3305**	**0.3709**	**0.2962**

This process enforces consistency with the known answer and avoids unconstrained expansions beyond the question context, thereby reducing the risk of hallucinations. Building on the generated evidence, we further distill a corresponding medical assertion and construct a matching retrieval cue, so that parametric knowledge can be incorporated into the MEC framework. Through this conversion, implicit experience encoded in model parameters is made explicit as citable evidence, enabling it to be jointly used with textual, graph-based, and imaging evidence under a unified structure.

2.5 Evidence from Medical Images

Medical images and their reports provide the most direct objective evidence for medical QA. However, their descriptions are highly structured and dense with negation and uncertainty, making them difficult to use on the same semantic scale as textual or graph-based knowledge. To address this, we reorganize imaging information into MEC-compliant imaging evidence capsules, enabling unified retrieval and constrained generation-time reasoning.

We use approximately 7,000 image–report pairs from MIMIC-CXR [8]. From each report, we extract medical assertions [6] and use the *findings* section as the citable evidence snippet, preserving the observational context without altering the original semantics. To improve retrievability and consistency, we normalize each imaging assertion into a fixed eight-tuple:

$$a_{\text{img}} = (m, p, o, l, f, s, e, u), \tag{7}$$

where m denotes the imaging modality, p the acquisition position, o the organ or structure, l the spatial region or laterality, f the core finding, s the severity or morphological descriptor, e the existence polarity, and u the uncertainty

Table 2. Comparison of different retrieval return fields, and VLM scales on Visual QA. Bold values indicate the best performance across models and datasets.

Datasets	Retrieval Evidence	Models			
		Qw-VL-3B	**Qw-VL-7B**	**MedG-4B**	**Lshu-7B**
SLAKE	Wo-RAG	0.5541	0.5811	0.4838	0.7865
	Findings	0.5608	0.5864	0.4905	0.7838
	Image+Findings	0.5716	0.5932	0.5014	0.7919
	MEC (Ours)	**0.5838**	**0.6027**	**0.5108**	**0.7986**
VQA-RAD	Wo-RAG	0.4971	0.5714	0.5886	0.6000
	Findings	0.5029	0.5771	0.5886	0.5943
	Image+Findings	0.5086	0.5829	0.5909	0.6057
	MEC (Ours)	**0.5257**	**0.5952**	**0.6000**	**0.6114**

Table 3. Tasks, models, datasets, and evaluation metrics used in experiments.

Task Type	Model	Datasets	Evaluation Metric
Text QA	Qwen2.5-3B-Instruct [28] Qwen2.5-7B-Instruct [28] Llama-3.1-8B-Instruct [3]	MecQA (In-domain) Med-FlashCards [4] HfMedQA	ROUGE-L [13]
Visual QA	Qwen2.5-VL-3B-Instruct [1] medgemma-1.5-4b-it [16] Qwen2.5-VL-7B-Instruct [1] Lingshu-7B [27]	VQA-RAD (CXR) [10] SLAKE (CXR) [14]	Accuracy

level. This representation covers both positive and negative findings and explicitly distinguishes confirmed observations from speculative conclusions, thereby reducing implicit inference during generation.

Based on these normalized imaging assertions, we generate corresponding retrieval cues and organize the assertion, retrieval cue, and the report's *findings* and *impression* into an imaging evidence capsule.

3 Experiments and Results

3.1 Experimental Setups

Datasets and Settings In our experiments, we construct a MEC retrieval corpus for medical multimodal QA, supporting both text-only QA and visual QA. The textual and structured evidence is built from medical textbook corpora aligned with MedQA [7] and UMLS [2], comprising approximately 500,000 medical assertions. The imaging evidence is derived from MIMIC-CXR [8], where about 50,000 imaging-related assertions are extracted from roughly 7,000 image–report pairs. Text-only QA uses the in-domain MecQA (sampled from our corpus) and cross-domain Med-FlashCards [4] and HfMedQA [3]; VQA uses VQA-

[3] https://huggingface.co/datasets/asmitha26/medical-data

RAD [10] and SLAKE [14]. Models and metrics are in Table 3. GPT-4o [5] is used to construct all evidence capsules.

Baselines We adopt several representative text-based RAG methods [11, 20] as baselines and integrate MEC into these frameworks under identical generation models and parameter settings to evaluate its evidence representation capability. For visual QA, we compare against no-retrieval settings and analyze the contribution of imaging evidence capsules through different MEC component combinations. All experiments are conducted with consistent model configurations and generation settings to ensure fair comparison.

3.2 Main Results Analysis

Text QA results are shown in Table 1. MEC consistently improves performance across different retrieval strategies and model scales, achieving the best results on all benchmarks. By organizing evidence into assertion-driven structured forms, MEC aligns retrieved content with QA semantics and reduces the implicit alignment burden during generation. This is particularly evident on medium-scale models, suggesting that structured evidence enhances multi-source knowledge efficiency. Visual QA results are in Table 2. When compared to no-retrieval or direct report text usage, MEC's organization of imaging assertions yields more stable accuracy improvements, with consistent trends across different VLMs. By structuring imaging data with explicit polarity and uncertainty markers, MEC enables seamless collaboration between imaging evidence and evidence from other sources, ensuring robust benefits across models.

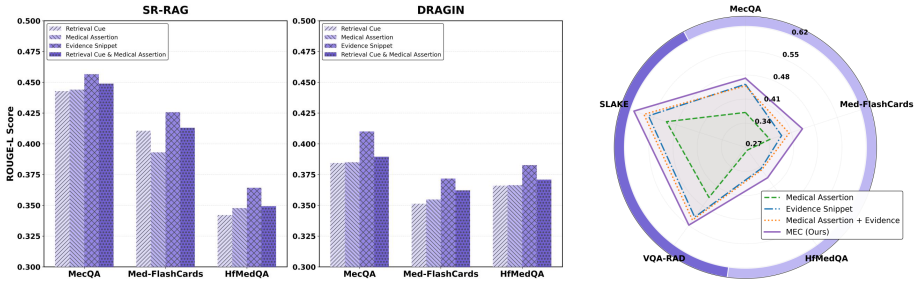

Fig. 2. Ablations on matching fields (bar) and return contents (radar). MEC performs best across text QA and medical VQA, with a better efficiency–accuracy trade-off.

3.3 Ablations and Case Discussion

Effect of Retrieval and Return Fields As shown in Fig. 2, we analyze the effect of MEC from the matching fields (bar chart) and the returned contents

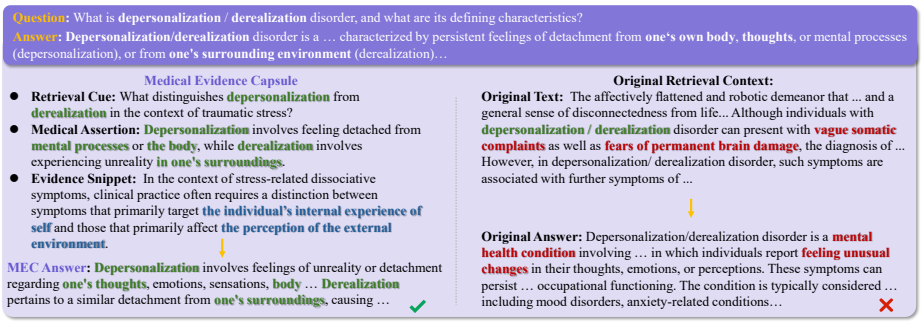

Fig. 3. Case study comparing Original retrieval with MEC retrieval. MEC provides a clearer reasoning scaffold, yielding more focused and traceable answers.

(radar chart). Using evidence snippets as the matching signal is typically the best, while matching only with medical assertions and retrieval cues yields only a slight performance drop. Moreover, since assertions and cues are more compact, this strategy substantially reduces the computational cost of vectorization and matching. Under a fixed matching strategy, the full MEC achieves the best results on all five benchmarks, outperforming returning only assertions or only evidence snippets. These results indicate that by jointly providing structured assertions, retrieval cues, and citable evidence snippets, MEC imposes more complete semantic and evidential constraints during generation, leading to a more robust trade-off between efficiency and effectiveness.

Case Study We further illustrate the practical effect of MEC through a case study, as shown in Fig. 3. Compared with retrieval based on raw text, the evidence returned under the MEC framework provides a clearer semantic scaffold during generation. The medical assertion directly captures the core discriminative information required by the question, the retrieval cue strengthens semantic alignment between the query and the evidence, and the evidence snippet supplies necessary contextual support and conditional constraints, thereby reducing the need for implicit filtering and alignment within lengthy contexts. In contrast, raw text retrieval often returns content that is descriptively rich but less task-relevant, increasing the reasoning burden and weakening the focus of the generated answer.

4 Conclusion

This paper proposes Medical Evidence Capsules (MEC), an evidence organization mechanism for medical multimodal QA, which explicitly constrain evidence retrieval and invocation through a compact structure, mitigating evidence granularity mismatch and improving the traceability of conclusions in medical multimodal QA. We further construct a large-scale retrieval corpus integrating

multi-source medical knowledge and evaluate MEC on medical text QA and visual QA benchmarks. Experimental results demonstrate that MEC consistently improves performance across different models and retrieval settings, validating its effectiveness in medical multimodal RAG.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62276089, by the Natural Science Foundation of Tianjin Municipality, China, under Grant 24JCJQJC00200, by the Natural Science Foundation of Hebei Province, China, under Grant F2024202064, and by the Ministry of Human Resources and Social Security, China, under Grant RSTH-2023-135-1.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
4. Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)
5. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
6. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
7. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. arXiv preprint arXiv:2009.13081 (2020)
8. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
9. Karim, A.R., Uzuner, O.: Masonnlp at mediqa-wv 2025: Multimodal retrieval-augmented generation with large language models for medical vqa. In: Proceedings of the 7th Clinical Natural Language Processing Workshop. pp. 84–94 (2025)
10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)

11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
12. Li, L., Zhou, X., Zhang, Y., Wu, X.: From retrieval to generation: Unifying external and parametric knowledge for medical question answering. arXiv preprint arXiv:2510.18297 (2025)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
15. Ning, Y., Sun, Y., Luo, L., Wang, Y., Pan, Y., Lin, H.: Medtrust-rag: Evidence verification and trust alignment for biomedical question answering. arXiv preprint arXiv:2510.14400 (2025)
16. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
17. Shaaban, M.A., Zarei, M.R., Khan, A., Akkasi, A., Yaqub, M., Komeili, M.: Towards multimodal retrieval-augmented generation for medical visual question answering. Research Square (2025). https://doi.org/10.21203/rs.3.rs-7752202/v1, preprint
18. Shang, Y., Liu, Y., Wang, H., Li, F., Sun, W., Chengyu, W., Zheng, Y.: Medusa: Cross-modal transferable adversarial attacks on multimodal medical retrieval-augmented generation. arXiv preprint arXiv:2511.19257 (2025)
19. Sohn, J., Park, Y., Yoon, C., Park, S., Hwang, H., Sung, M., Kim, H., Kang, J.: 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). pp. 12739–12753 (2025)
20. Su, W., Tang, Y., Ai, Q., Wu, Z., Liu, Y.: Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12991–13013 (2024)
21. Sun, L., Zhao, J.J., Han, W., Xiong, C.: Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. 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. 643–655 (2025)
22. Wang, C., Chen, Y.: Evaluating large language models for evidence-based clinical question answering. arXiv preprint arXiv:2509.10843 (2025)
23. Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., Jin, Y., Grau, V.: Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 28443–28467 (2025)
24. Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., Yao, H.: Rule: Reliable multimodal rag for factuality in medical vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1093 (2024)

25. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233–6251 (2024)
26. Xu, S., Pang, L., Yu, M., Meng, F., Shen, H., Cheng, X., Zhou, J.: Unsupervised information refinement training of large language models for retrieval-augmented generation. arXiv preprint arXiv:2402.18150 (2024)
27. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
28. 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., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)