

# Evaluating Domain-Agnostic Retrieval-Augmented Generation for Gastrointestinal Medical VQA

Tobias Abraham Haider, Poj Netsiri, Tuvshin Selenge, Felix Kapfer and Elina Messeritsch

## Abstract

Medical visual question answering (MedVQA) for gastrointestinal imaging requires accurate visual reasoning and clinically consistent answer generation. This project investigates whether standard, domain-agnostic Retrieval-Augmented Generation (RAG) methodologies can effectively improve performance on highly specialized medical tasks. Using a fine-tuned Qwen2.5-VL model as the foundation, we propose and evaluate RAG pipelines that do not rely on complex medical ontologies. Our primary experiments evaluate BM25 lexical retrieval and a multimodal hybrid search strategy (BM25 + ColQwen) on the ImageCLEFmed-MEDVQA-GI 2026 dataset. We also assess the impact of neural reranking and evidence formatting during system development. Our results demonstrate that off-the-shelf RAG components improve generative consistency over standard non-RAG baselines. While bespoke, heavily fine-tuned models hold an edge in certain lexical metrics, general-purpose retrieval strategies achieve highly competitive performance, offering a lightweight alternative to extensive domain-specific engineering.

## 1. Introduction

Medical visual question answering (MedVQA) for gastrointestinal imaging requires accurate visual reasoning, medical terminology understanding, and clinically consistent answer generation from endoscopic images and textual questions. This work investigates whether standard, domain-agnostic Retrieval-Augmented Generation (RAG) methodologies can effectively improve performance on highly specialized medical tasks without the need for complex medical ontologies.

Specifically, we evaluate these methods within the context of the ImageCLEFmed-MEDVQA-GI 2026 challenge [1, 2]. Using a fine-tuned Qwen2.5-VL model as the foundation, we propose and evaluate multiple RAG pipelines using BM25 lexical retrieval [3], neural reranking [4], and multimodal hybrid search strategies. Our approach establishes that general-purpose retrieval strategies can yield competitive improvements in clinical VQA, providing a practical alternative to designing entirely bespoke models for distinct medical sub-tasks.

## 2. Related Work

**Medical Visual Question Answering in Endoscopy.** Recent years have seen rapid development in MedVQA, largely driven by the shift from traditional classification algorithms to generative, open-ended vision-language models [5]. In the specialized domain of gastrointestinal endoscopy, datasets like Kvasir-VQA [6] and benchmarks such as EndoBench [7] have established critical baselines for evaluating model reasoning and spatial localization. Despite these advances, adapting generalized vision-language models to consistently output strict, domain-specific terminology without hallucination remains a significant challenge.

**Retrieval-Augmented Generation (RAG) for Medical Multimodality.** To mitigate factual inaccuracies and hallucinations inherent in standalone generation approaches, RAG has emerged as a standard solution [8]. Recent frameworks increasingly leverage multimodal RAG architectures, combining text-based retrieval with visual embeddings to ground AI responses in external, factual knowledge

*CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany*

✉ e11833743@student.tuwien.ac.at (T. A. Haider); e12402153@student.tuwien.ac.at (P. Netsiri);

e11815791@student.tuwien.ac.at (T. Selenge); e12429669@student.tuwien.ac.at (F. Kapfer); e12022014@student.tuwien.ac.at (E. Messeritsch)

ORCID 0009-0007-3722-3717 (T. A. Haider)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

bases. Studies demonstrate that augmenting VLMs with retrieved medical reports significantly improves factual consistency over non-retrieval baselines [8, 5]. Our work builds upon this foundation by investigating whether computationally lightweight, domain-agnostic RAG strategies can scale to specialized queries without relying on bespoke medical ontologies.

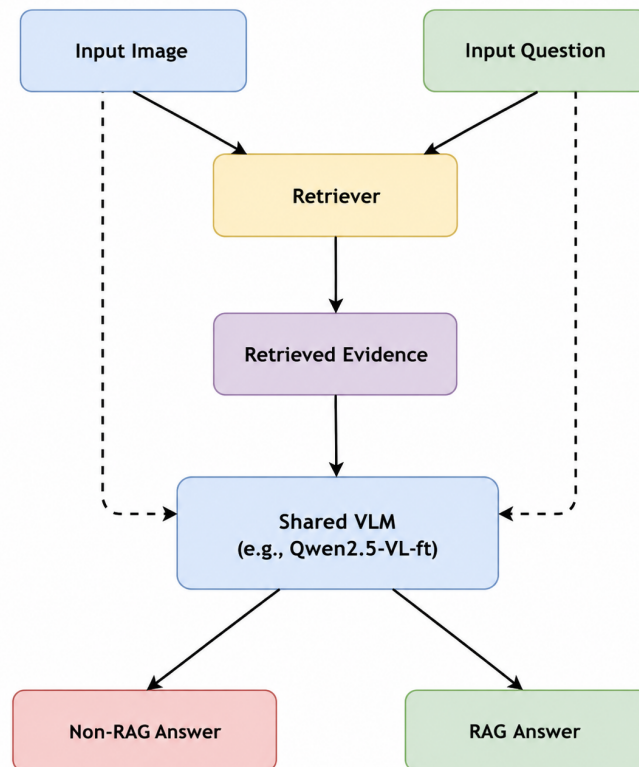
### 3. Base System Design

The proposed system is designed as a multimodal MedVQA pipeline combining a fine-tuned Qwen2.5-VL model with a lightweight RAG module. Given an input image and clinical question, the system retrieves semantically relevant evidence from the dataset corpus using BM25 retrieval. The retrieved evidence, image, and question are then provided to the VLM for answer generation.

To bridge the gap between open-ended VLM generation and strict dataset evaluation metrics, we implemented a rule-based normalization layer. This layer uses regular expressions and keyword matching to strip conversational prefixes (e.g., "The answer is:") and standardize outputs. Specifically:

- **Binary Logic:** Infers Yes/No responses by scanning for negative markers (e.g., "absent") or positive markers (e.g., "identified").
- **Procedures:** Maps localized anatomical terms to broader procedure labels (e.g., mapping "cecum" to "colonoscopy").
- **Formatting:** Standardizes color names and merges multiple spatial locations using semicolons to match dataset conventions.

The overall system design and inference pipeline are illustrated in Figure 1.



**Figure 1:** System design for the RAG and Non-RAG VQA pipelines.

## 4. Experimental Setup

To identify the optimal configuration for the final competition submission, we conducted a series of targeted internal experiments to evaluate prompting strategies, evidence formatting, and retrieval techniques.

### 4.1. Non-RAG Baseline and Prompt Tuning

Directing large language models to consistently output the strict, dataset-style labels required in medical domains is notoriously difficult. To establish a baseline, we evaluated generic models (LLaVA-1.5-7B) against our domain-adapted adapter (Qwen2.5-VL-7B-Instruct fine-tuned on Kvasir-VQA-x1). Through iterative testing, we determined that explicit system prompting was essential. We tested concise, label-only, and medically contextualized prompts. The domain-adapted Qwen model utilizing a medically contextualized prompt (You are a clinical assistant for gastrointestinal endoscopy images. Use only the image and answer concisely with medically relevant wording.) consistently produced answers closest to the target distribution and was selected as the foundation for the RAG integration.

### 4.2. Evidence Format Representation

We investigated how different evidence representations influence answer generation within the RAG pipeline. The retrieved BM25 evidence was formatted in three ways: (1) a baseline question-answer (QA) format, (2) a natural language sentence format, and (3) a structured JSON format. Internal validation showed that the natural language sentence format provided the highest stability and token overlap, slightly outperforming the JSON and QA structures. Consequently, the sentence format was adopted for the final pipeline.

### 4.3. Neural Reranking (BGE)

Neural reranking is an industry standard used to refine initial lexical search results. We explored incorporating a cross-encoder (BAAI/bge-reranker-base) [4] to rescore candidate snippets. However, this setup was not sufficiently optimized or prompted for the strict dataset-style labels required by the challenge. Because the reranker was misaligned with the final submission targets, it severely underperformed in early testing and was excluded from our official RAG pipeline.

### 4.4. Hybrid RAG (BM25 + ColQwen)

Standard visual retrieval often relies on global image embeddings, which struggle when queries target highly localized features such as surgical instruments or specific polyps. To address this, we implemented a hybrid retrieval strategy fusing lexical similarity (BM25) with fine-grained visual similarity using ColQwen patch-wise image embeddings [9, 10].

For each question aspect, BM25 retrieved candidate records from the text corpus, while the most similar visual patches provided a localized visual score. These score streams were min-max normalized and fused using a weighted sum.

The fusion weights (0.2 for text, 0.8 for visual) were determined empirically through qualitative observation rather than systematic ablation. We manually inspected retrieved results across different weight distributions to select the final parameters for the competition. A formal ablation study detailing the influence of visual retrieval on specific question aspects remains future work.

## 5. Results and Discussion

We evaluated our finalized RAG and Hybrid RAG pipelines against the ImageCLEFmed-MEDVQA-GI 2026 test set. The dataset-provided answers were treated as ground truth, and performance was

measured using the official competition metrics: BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR.

Table 1 details the performance of our submitted systems compared to the top external baselines on the public leaderboard. In preliminary internal validation, the RAG pipeline consistently outperformed the standalone Non-RAG baseline in token overlap and precision, which informed our decision to exclusively submit the RAG variants to the official competition.

**Table 1**

ImageCLEFmed-MEDVQA-GI 2026 Public Validation Scores. Our proposed methods are denoted as "Ours". Top external baseline submissions from the [competition leaderboard](#) are included for comparison.

| Method                               | BLEU        | ROUGE-1     | ROUGE-2     | ROUGE-L     | METEOR      |
|--------------------------------------|-------------|-------------|-------------|-------------|-------------|
| Proposed RAG (Ours - cn224)          | <b>0.46</b> | 0.84        | 0.09        | 0.84        | 0.45        |
| Proposed Hybrid RAG (Ours - tobsel7) | 0.29        | 0.84        | 0.10        | 0.84        | 0.46        |
| External Baseline (saademaad)        | 0.43        | <b>0.89</b> | 0.13        | <b>0.89</b> | 0.50        |
| External Baseline (nhattan9999t)     | 0.24        | 0.81        | 0.10        | 0.80        | 0.44        |
| External Baseline (krissTewari)      | 0.14        | 0.40        | <b>0.27</b> | 0.37        | <b>0.51</b> |

The public leaderboard results reflect the reality that bespoke, heavily fine-tuned models often capture specific generation metrics better than generalized models. While our standard RAG approach achieved the highest BLEU score (0.46), external baseline submissions outperformed our methods in ROUGE and METEOR. Specifically, the external submission saademaad achieved the highest ROUGE-1 and ROUGE-L scores, while krissTewari achieved the highest ROUGE-2 and METEOR scores.

This indicates that while our general-purpose RAG pipeline provides a strong foundation and excels at exact n-gram overlap, highly tailored domain-specific models or custom generation strategies may capture semantic nuance more effectively. The Hybrid RAG approach maintained high ROUGE-1 and ROUGE-L scores but saw a drop in BLEU, suggesting that the integration of visual patch embeddings may have introduced variations that penalized exact string matching.

## 6. Conclusion

This work investigated domain-agnostic Retrieval-Augmented Generation for gastrointestinal Med-VQA. Building, retraining, and custom-prompting individual models for every emerging clinical task is resource-intensive and often unfeasible. By applying industry-standard retrieval methodologies [3, 9], we demonstrate that RAG consistently improves answer consistency over standard zero-shot baselines. While highly specialized models still dominate certain leaderboard metrics, these general-purpose approaches represent a practical, scalable path forward for developing clinical tools without the dependency on deep domain-specific engineering and costly retraining cycles.

## Declaration of AI Use

The core scientific contributions, experimental design, data analysis, and original content of this paper were produced entirely by the authors. Large language models were used for this paper as writing assistants to harmonize stylistic consistency and improve the fluency of the text. All AI-generated edits were thoroughly reviewed, verified, and approved by the authors prior to submission.

## References

- [1] S. A. Hicks, S. Gautam, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2026 – medical visual question answering for gastrointestinal tract, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Berlin, Germany, 2026.

- [2] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [3] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, volume 4, Now Publishers Inc, 2009.
- [4] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-pack: Packed resources for general chinese embeddings, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 641–649. URL: <https://doi.org/10.1145/3626772.3657878>. doi:10.1145/3626772.3657878.
- [5] M. B. Noshin, M. Dutta, M. N. Kaysar, R. H. Sajib, M. J. Hossen, D. Nandi, A. A. Jubair, M. Rahman, Medical visual question answering with multimodal: a systematic mini review (2023–2026), Frontiers in Digital Health Volume 8 - 2026 (2026). URL: <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2026.1848710>. doi:10.3389/fdgth.2026.1848710.
- [6] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy, arXiv preprint arXiv:2506.09958 (2025). URL: <https://arxiv.org/abs/2506.09958>.
- [7] S. Liu, B. Zheng, W. Chen, Z. Peng, Z. Yin, J. Shao, J. Hu, Y. Yuan, Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis, 2025. URL: <https://arxiv.org/abs/2505.23601>. arXiv:2505.23601.
- [8] A. H. M. R. Karim, O. Uzuner, Multimodal retrieval-augmented generation with large language models for medical vqa, 2025. URL: <https://arxiv.org/abs/2510.13856>. arXiv:2510.13856.
- [9] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, P. Colombo, Colpali: Efficient document retrieval with vision language models, 2025. URL: <https://arxiv.org/abs/2407.01449>. arXiv:2407.01449.
- [10] P. Straßer, B. Trent, Colpali & elasticsearch: How to search complex documents, <https://www.elastic.co/search-labs/blog/elasticsearch-colpali-document-search>, 2025.