# Iterative Multimodal Retrieval-Augmented Generation for Medical Question Answering

**Xupeng Chen** [*1] **Binbin Shi** [2] **Chenqian Le** [1] **Jiaqi Zhang** [3] **Kewen Wang** [4] **Ran Gong** [1] **Jinhan Zhang** [1] **Chihang Wang** [1]

## Abstract

Medical retrieval-augmented generation (RAG) systems typically operate on text chunks extracted from biomedical literature, discarding the rich visual content (tables, figures, structured layouts) of original document pages. We propose MEDVRAG, an iterative multimodal RAG framework that retrieves and reasons over PMC document page images instead of OCR'd text. The system pairs ColQwen2.5 patch-level page embeddings with a sharded MapReduce LLM filter, scaling to ∼350K pages while keeping Stage-1 retrieval under 30 ms via an offline coarse-to-fine index ($C$=8 centroids per page, ANN over centroids, exact two-way scoring on the top-$R$ shortlist). A vision-language model (VLM) then iteratively refines its query and accumulates evidence in a memory bank across ≤3 reasoning rounds, with a single iteration costing ∼15.9 s and the full three-round pipeline ∼47.8 s on 4×A100. Across four medical QA benchmarks (MedQA, MedMCQA, PubMedQA, MMLU-Med), MEDVRAG reaches 78.6% average accuracy. Under controlled comparison with the same Qwen2.5-VL-32B backbone, retrieval contributes a +5.8-point gain over the no-retrieval baseline; we also note a +1.8-point edge over MedRAG + GPT-4 (76.8%), with the caveat that this is a cross-paper rather than head-to-head comparison. Ablations isolate +1.0 from page-image vs. text-chunk retrieval, +1.5 from iteration, and +1.0 from the memory bank.

[1]New York University, New York, USA [2]Tsinghua University, Beijing, China [3]Independent Researcher [4]Chinese Academy of Sciences, Beijing, China. Correspondence to: Xupeng Chen <xc1490@nyu.edu>.

*Accepted to the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMM-QA)*, Seoul, South Korea. 

## 1. Introduction

Knowledge-intensive multimodal question answering—answering questions whose evidence lives in document pages with mixed text, tables, and figures—stresses both retrieval and reasoning under non-trivial compute budgets. Medical QA is a particularly demanding instance: the underlying biomedical literature is large (∼350K relevant PMC pages in our setup), pages routinely encode dosage tables and diagnostic flowcharts that linearize poorly, and many questions require synthesizing evidence from multiple sources, motivating an iterative retrieve-reason loop rather than a single pass. We use medical multiple-choice QA as the testbed but design the system around the broader goal of *efficient* multimodal knowledge-intensive QA. Large language models (LLMs) achieve strong accuracy on medical licensing exams—GPT-4-base 86.1% on MedQA under 5-shot prompting (Nori et al., 2023), Med-PaLM 2 86.5% on MedQA from a MedQA-specific instruction-finetuned variant (Singhal et al., 2025)—but remain limited by static parametric knowledge and hallucination (Ji et al., 2023); clinical decision support (Singhal et al., 2023) would require both grounded evidence and bounded latency. RAG (Lewis et al., 2020) addresses the grounding side by retrieving from external corpora; in medicine, MedRAG (Xiong et al., 2024), Self-BioRAG (Jeong et al., 2024), i-MedRAG (Xiong et al., 2025), and RAG2 (Sohn et al., 2025) have demonstrated substantial gains.

The medical RAG baselines we compare against, however, share one assumption: retrieval over text. Biomedical PDFs are OCR'd or parsed into chunks, and only the extracted text is indexed. This systematically discards visual content—dosage tables, diagnostic flowcharts, anatomical diagrams—that frequently conveys information more effectively than its linearized form. Single-pass retrieval is also limited for multi-hop questions where mechanism, contraindications, and monitoring guidelines reside on distinct pages.

We propose **MEDVRAG** (Medical Visual RAG), which addresses both issues: (i) **page-level multimodal retrieval** indexes original 350K PMC document pages with a vision-language embedding model rather than text chunks; (ii) **it-**

**erative reasoning with memory** lets a VLM autonomously refine its query and accumulate findings across up to three rounds. To our knowledge MEDVRAG is among the first multimodal RAG frameworks for biomedical multiple-choice QA operating on PMC page images with iterative VLM reasoning—general-domain page-image RAG (Faysse et al., 2024) targets document VQA, and the medical RAG systems above operate on text.

**Contributions.** (1) A two-stage retrieval pipeline combining ColQwen2.5 patch-level page embeddings with a sharded MapReduce LLM filter that scales to 350K biomedical pages at ∼30 ms Stage-1 latency through an offline coarse-to-fine index ($C$=8 centroids/page, ANN over centroids, exact two-way scoring on the top-$R$ shortlist). (2) An iterative VLM reasoner with a structured memory bank that issues query refinements and accumulates per-round findings across up to three rounds; per-iteration latency ∼15.9 s on 4×A100. (3) A controlled head-to-head comparison against same-backbone text-only RAG, no-RAG, and progressively-ablated variants on four standard medical QA benchmarks, with per-stage latency, per-iteration accuracy, and a four-class error taxonomy.

## 2. Related Work

**Medical RAG.** MedRAG (Xiong et al., 2024) systematically benchmarks RAG on five corpora and four retrievers; Self-BioRAG (Jeong et al., 2024) adds self-reflection to decide when to retrieve; i-MedRAG (Xiong et al., 2025) issues iterative follow-up queries; RAG2 (Sohn et al., 2025) adds rationale-guided retrieval. All operate on text chunks.

**Multimodal page-image retrieval.** ColPali (Faysse et al., 2024) introduced multi-vector page-image retrieval over a vision-language backbone, with later variants (ColQwen2.5) replacing the encoder. Concurrent general-domain page-image RAG efforts explore document-VQA-style retrieval, but to our knowledge none targets medical benchmarks or couples retrieval with iterative reasoning over a memory bank.

**Vision-language models for medical content.** LLaVA-Med (Li et al., 2023) and Med-Flamingo (Moor et al., 2023) are biomedical VLMs trained for visual QA over images such as radiographs; Qwen2.5-VL (Bai et al., 2025) demonstrates strong general document and layout understanding that we leverage for biomedical pages without further fine-tuning.

## 3. Method

**System overview.** MEDVRAG consists of four components (Fig. 1): (1) a page-level embedding and indexing module over the PMC corpus; (2) a two-stage retrieval pipeline (embedding similarity, then LLM filtering); (3) an iterative reasoning engine driven by a VLM; and (4) a memory bank that accumulates findings across rounds. We describe each component below.

**Knowledge base.** We render each PMC OA article PDF page-by-page at 300 DPI; near-duplicate pages are dropped (cosine $> 0.97$). The final corpus is ∼350K pages from ∼18K articles. Text overlap with each benchmark's source articles is excluded by DOI/PMID; for MMLU-Med (no source DOIs) we additionally check verbatim question text appears in zero retained pages.

**Page-level embedding.** Each page is encoded with the ColPali multi-vector page-image retriever (Faysse et al., 2024)—specifically the ColQwen2.5 community variant that swaps PaliGemma for a Qwen2.5-VL backbone—into $n$ patch-level vectors. We PCA-reduce these to $d$=128 (from 768) for index size; a short text summary (∼120 tokens) is also generated offline by Qwen2.5-VL-7B for use in Stage 2.

**Stage 1 (embedding retrieval).** Given a question $q$ encoded into $m$ token vectors, we score each page by a two-way late-interaction sum

$$S(q,p) = \tfrac{1}{m}\sum_i \max_j q_i^\top v_j \;+\; \tfrac{1}{n}\sum_j \max_i q_i^\top v_j$$

and return the top $N_1$=2,000 pages. A coarse-to-fine strategy ($C$=8 centroids per page, ANN over the centroid index, then exact two-way scoring on the top candidates) keeps Stage 1 latency under 30 ms.

**Stage 2 (LLM filtering).** Since $N_1 \times \bar{\ell}_{\text{sum}} \approx 240$K tokens exceeds the 32K context, we use sharded MapReduce over Qwen3-30B-A3B: 8 parallel map calls each score $B$=256 summaries (∼31K tokens) and emit top-25; one reduce call ranks the surviving 200 into the final $N_2$=100.

**Iterative reasoning.** Per iteration $t \in \{1,2,3\}$, the VLM (Qwen2.5-VL-32B-Instruct) receives the question + options, the highest-ranked **10 page images** from the $N_2$=100 filtered set at native dynamic resolution, the top-20 page summaries, and the memory bank. It either emits an answer (`<answer>...</answer>`) or a refined query plus running notes; refined queries trigger fresh Stage-1+Stage-2 retrieval. The memory bank stores `iteration`, per-round `key_findings`, and `reasoning_history`; verbatim prompts and JSON schema are in Appendix A.

## 4. Results

**Setup.** We evaluate on the MIRAGE (Xiong et al., 2024) splits: MedQA (Jin et al., 2021) 1,273 USMLE questions, MedMCQA (Pal et al., 2022) 4,183, PubMedQA (Jin et al., 2019) 500 expert-labeled test, MMLU-Medical (Hendrycks et al., 2021) 1,089 across six clinical subdomains. All models are evaluated zero-shot, regex-extracting the option

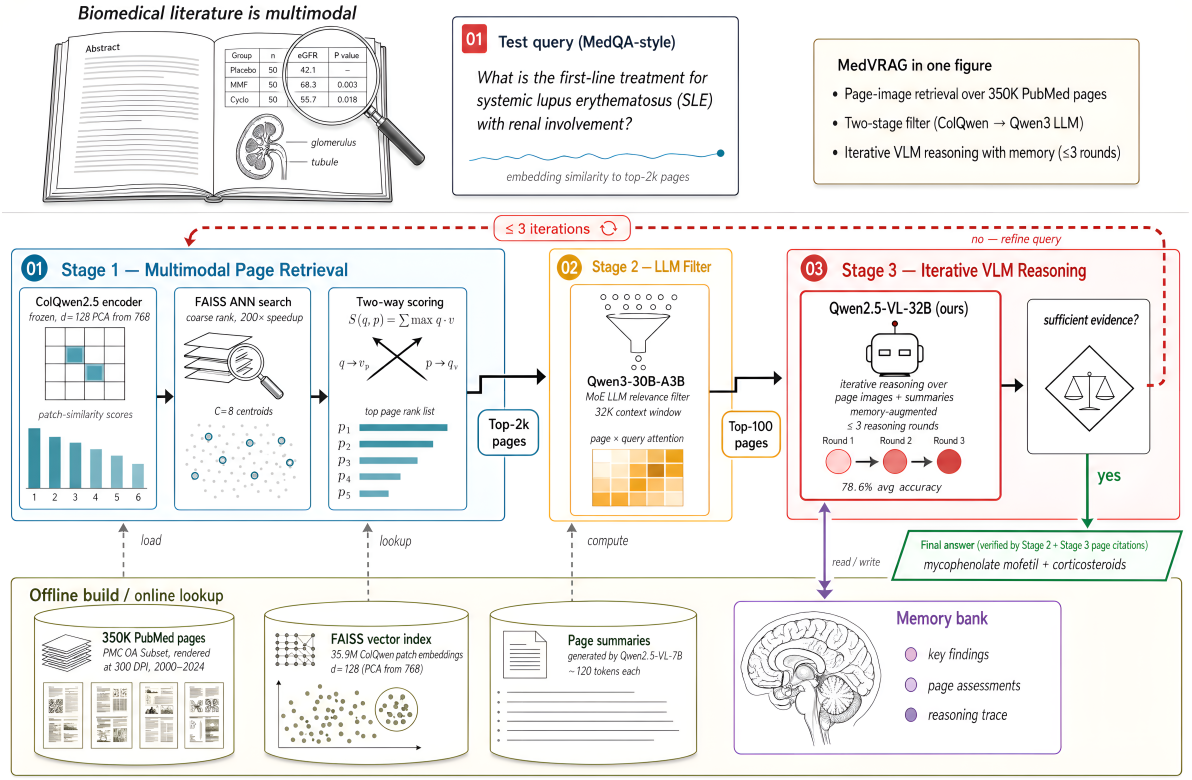


*Figure 1.* MEDVRAG pipeline. A medical question is encoded by ColQwen2.5 and matched against a FAISS index over ∼350K PMC pages; an LLM filters candidates by summary (sharded MapReduce); a VLM iteratively reasons over the highest-ranked 10 page images plus top-20 summaries from the $N_2$=100 filtered set, refining queries and accumulating findings in a memory bank across $\leq 3$ rounds.

letter; unparseable outputs count as incorrect. Hardware: 4×NVIDIA A100 80GB.

Table 1 reports our main results. MEDVRAG (full) achieves 78.6% average accuracy. Under controlled comparison with the same Qwen2.5-VL-32B backbone, it improves over no-retrieval by +5.8 points on average, with positive gains on every dataset (largest: MedQA +8.0, MedMCQA +6.8). Even the single-round variant (76.1%) approaches MedRAG + GPT-4 (76.8%) using a substantially smaller, open-weight backbone. Among no-RAG baselines, Qwen2.5-VL-32B (72.8%) is comparable to text-only Qwen3-30B (74.8%), so the VLM backbone does not compromise text reasoning. We caution that cross-paper comparisons against published † baselines use different prompts, temperatures, and context windows; the most internally controlled comparison is the additive ablation sequence inside Table 1, in which we toggle one component at a time while reusing the same backbone, prompts, decoding settings, and evaluation protocol.

**Ablation (rows in additive sequence).** Each row toggles exactly one component vs. the row above. Page images vs. OCR text chunks (text-only → single-round, both with the Stage-2 filter) yields +1.0. Iteration adds +1.5. The memory bank adds +1.0, for a total +5.8 over no RAG. Note that the modality gain is a single comparison against one BGE-large text retriever; stronger configurations (E5-Mistral, cross-encoder rerankers, Nougat-style layout-aware parsing) may close the gap, and we leave that to future work.

**Iteration efficiency.** Across datasets, 61.4% of questions are answered in one round, 24.8% in two, and 13.8% use all three rounds (Fig. 2). Round 1 questions are easier (mean R1 accuracy 81.0%); later-round questions are harder but still benefit from iteration (mean R2 74.5%, R3 70.2%; Fig. 3). Table 2 reports per-stage latency: a single-iteration query takes ∼15.9 s and three iterations ∼47.8 s, dominated by the VLM call (∼12.1 s/iter); the embedding-based retrieval stage stays under 30 ms thanks to the coarse-to-fine strategy.

**Accuracy–compute tradeoff.** On MedQA, Round 2 recovers ∼7.6% of initial errors but Round 3 only ∼2.5%, so the marginal accuracy per second drops sharply between rounds 2 and 3. Treating the iteration cap as a deployment knob rather than a fixed value, an early-stopping policy of *stop after Round 2 (or earlier if the VLM emits an answer)* captures roughly $7.6/(7.6{+}2.5) \approx 75\%$ of the recoverable errors that the full 3-round pipeline would correct, at ∼67% of the worst-case wall time (∼31.7 s vs. ∼47.8 s). We report

*Table 1.* Accuracy (%) on medical QA benchmarks. Bold marks full MEDVRAG; not a global best (Med-PaLM 2 wins MedQA/PubMedQA on a matched 3-dataset average 80.2 vs. 75.3). † = published results, taken directly from the original publications. * HuatuoGPT-o1 reports 63.1% on MMLU-Pro (Medical), not the MIRAGE MMLU-Med used here; we omit the column to avoid a non-comparable cell. Our controlled baselines (Qwen3-30B, Qwen3-8B, and the MEDVRAG ablation rows) are run under identical zero-shot conditions, prompt template, and answer-extraction procedure.

| Model | MedQA | MedMCQA | PubMedQA | MMLU-Med | Avg. |
|---|---|---|---|---|---|
| *No-RAG baselines* | | | | | |
| GPT-3.5† (Xiong et al., 2024) | 60.2 | 62.7 | 78.2 | — | — |
| GPT-4-base 5-shot† (Nori et al., 2023) | 86.1 | 73.7 | 77.4 | — | — |
| Med-PaLM 2† (Singhal et al., 2025) | 86.5 | 72.3 | 81.8 | — | — |
| HuatuoGPT-o1-8B† (Chen et al., 2024) | 72.6 | 60.4 | 79.2 | —* | — |
| Meditron-70B† (Chen et al., 2023) | ∼70 | ∼66 | 81.6 | 77.6 | 73.8 |
| Yi-34B (OpenMedLM)† (Maharjan et al., 2024) | 72.6 | 68.3 | 77.3 | 81.7 | 75.0 |
| Qwen3-30B (Yang et al., 2025) | 71.5 | 65.1 | 77.7 | 85.0 | 74.8 |
| Qwen3-8B (Yang et al., 2025) | 63.5 | 59.3 | 74.7 | 78.8 | 69.1 |
| *Text-only RAG* | | | | | |
| MedRAG + GPT-3.5† | 66.6 | 58.0 | 67.4 | 75.5 | 66.9 |
| MedRAG + GPT-4† | 82.8 | 66.7 | 70.6 | 87.2 | 76.8 |
| Self-BioRAG-7B† | 43.6 | 42.1 | — | 53.9 | — |
| MEDVRAG *(ours; rows in additive sequence)* | | | | | |
| – no RAG | 71.4 | 62.4 | 73.7 | 83.7 | 72.8 |
| – text-only retrieval | 74.8 | 66.2 | 73.3 | 86.2 | 75.1 |
| – single-round (page images) | 75.8 | 67.0 | 74.9 | 86.8 | 76.1 |
| – + iterative reasoning (no memory) | 77.6 | 67.5 | 77.0 | 88.4 | 77.6 |
| **– + memory bank (full system)** | **79.4** | **69.2** | **77.2** | **88.6** | **78.6** |

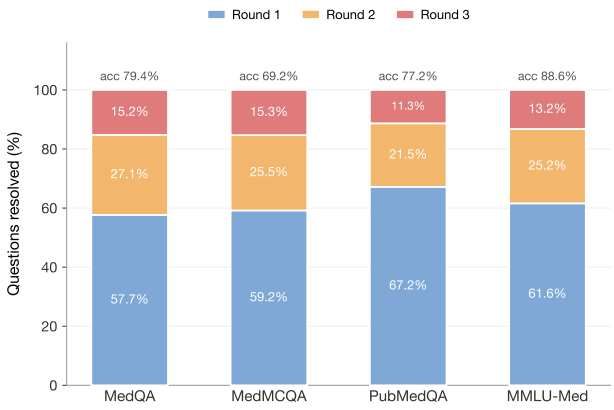


*Figure 2.* Iteration distribution by dataset. Most questions (57.7–67.2%, mean 61.4%) are resolved in Round 1; 11.3–15.3% (mean 13.8%) use all three rounds.

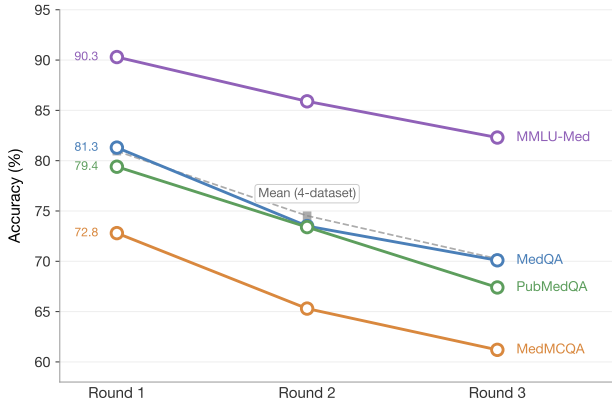


*Figure 3.* Accuracy conditioned on iteration count, per dataset. The 72.9% multi-iteration figure is the within-dataset weighted average over R2+R3.

the 3-iteration cap throughout the rest of the paper for a like-for-like comparison with prior multi-round RAG systems, and leave a learned policy that conditions on calibrated confidence to future work.

**Retrieval quality (MedQA).** Using a Qwen3-30B-A3B judge with 100 first-author-spot-checked labels (kappa $\approx$ 0.81), Evidence Recall@$N_2$=100 is 72.4% and Recall@$N_1$=2000 is 89.4%. A summary-supported metric (judge sees the cited page summary, not the image) reports 85.3% of correct answers as supported; we caution this is summary-grounded and uses a same-family judge, so it is not a verified visual-grounding rate.

**Error analysis.** MEDVRAG (full system) errors on MedQA fall into four mutually exclusive primary causes (single-rater): (i) *retrieval miss* (∼35.9%) — no relevant page in the top-100; (ii) *filter miss* (∼19.7%) — relevant page in top-2000 dropped by Stage-2; (iii) *VLM misinterpretation* (∼29.5%) — relevant page retrieved but reasoned incorrectly; (iv) *iteration drift* (∼14.9%) — refined query moves off-topic.

*Table 2.* End-to-end latency per query on 4×A100 (wall-clock, mean of 50 MedQA queries). Stage-1 retrieval and Stage-2 filter both repeat at every iteration when the VLM emits a refined query.

| Stage | Time |
|---|---|
| ANN search (CPU) | ∼1.9 ms |
| Coarse ranking (CPU) | ∼4.6 ms |
| Fine-grained scoring (GPU) | ∼20.9 ms |
| LLM filter (Qwen3-30B-A3B, sharded) | ∼3.8 s |
| VLM reasoning per iteration | ∼12.1 s |
| **Total (1 iteration)** | **∼15.9 s** |
| **Total (3 iter., 3× filter+VLM)** | **∼47.8 s** |

**Case study (paraphrased).** *Question*: A 45-year-old woman with SLE (joint pain, butterfly rash, elevated anti-dsDNA) is started on hydroxychloroquine. Which side effect is most important to monitor? *Round 1* retrieved 10 hydroxychloroquine pharmacology pages covering indications and dosing but not toxicity monitoring; the VLM emitted the refined query `"hydroxychloroquine retinal toxicity ophthalmologic screening guidelines"`. *Round 2* retrieved 10 ophthalmology pages including an AAO-style monitoring-schedule table; the VLM selected **Retinal toxicity**, citing that page. Full trace with page IDs and image crops will be released with the code.

## 5. Discussion and Limitations

**Why page-level retrieval helps.** The ablation gain decomposes into three mechanisms. (i) *Tabular data*: medical pages routinely encode dosage tables, diagnostic criteria, and reference ranges; text-only RAG must linearize these and frequently loses row–column structure. (ii) *Figures and diagrams*: in a manual sample of 100 MedQA questions where MEDVRAG answers correctly and text-only does not, we classified 20 as figure/diagram-dependent, 39 as table-dependent, and 41 as text-with-layout (single-rater, mutually exclusive primary cause). (iii) *Structural context*: section headings and spatial proximity provide cues even on prose-only pages. Per-dataset modality gains (Table 1, text-only → single-round) are: PubMedQA +1.6, MedQA +1.0, MedMCQA +0.8, MMLU-Med +0.6 (avg. +1.0). MMLU-Med shows the smallest effect, consistent with its short, fact-recall-style questions that rarely require layout cues; PubMedQA's larger gain is perhaps surprising given its single-abstract evidence design and may reflect that page-level rendering preserves abstract structure (headings, conclusion sentences) more faithfully than 512-token chunking.

**Why iteration helps.** 38.6% of questions trigger more than one round (Fig. 2); these are typically multi-hop scenarios that require synthesizing mechanism, contraindications, and monitoring schedules from distinct pages, mirroring how clinicians refine literature searches. The per-iteration accuracy curve (Fig. 3) shows that round-2 and round-3 questions are intrinsically harder—a forced single-round re-run on this subset would isolate the iteration contribution from question difficulty and is left to future work.

**When text-only RAG hurts.** Xiong et al. (2024) report MedRAG + GPT-4 *degrades* MedQA-US (84.0→82.8) and MMLU-Med (89.4→87.2) under conflicting text-chunk context; in our own ablation, text-only retrieval drops PubMedQA from 73.7 to 73.3, again consistent with single-source questions where extra chunks introduce noise. MEDVRAG avoids regression on every dataset, suggesting that page-level evidence plus LLM filtering is more robust to context conflict than naive text-chunk RAG—at least against the BGE-large baseline used here.

**Limitations.** All numbers in Table 1 are from a single seeded run with low-temperature decoding; without bootstrap CIs, multi-seed runs, or significance tests, small deltas (+0.2 on PubMedQA/MMLU-Med from memory, +1.0 modality, +1.0 memory, +1.8 cross-paper) should be read as suggestive rather than statistically validated. The text-only baseline uses BGE-large embeddings on 512-token OCR'd chunks; stronger retrievers (E5-Mistral, cross-encoder rerankers, layout-aware parsing) are not evaluated. The corpus is PMC-only and the evaluation is multiple-choice; the system is *not* clinically validated, and free-text generation, calibrated uncertainty, harmful-answer analysis, and clinician-in-the-loop review are all required before deployment. The system's two design properties most relevant to future deployment are (i) retrieved evidence remains in original page-image form and is directly inspectable, and (ii) the memory bank produces a per-round audit trail; both are necessary, not sufficient.

**Conclusion.** MEDVRAG reaches 78.6% average across four standard medical QA benchmarks and improves over the same-backbone no-RAG baseline by +5.8 points by combining ColQwen2.5 page-image retrieval, Stage-2 LLM filtering, and iterative VLM reasoning with a memory bank. The +1.0 modality gain (over a single text baseline) and +1.0 memory gain are suggestive but small; statistical validation, stronger text baselines, and free-text/clinical evaluation are all open work. Future work will explore expanding the knowledge base to include clinical guidelines and textbooks, reducing inference latency through selective iteration, and evaluating on open-ended clinical QA tasks beyond multiple-choice benchmarks.

## Data and Code Availability

The PubMed page corpus is built from the PMC Open Access Subset (`https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`). All evaluation bench-

marks are public: MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU-Med (Hendrycks et al., 2021). Verbatim prompts, the memory-bank JSON schema, the answer-extraction regex, all decoding settings, and the BGE-large text-baseline configuration are inlined in Appendix A–B, so the system is fully specified at submission time. Source code, FAISS indices, page summaries, and reproduction scripts will be released under MIT alongside the camera-ready version; an anonymized code+config snapshot is available to reviewers during review on request through the workshop program chairs.

## References


Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., and Wang, B. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024.

Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., and Colombo, P. ColPali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

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. In *Proceedings of EMNLP-IJCNLP*, pp. 2567–2577, 2019.

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. In *Advances in Neural Information Processing Systems*, pp. 9459–9474, 2020.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, 2023.

Maharjan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M. G., Barnes, G., Thapa, R., Mao, Q., and Das, R. OpenMedLM: Prompt engineering can outperform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14:14156, 2024.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. Med-Flamingo: A multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 2023.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

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 (CHIL)*, 2022.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Sohn, J., Park, Y., Yoon, C., Park, S., Hwang, H., Sung, M., Kim, H., and Kang, J. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 12739–12753, 2025. doi: 10.18653/v1/2025.naacl-long.635. URL https://aclanthology.org/2025.naacl-long.635/.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL*, pp. 6233–6251, 2024.

Xiong, G., Jin, Q., Wang, X., Zhang, M., Lu, Z., and Zhang, A. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

## A. Prompts and Output Schema

**Stage-2 filter prompt** (Qwen3-30B-A3B, sharded MapReduce; `target_k`=25 in map shards, 100 in reduce):

> *You are a medical document retrieval expert. Given a medical question and candidate page summaries, select the* {`target_k`} *most relevant pages. Question:* {`question`}*. Candidate page summaries:* {`summaries`}*. Select the* {`target_k`} *most relevant pages by listing their numbers in order of relevance (most relevant first). Output ONLY the page numbers inside* `<selected_pages>` *tags.*

**VLM reasoning prompt** (Qwen2.5-VL-32B-Instruct):

> *You are a medical QA expert. Answer the multiple-choice question based on the provided document pages. Question:* {`question`}*. Options:* {`options`}*.* {`memory_section`}*. Retrieved page summaries:* {`summaries`}*. (The actual page images are also provided for your reference.) Instructions: If you have enough information to answer, output your answer inside* `<answer>` *tags with a brief justification. If you need more information, output a refined search query inside* `<query_update>` *tags and summarize your current findings inside* `<notes>` *tags. This is iteration* {`iteration`}/{`max_iterations`}*.* {`force_msg`}*.*

**Memory schema (JSON):** {`iteration: int, key_findings: list[str (each prefixed [Round k])], reasoning_history: list[{iteration, notes}]`}.

**Decoding:** VLM reasoning $T$=0.1, max_new_tokens = 2048; Stage-2 filter $T$=0, max_new_tokens = 1024; page-summary (offline) $T$=0.

**Answer extraction (regex):** `<answer>\s*([A-Da-d]|yes|no|maybe)\b`; outputs without `<answer>` tags count as incorrect.

## B. Implementation Details

**Models.** ColQwen2.5 ($d$=128 after PCA from 768) for page embedding; Qwen2.5-VL-7B for offline page summaries (chosen over the 32B variant for indexing throughput; mean summary length ∼120 tokens); Qwen3-30B-A3B for Stage-2 filter; Qwen2.5-VL-32B-Instruct for iterative reasoning.

**Retrieval cutoffs.** $N_1$=2,000 candidates from Stage 1, $N_2$=100 pages after the Stage-2 LLM filter, max 3 iterations (hard cap, enforced regardless of VLM decision).

**Hardware.** 4×NVIDIA A100 80GB. **Random seed:** 42 for the single reported run. **Image preprocessing:** Qwen2.5-VL native dynamic-resolution preprocessor with min/max pixels = $256/1280 \cdot 28^2$. **Decoding:** VLM reasoning $T$=0.1, max_new_tokens = 2048; Stage-2 filter $T$=0, max_new_tokens = 1024; offline page summary $T$=0.

**Knowledge-base construction.** The PMC Open Access Subset is restricted to articles published between 2000 and 2024. Each PDF is rendered page-by-page at 300 DPI to PNG. Near-duplicate pages are removed by ColQwen2.5 cosine-similarity thresholding at 0.97. The final corpus contains ∼350K pages from ∼18K articles. The FAISS index over PCA-reduced patch vectors holds ∼35.9M patch-level entries (avg. ∼103 patches per page). Leakage check: articles overlapping with MedQA/MedMCQA/PubMedQA source articles are excluded by DOI/PMID; for MMLU-Med (no source DOIs) we additionally verified zero verbatim question text matches in retained pages.

**Coarse-to-fine retrieval.** Stage 1's two-way late-interaction sum is computed in two passes to avoid the naive $O(m \cdot n \cdot N)$ cost. (i) *Offline*: each page's $n$ patches are k-means-clustered into $C$=8 centroids; centroids and a page-id pointer are stored in a FAISS ANN index. (ii) *Online*: the $m$ query token vectors are searched against the centroid index for top-$R$ candidate pages ($R \ll N$); the exact two-way score is then computed only on those $R$ candidates with their full $n$ patches. The asymptotic cost drops to $O(m \cdot C \cdot N + m \cdot n \cdot R)$, which in practice runs in <30 ms on CPU+GPU.

**Text-only retrieval baseline (Table 1 "text-only retrieval" row).** Pages are OCR'd, chunked into 512-token segments with 64-token stride, embedded with BGE-large-en-v1.5 and indexed in a separate FAISS index. Stage 1 returns the top-2,000 chunks by cosine similarity; Stage 2 (same Qwen3-30B-A3B sharded MapReduce filter) selects $N_2$=100 chunks; the VLM then receives chunks-as-text rather than page images. All other settings (iteration count, prompts, decoding) are held identical to the full system, isolating the modality effect.

**Memory-bank schema.** Stored as a JSON object with three fields:

- `iteration`: integer round counter (1, 2, or 3).
- `key_findings`: list of medical facts and relationships extracted during reasoning, each prefixed `[Round k]` so the VLM can attribute the round-of-origin in later iterations.
- `reasoning_history`: list of {`iteration`, `notes`} pairs giving a per-round summary of the reasoning trajectory.

The bank serves two purposes: (i) it prevents information loss across iterations as the context grows, and (ii) it provides a structured, auditable record of which evidence was consulted when, useful for retrospective inspection.

**Dataset details.** MedQA: 1,273 USMLE 4-option questions covering clinical medicine, pharmacology, pathology. MedMCQA: 4,183 validation questions from AIIMS/NEET, 21 medical subjects. PubMedQA: 500 expert-labeled test questions following the MIRAGE split (yes/no/maybe). MMLU-Medical: 1,089 questions across six subdomains—Clinical Knowledge, Medical Genetics, Anatomy, Professional Medicine, College Biology, College Medicine.

**Per-dataset iteration distribution** (numbers behind Fig. 2): MedQA {R1 57.7%, R2 27.1%, R3 15.2%}; MedMCQA {R1 59.2%, R2 25.5%, R3 15.3%}; PubMedQA {R1 67.2%, R2 21.5%, R3 11.3%}; MMLU-Med {R1 61.6%, R2 25.2%, R3 13.2%}.

**Per-iteration accuracy** (numbers behind Fig. 3, equal-weight 4-dataset mean): R1 81.0%, R2 74.5%, R3 70.2%. The within-dataset weighted average over R2+R3 is 72.9% (the multi-iteration figure quoted in the main text).

**Code release.** Source code, FAISS indices, page summaries, and reproduction scripts will be released under MIT alongside the camera-ready version. An anonymized code+config snapshot is available to reviewers during review on request via the workshop program chairs.