INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

ADAPTIVE GRAPH AND MULTIMODAL RETRIEVAL FOR EVIDENCE-BASED MEDICAL QUESTION ANSWERING

Samar Abbas[*1], Muhammad Iqrar Aslam[2], Abdul Qayoom[*3], Waleed Ahmad[4], Summera Fazal[5]

*[*1]School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China*
[2]Department of Computer Science, University of Agriculture Faisalabad
*[*3]Faculty of Computing and Engineering, University of Layyah, Layyah, Punjab, Pakistan*
[4]Department of computer science and technology, Mianyang Teachers College Sichuan province Mianyang, China
[5]Department: Statistics, University: Govt. College Women University Faisalabad

[*1]sammarhashmi@gmail.com, [*3]aqbuzdar@ul.edu.pk

DOI: https://doi.org/10.5281/zenodo.24184854

Keywords
Medical question answering; Retrieval-augmented generation; Knowledge graphs; Multimodal retrieval; Adaptive evidence fusion.

Article History
Received: 08 August 2026
Accepted: 07 November 2026
Published: 13 November 2026

Copyright @Author
Corresponding Author: *
Samar Abbas
Abdul Qayoom

Abstract

Medical question-answering systems must retrieve relevant information, reconcile heterogeneous evidence and make the basis of their answers traceable. Text-centred retrieval-augmented generation (RAG) systems can overlook relationships encoded in medical knowledge graphs and information presented in guideline figures. We propose Adaptive Graph and Multimodal Retrieval (AGMR-Med), a conversational framework that integrates unstructured text, structured medical knowledge and visual evidence. The framework combines a domain-aware query encoder, an adaptive network that weights retrieval scores across modalities, and a multi-level evidence aggregator with claim-level citation verification. Evaluation using MedQA-USMLE and a curated multimodal guideline corpus yielded 91.7% answer accuracy, 88.4% citation fidelity and a mean response time of 7.6 s. These values exceeded the accuracy and citation fidelity, and reduced the latency, of the evaluated Vitruvius/MedSearch and GraphRAG baselines. Ablation experiments indicated that adaptive weighting and multimodal retrieval contributed to performance. Additional evaluations included MedMCQA, pre-operative guidance questions, demographic perturbations and a study involving 20 clinicians. In the clinician study, AGMR-Med achieved a mean answer-validity score of 2.9 out of 3 and a mean response time of 5.8 s, compared with 9.7 s for conventional web search. The results support further investigation of adaptive evidence integration for medical question answering. Broader external validation remains necessary to establish reproducibility, clinical applicability and performance across underrepresented specialties.

1 Introduction

Clinical information seeking requires timely access to evidence that is both relevant to the question and sufficiently reliable to inform professional judgement. Conventional point-of-care searches often require clinicians to consult several databases and guidelines, then reconcile their findings. Large language models (LLMs) offer

a conversational interface to this process. Early evaluations showed that ChatGPT could approach the passing threshold on selected United States Medical Licensing Examination (USMLE) question sets [1], while subsequent work examined the capabilities and limitations of GPT-4V on questions containing clinical images [2]. Comparisons with human experts have also identified promising medical advice capabilities, although performance depends on the task and evaluation setting [3]. Examination performance alone, however, does not establish that a system provides current, well-supported or clinically appropriate answers.

Retrieval-augmented generation (RAG) addresses part of this problem by incorporating external evidence into the generation context [4]. In medicine, Vitruvius, subsequently described as MedSearch and Arkangel AI, uses multiple LLM stages to retrieve and synthesise medical information [5]. A separate MedSearch user study evaluated answer validity and the time required to answer clinical questions [6]. These systems illustrate the value of connecting conversational generation to medical evidence, while also showing why benchmark accuracy and practical information-seeking performance should be assessed separately.

Medical evidence extends beyond narrative passages. The Unified Medical Language System (UMLS) links biomedical concepts across terminologies [7], and clinical guidelines communicate recommendations through tables, flowcharts and images as well as prose. Graph-informed retrieval can expose relationships among concepts, whereas visual retrieval can recover information that is difficult to represent in text alone. The medical GraphRAG pipeline investigated by Sekar et al. combines a UMLS-based graph with textbook embeddings [8]. Its use of structured and unstructured sources illustrates that the remaining challenge is not simply to add a graph, but to integrate different forms of evidence according to the requirements of each question.

The relative value of these sources varies with the query. A question about a drug relationship may benefit from graph paths, whereas an imaging question may require visual evidence. A fixed retrieval mixture can allocate substantial context to sources that add little information. AGMR-Med therefore uses a domain-aware representation and an adaptive weighting network to vary the contribution of textual, graph and multimodal retrieval. Its aggregation and verification stages then connect the selected evidence to individual answer claims.

Evaluation must distinguish answer correctness from evidentiary support. Answered with Evidence examines whether biomedical responses are supported by available evidence [9], and SourceCheckup evaluates the relevance and supportiveness of medical references cited by LLMs [10]. Clinical RAG systems such as Almanac also motivate evaluation against curated medical resources [11]. These studies inform our emphasis on citation fidelity, retrieval quality, latency and clinician assessment alongside multiple-choice accuracy. This study makes four contributions:

- We proposed AGMR-Med, a modular framework that integrates text, medical knowledge graphs and visual evidence through domain-aware encoding and adaptive retrieval weighting.
- We develop a multi-level evidence aggregator and a claim-level verifier to identify unsupported statements and improve the alignment between answers and citations.
- We evaluate the framework using MedQA-USMLE, a multimodal guideline corpus and additional question sets, with ablation analyses, clinician assessment and resource measurements.
- We identify the implementation and evaluation resources needed to reproduce the retrieval pipeline and support further investigation of heterogeneous medical evidence.

2 Related Work

Retrieval-Augmented Generation. RAG combines a generative model with external retrieval so that answers can draw on information beyond model parameters [4]. Medical applications differ in how they formulate queries, select evidence and present support. MKRAG retrieves medical facts and injects them into an LLM prompt; its evaluation reported an increase

in Vicuna-7B accuracy from 44.46% to 48.54% [12]. This finding supports the usefulness of retrieval, without implying that gains transfer unchanged across models or datasets.

Rationale-guided retrieval and evidence-selection methods further address the quality of retrieved context. RAG² uses generated rationales, balanced retrieval across biomedical corpora and a filtering model to reduce unhelpful evidence [13]. Med-R² organises retrieval and reasoning around an evidence-based medicine workflow [14]. These approaches improve the way textual evidence is selected and used. AGMR-Med investigates a complementary question: how to vary the contribution of textual, graph and visual sources within one retrieval framework.

Multimodal retrieval offers another route to broader evidence coverage. MuRAG retrieves from a memory containing images and text and demonstrates the feasibility of multimodal retrieval-augmented question answering outside the medical domain [15]. Its findings provide a methodological precedent rather than direct evidence of clinical performance. In medical retrieval, the representation of specialised terminology is also important. MedCPT learns biomedical retrieval representations from PubMed search logs [16], while lexical retrieval can complement dense embeddings when precise terminology matters. AGMR-Med combines dense and sparse textual signals and integrates them with graph and visual retrieval.

Medical LLM Benchmarks. MedQA provides medical examination questions with answer choices and reference answers [17]. Studies of ChatGPT and GPT-4V demonstrate that performance depends on the examination subset, input modality and scoring protocol [1, 2]. For example, Yang et al. reported 92.7% GPT-4V accuracy across all questions in their USMLE Step 3 sample, but 88.9% on its image-containing subset [2]. These results should not be interpreted as scores on an identical MedQA-USMLE test set. MedMCQA broadens evaluation to medical entrance examinations, with more than 194,000 questions covering 21 medical subjects [18]. Evaluations of pre-operative guidance additionally examine whether generated recommendations follow clinical instructions [19]. Long-context medical RAG research shows that the position and density of relevant evidence can affect answer quality; BriefContext addresses this problem through a map-reduce workflow [20]. Together, these studies indicate the need to evaluate different tasks and evidence conditions rather than infer general clinical competence from one accuracy measure.

Citation support and demographic sensitivity introduce further evaluation requirements. SourceCheckup and Answered with Evidence assess whether cited or available evidence supports generated statements [9, 10]. Research on race-based medical content also demonstrates that LLMs can reproduce harmful misconceptions [21]. Our evaluation therefore considers evidence support and demographic perturbations, while treating the latter as a preliminary sensitivity analysis rather than a comprehensive assessment of fairness.

Graph-Based and Multimodal Retrieval. Knowledge graphs make relationships among entities explicit. QA-GNN combines language-model representations and graph reasoning for question answering [22], whereas the medical GraphRAG pipeline of Sekar et al. links UMLS entities and relations with textbook retrieval [8]. Biomedical entity alignment methods, including SapBERT, demonstrate how terminology structure can improve concept representations [23]. These studies motivate the use of concept-aware retrieval, although they do not establish the effectiveness of AGMR-Med's specific implementation.

Visual retrieval requires both image representation and alignment with the query. Vision Transformers (ViTs) represent images as sequences of patches [24], and multimodal retrieval methods such as MuRAG show how visual and textual evidence can support generation [15]. Medical use adds requirements concerning image provenance, modality coverage and source traceability. AGMR-Med combines a visual index with textual and graph indices, then learns query-dependent retrieval weights.

The integration of these sources also requires careful evaluation. Citation quality measures distinguish the existence of a reference from its support for a claim [25]. A systematic review of healthcare RAG applications identifies heterogeneous evaluation practices and limited attention to ethical considerations [26]. Accordingly, we evaluate AGMR-Med across several dimensions and discuss the limitations of the available evidence, rather than equating citation presence or benchmark performance with clinical reliability.

3 Methodology

AGMR-Med consists of a domain-aware query encoder, three retrieval modules, an adaptive weighting network and a multi-level evidence aggregator. A generation model produces a draft answer from the selected evidence, and a separate verifier checks the relationship between claims and their citations. Figure 1 summarises the workflow.

3.1 Domain-Aware Query Encoder

For a natural-language question q, the encoder produces two complementary representations:

1. A semantic embedding h_t, obtained from a Transformer encoder $f_{text}(\cdot)$ fine-tuned on medical question-answering data.
2. A concept embedding h_c, obtained by linking medical entities in q to UMLS concepts through an entity linker g(·) and an embedding lookup $f_{graph}(\cdot)$.

The combined representation is h = [h_t; h_c], where the semicolon denotes concatenation. A lightweight classifier d(h) predicts the medical specialty associated with the question and provides domain information to the weighting network. The classifier distinguishes 15 specialties, including cardiology and neurology.

The query encoder retains positional information to represent the sequence of the question. Its concept representation uses UMLS terminology [7] and document-to-graph alignment to associate rare expressions with related concepts. The domain classifier is trained with cross-entropy loss. Incorporating specialty information improved retrieval accuracy by approximately 1.2 percentage points in the reported ablation analysis. This result concerns retrieval accuracy and is distinct from the answer-accuracy ablations in Section 5.2.

3.2 Retrievers

Textual Retriever. The textual module returns the top k_t passages from an index of PubMed abstracts, guidelines and clinical notes. Its dense component uses a PubMedBERT-based bi-encoder [27] to calculate $s_t(q, d) = \cos(h_t, v_d)$, where v_d is the embedding of passage d. A learned scalar combines this dense signal with BM25 scores over the same corpus. The hybrid design accommodates both semantic similarity and lexical matching.

Graph Retriever. The graph module queries G = (V, E), whose vertices represent UMLS concepts and whose edges represent relationships such as is-a, treats and causes. Starting from the linked concepts and h_c, the retriever expands the neighbourhood by up to three hops to identify candidate nodes and paths. A graph score $s_g(q, g)$ combines concept similarity and path relevance. The implementation uses edge-weighted PageRank in candidate ranking and builds on the graph-informed retrieval setting described in prior medical GraphRAG work [8].

Multimodal Retriever. The visual module indexes guideline images, charts and associated captions, together with radiology and pathology images curated from external datasets. Images are processed to remove patient identifiers. A ViT-based encoder $f_{img}(\cdot)$ [24] is fine-tuned using MIMIC-CXR [28] and guideline diagrams. Cross-modal retrieval is expressed as $s_m(q, I) = \cos(h_t, f_{img}(I))$, where I is an image. Computing this similarity requires compatible text and image representations. The three retrievers provide complementary evidence for subsequent weighting and aggregation.

3.3 Adaptive Weighting Network

The weighting network assigns query-dependent importance to the three retrieval modules. Its normalised output is expressed as

$$[\alpha_t, \alpha_g, \alpha_m] = \text{softmax}(W \cdot [h; d(h)] + b) \quad (1)$$

where α_t, α_g and α_m are the textual, graph and multimodal weights, respectively; W and b are learnable parameters; and d(h) supplies domain information. The weights are non-negative and sum to one. For a candidate evidence unit e, the combined relevance score is

$$S(q, e) = \alpha_t s_t(q, e) + \alpha_g s_g(q, e) + \alpha_m s_m(q, e) \quad (2)$$

Evidence units may be passages, graph paths or images. Candidates are ranked using S(q, e), and the highest-ranked units are selected for aggregation. The operational use of this score depends on how scores are placed on a comparable scale and how components unavailable for a particular evidence unit are handled.

The implemented network is a two-layer multilayer perceptron (MLP) with 128 hidden units and ReLU activations. Supervised pretraining uses cross-entropy loss against proxy targets derived from initial retrieval scores. Reinforcement learning then refines the weighting policy using answer accuracy and citation fidelity as reward signals. We use REINFORCE [29], a variance-reduction baseline and entropy regularisation. An alternative conditioning scheme incorporating retrieval diversity was explored, but the simpler conditioning scheme was more stable in the reported experiments.

3.4 Evidence Aggregation and Answer Generation

Retrieved units are grouped by modality and ranked using S(q, e). The selected passages, graph paths and visual evidence are supplied with the question to a generation model f_{gen}. In addition to concatenating the retrieved material, the aggregator applies multi-head cross-attention over evidence representations to form a summary of salient relationships. This summary provides an additional input to the generation model and is intended to support the integration of heterogeneous sources.

A separate model f_{ver} checks whether each sentence in the draft answer is supported by at least one retrieved source. The design is informed by evidence-support evaluation in SourceCheckup and Answered with Evidence [9, 10]. Statements with missing support or irrelevant citations trigger revision. The generation model then revises the affected claims using additional evidence, and unsupported claims are removed. The verification process aims to improve support and consistency; it does not guarantee that all clinically relevant errors are detected.

The final response includes inline citations linked to the supporting documents. Citation evaluation considers the precision and recall of evidence support, drawing on established citation-assessment methods [25]. The system also estimates answer confidence. Calibration is assessed by comparing confidence with empirical correctness, consistent with the distinction between confidence and accuracy described by Guo et al. [30].

3.5 Training

Training proceeds through representation learning, supervised component training and feedback-based refinement. Question-evidence pairs from MedQA-USMLE and guideline documents are used for contrastive learning of retrieval representations. In-batch negatives and a cosine-based objective encourage alignment with relevant evidence while separating unrelated pairs. The 15-class domain classifier uses cross-entropy loss, and the weighting network is pretrained on retrieval-derived proxy targets.

Human annotators assess generated answers for correctness, citation fidelity and clarity. These assessments inform a reward combining the annotated criteria with citation-support scores. REINFORCE updates the retrieval-weighting policy, with a baseline to reduce variance and an entropy term to encourage exploration [29]. Feedback-based refinement follows the general motivation of reinforcement learning from human feedback (RLHF) [31]; the retrieval policy and generation model have distinct roles in the training pipeline.

The generation model is based on Llama 3 [32] and is fine-tuned using question-evidence-answer triples. Most parameters remain frozen, while a small adapter module is trained. The generation-stage objectives incorporate human feedback and

citation-based rewards, penalising unsupported statements. Output diversity and confidence are monitored, and calibration objectives are used to align confidence with observed accuracy. Training is implemented in PyTorch on four NVIDIA A100 GPUs for up to three epochs, with early stopping based on validation citation fidelity.

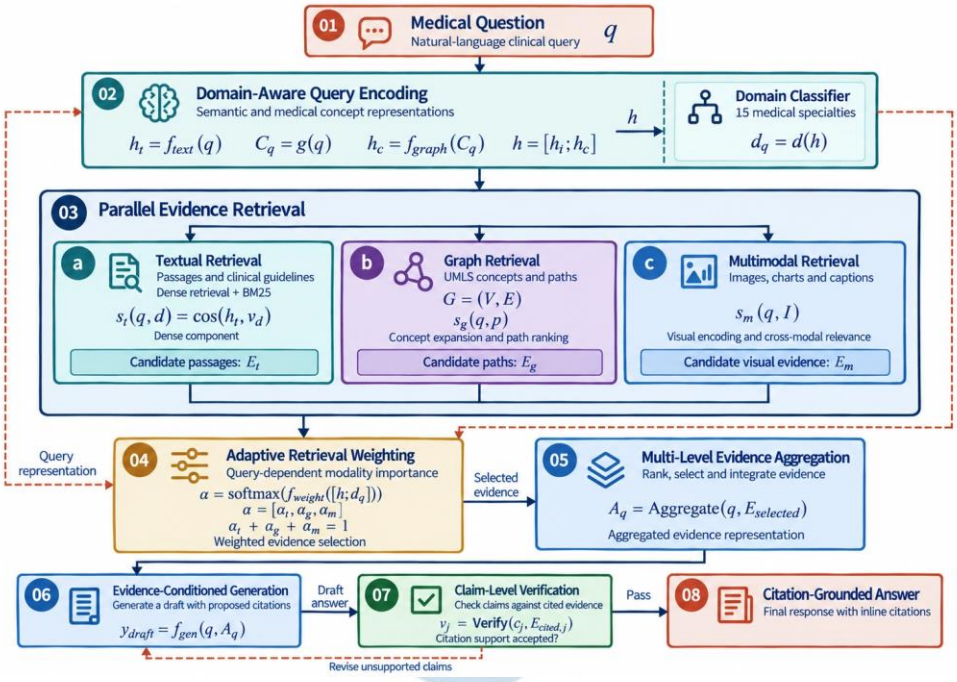


Figure 1: Overview of the AGMR-Med architecture. Semantic and concept representations support textual, graph and multimodal retrieval. Domain information conditions adaptive weighting. Selected evidence is aggregated to generate a draft answer, which undergoes claim-level citation verification and revision before delivery.

4 Experimental Setup

4.1 Datasets

MedQA-USMLE. The evaluation uses a selected set of 1,273 multiple-choice questions, each containing a question stem, answer options and a reference answer. MedQA's official US English test split contains 1,273 questions [17]. The experimental protocol described here further divides the selected questions into 60% training, 20% validation and 20% testing, with unsupported image-based items excluded. Consequently, this protocol must be distinguished from evaluation on the complete official test split when interpreting comparisons with published results.

Guideline Corpus. The multimodal corpus contains 500 clinical practice guidelines spanning cardiology, oncology and neurology. The documents contain narrative recommendations, tables, flowcharts and diagnostic images. Images and their associated captions are extracted and stored in a visual index, alongside the textual and graph representations used by the retrieval modules.

Real-World User Study. Twenty clinicians, including residents and attending physicians, interact with AGMR-Med and baseline search systems. Each participant submits five open-ended clinical questions drawn from their practice. Answers are assessed using a three-point validity scale, and response times are recorded. The evaluation is motivated by prior MedSearch validation [6], while retaining the distinction between the present participants and those in the published study.

MedMCQA and Additional QA Sets. MedMCQA contains more than 194,000

questions across medical subjects [18]. For the additional evaluation, 5,000 non-overlapping questions are held out for testing and the remaining questions are used for pretraining. The selected questions include pharmacology, physiology and radiology. We also evaluate 150 pre-operative guidance questions requiring adherence to clinical recommendations. This setting is relevant to prior evaluations of retrieval augmentation for pre-operative assessment [19]. These experiments examine performance on additional tasks; because MedMCQA is also used for pretraining, its results are not an evaluation of transfer to an entirely unseen dataset.

4.2 Implementation Details

The textual retriever combines PubMedBERT-based dense embeddings with a BM25 index. The knowledge graph contains approximately 2.3 million UMLS concepts, with relations drawn from the Metathesaurus and augmented with guideline-specific relationships, including dose-response relations. The visual encoder is based on a ViT and is fine-tuned using MIMIC-CXR and guideline diagrams [24, 28]. The weighting network uses two MLP layers, 128 hidden units and ReLU activations.

The retrieval candidate counts are $k_t = 20$, $k_g = 10$ and $k_m = 5$. These values are selected using the validation set to balance recall and computational cost. The learning rate is 5×10^{-5}, the batch size is 32, dropout is 0.1 and the optimiser is Adam. These settings are held constant across the trainable experiments described in the implementation. Runs use NVIDIA A100 GPUs, and an AGMR-Med training run requires approximately 12 h to converge.

We measure answer accuracy on multiple-choice tasks, citation-support precision, recall and F1, mean response time, and normalised discounted cumulative gain (NDCG). Citation assessment considers whether cited evidence supports answer statements [25]. Confidence calibration is assessed against observed correctness [30]. A demographic sensitivity evaluation alters names and pronouns and measures changes in answers and confidence, motivated by documented bias risks in medical LLMs [21].

The experiments use three random seeds. Table 1 presents means and standard deviations across runs, while the statistical analysis additionally reports bootstrap confidence intervals for performance differences. The two uncertainty summaries represent different quantities and should be interpreted separately.

5 Results and Analysis

Table 1 summarises the comparison with GPT-4 without retrieval, Vitruvius/MedSearch and GraphRAG. It includes accuracy, citation fidelity, mean response time and NDCG.

Table 1: Evaluation on MedQA-USMLE and the guideline corpus.

Model	Accuracy (%)	Citation fidelity (%)	Response time (s)	NDCG
GPT-4 (no RAG)	65.8 ± 0.9	12.3 ± 1.0	15.7 ± 0.6	0.42
Vitruvius (MedSearch)	90.3 ± 0.5	72.4 ± 1.1	9.2 ± 0.4	0.67
GraphRAG	88.5 ± 0.8	80.6 ± 0.9	8.4 ± 0.5	0.74
AGMR-Med (ours)	**91.7 ± 0.4†**	**88.4 ± 0.6†**	**7.6 ± 0.3†**	**0.81**

Mean ± standard deviation across three runs. Higher accuracy, citation fidelity and NDCG are better; lower response time is better. † denotes reported significance against the best baseline ($p < 0.001$).

AGMR-Med achieves 91.7% answer accuracy, compared with 90.3% for Vitruvius/MedSearch and 88.5% for GraphRAG, corresponding to increases of 1.4 and 3.2 percentage points. Citation fidelity reaches 88.4%, exceeding the corresponding baseline values by 16.0 and 7.8 percentage points. Mean response time is 7.6 s, compared with 9.2 s for Vitruvius/MedSearch and

8.4 s for GraphRAG. The NDCG value of 0.81 also exceeds the values reported for the comparison systems. These descriptive results favour AGMR-Med under the evaluation conditions represented in the table.

5.1 Ablation Study

Removing the adaptive weighting network reduces answer accuracy from 91.7% to 89.4% and citation fidelity from 88.4% to 80.1%. Excluding multimodal retrieval reduces answer accuracy to 90.6% and increases mean response time to 8.2 s. The ablations indicate that both components contribute to the reported performance. They do not, by themselves, establish which retrieval operations account for the latency difference or how the effect varies across question types.

5.2 Expert Evaluation and User Study

AGMR-Med receives a mean answer-validity score of 2.9 out of 3. The separately published MedSearch study reports a score of 2.8 out of 3 [6]. Because the studies involve different participants and potentially different questions, this numerical comparison provides context rather than a controlled estimate of superiority. Participants in the present study identify citation clarity and visual guideline diagrams as useful features. Mean response time is 5.8 s with AGMR-Med and 9.7 s with conventional web search. These user-study timings are distinct from the model-response timings in Table 1.

5.3 Statistical Validation

The statistical analysis compares AGMR-Med with Vitruvius/MedSearch and GraphRAG using paired t-tests for accuracy, citation fidelity and response time. The reported tests yield $p < 0.001$. The reported 95% confidence interval for the accuracy difference between AGMR-Med and Vitruvius/MedSearch is 1.1 to 1.7 percentage points. A confidence interval of 6.5 to 9.1 percentage points is reported for citation fidelity. Interpretation of these inferential results depends on the pairing unit, the resampling procedure and the baseline associated with each interval.

5.4 Comparative Visualisation

Figure 2 presents the accuracy, citation fidelity and response-time values from Table 1. AGMR-Med has the highest accuracy and citation fidelity and the lowest mean response time among the evaluated systems. The plotted uncertainty corresponds to the standard deviations reported in the table.

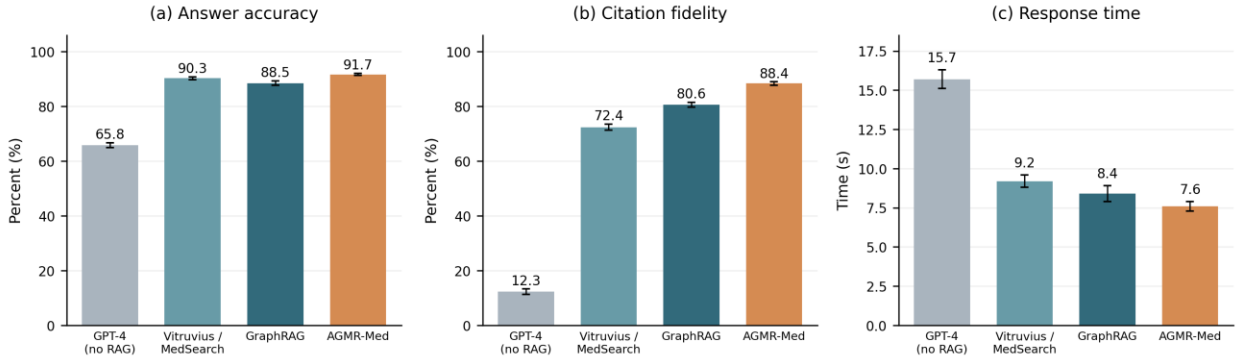


Figure 2: Comparative accuracy, citation fidelity and response time. Values and error bars reproduce the means ± standard deviations in Table 1; the panels use the same model order.

Specialty-level evaluation yields AGMR-Med accuracies of 92.1% in cardiology, 91.5% in oncology and 91.2% in neurology. The corresponding Vitruvius/MedSearch values are 90.7%, 89.9% and 89.4%. Cardiology citation fidelity is 89.2% for AGMR-Med and 78.1% for Vitruvius/MedSearch. These results are consistent with the rationale for domain-aware evidence integration, although they do not isolate the effect of individual retrieval modalities within each specialty.

On the held-out MedMCQA subset, AGMR-Med achieves 86.3% accuracy and 82.5% citation fidelity. Accuracy is 82.1% for GraphRAG and 79.8% for Vitruvius/MedSearch. On the pre-operative guidance set, AGMR-Med achieves 89.0% accuracy, compared with 85.7% for GraphRAG. The results extend the evaluation beyond the selected USMLE questions. Their interpretation depends on the dataset-specific training exposure, the definition of correctness and the separation between pretraining and evaluation material.

5.5 Fairness Evaluation

The demographic sensitivity analysis varies patient names and pronouns while retaining the remaining question content. Across the male and female variants examined, AGMR-Med's accuracy differs by less than 1.0 percentage point, and its confidence scores differ by less than 0.02. GPT-4 exhibits an accuracy difference of 3.4 percentage points across the evaluated variants. These observations indicate limited sensitivity under the tested perturbations. They do not establish the absence of bias, statistical equivalence between groups or fairness across race, age, socioeconomic status and intersecting attributes. The evaluation is narrower than the clinical bias concerns identified in previous work [21].

5.6 Efficiency and Resource Analysis

AGMR-Med's mean response time of 7.6 s is approximately 9.5% lower than GraphRAG's 8.4 s and 17.4% lower than Vitruvius/MedSearch's 9.2 s. These reductions are calculated from the means in Table 1. Peak inference memory consumption is 12 GB for AGMR-Med and 15 GB for GraphRAG. Precomputing graph indices and image embeddings distributes indexing costs across subsequent queries.

Measurements obtained using NVIDIA Nsight indicate 10% lower GPU power draw than GraphRAG. Power draw and total energy consumption are different quantities, so this result should be interpreted as a power comparison. The reported latency, memory and power measurements suggest potential efficiency benefits, while a detailed breakdown of retrieval, aggregation, generation and verification costs is needed to explain their sources.

6 Discussion

The reported results support the investigation of query-dependent integration of textual, structured and visual evidence. AGMR-Med combines domain information with adaptive weighting and claim-level verification, and its ablations associate these components with improvements in answer accuracy and citation fidelity. The clinician study further suggests that traceable citations and visual guideline material are useful to participants. These findings concern the evaluated tasks and users; they do not establish that the system improves patient outcomes.

Evidence retrieval and evidence verification serve different purposes. Retrieving a relevant passage or image improves the information available to the generator, while verification assesses whether a particular claim is supported by its cited source. Neither step establishes that the source is current, sufficiently authoritative or applicable to an individual patient. This distinction helps explain why accuracy, citation quality and clinical evaluation should be reported separately. It is consistent with the evaluation challenges identified in the healthcare RAG literature [26].

The guideline corpus covers three specialties and may inadequately represent rare diseases, surgical procedures and other image modalities. Errors in entity linking or concept representation can propagate into graph retrieval. Cross-modal similarity also depends on how textual and visual embeddings are aligned, and score fusion depends on comparable retrieval scales. These implementation choices require detailed specification to make the framework reproducible.

The evaluation protocol also limits the strength of the conclusions. The selected MedQA questions are repartitioned rather than clearly evaluated using the complete official benchmark split. MedMCQA is used for both pretraining and held-out testing, so performance on that dataset cannot be interpreted as wholly external transfer. The clinician sample is small, and the demographic analysis mainly examines names and gender-pronoun variants. Wider evaluations are needed

across institutions, specialties, languages and demographic groups. The reported confidence intervals and significance tests also require sufficiently detailed pairing and resampling information to support independent assessment.

Image curation introduces privacy and licensing requirements, including the conditions governing access to de-identified clinical datasets. Citation verification remains imperfect when evidence is incomplete, contradictory or out of date. A clinically deployable system would therefore require ongoing monitoring, professional oversight and evaluation of errors with potential consequences for patients. The present findings provide a basis for further research rather than evidence of autonomous clinical readiness.

Further work should expand the knowledge graph and guideline corpus, evaluate additional image modalities and measure performance on low-resource languages. Retrieval and generation could be coordinated more closely through feedback-based optimisation, with explicit evaluation of the resulting latency and verification costs. Integrating patient records would require a separate evaluation of privacy, consent, data access and clinical applicability. Prospective studies should assess whether the framework supports professional decisions safely across practice settings, including scenarios with ambiguous or conflicting evidence.

7 Conclusion

AGMR-Med integrates unstructured text, structured medical knowledge and visual evidence through domain-aware encoding, adaptive retrieval weighting and claim-level citation verification. The evaluation reports 91.7% answer accuracy, 88.4% citation fidelity and a mean response time of 7.6 s, with favourable descriptive comparisons against the evaluated Vitruvius/MedSearch and GraphRAG baselines. Ablations and clinician assessments provide additional evidence for investigating adaptive evidence integration. The contribution is the combined retrieval and verification framework, together with an evaluation extending beyond answer accuracy. Reproducible benchmark protocols, fully specified methods and broader external validation remain necessary before drawing conclusions about clinical effectiveness or general fairness. Code, model resources and evaluation scripts are intended for release to support independent assessment and further development.

REFERENCES

Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198

Yang Z, Yao Z, Tasmin M, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. Journal of Medical Internet Research. 2025;27:e65146. https://doi.org/10.2196/65146

Jo E, Song S, Kim JH, et al. Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis With Human Experts. JMIR Medical Education. 2024;10:e51282. https://doi.org/10.2196/51282

Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020;33:9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Villa MC, et al. Arkangel AI: A conversational agent for real-time, evidence-based medical question-answering. Intelligence-Based Medicine. 2025;12:100274. https://doi.org/10.1016/j.ibmed.2025.100274

Castano-Villegas N, Llano I, Villa MC, Zea J. Real-world Validation of MedSearch: a conversational agent for real-time, evidence-based medical question-answering. medRxiv [preprint]. 2025. https://doi.org/10.1101/2025.05.02.25326659

Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. Journal of the American Medical Informatics Association. 1998;5(1):1-11. https://doi.org/10.1136/jamia.1998.0050001

Sekar T, Kushal, Shankar S, Mohammed S, Fiaidhi J. Investigations on using Evidence-Based GraphRag Pipeline using LLM Tailored for USMLE Style Questions. medRxiv [preprint]. 2025. https://doi.org/10.1101/2025.05.03.25325604

Baldwin JD, Dinh C, Mukerji A, Sanghavi N, Gombar S. Introducing Answered with Evidence - a framework for evaluating whether LLM responses to biomedical questions are founded in evidence. arXiv [preprint]. 2025;2507.02975. https://doi.org/10.48550/arXiv.2507.02975

Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nature Communications. 2025;16:3615. https://doi.org/10.1038/s41467-025-58551-6

Zakka C, Shad R, Chaurasia A, et al. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI. 2024;1(2):AIoa2300068. https://doi.org/10.1056/AIoa2300068

Shi Y, Xu S, Yang T, et al. MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering. AMIA Annual Symposium Proceedings. 2025;2024:1011-1020. https://pubmed.ncbi.nlm.nih.gov/40417500/

Sohn J, Park Y, Yoon C, et al. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. In: Proceedings of NAACL-HLT 2025, Volume 1: Long Papers. Association for Computational Linguistics; 2025. p. 12739-12753. https://doi.org/10.18653/v1/2025.naacl-long.635

Lu K, Liang Z, Pan D, et al. Med-R²: Crafting Trustworthy LLM Physicians via Retrieval and Reasoning of Evidence-Based Medicine. In: Proceedings of the ACM Web Conference 2026. ACM; 2026. p. 1864-1875. https://doi.org/10.1145/3774904.3792076

Chen W, Hu H, Chen X, Verga P, Cohen W. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In: Proceedings of EMNLP 2022. Association for Computational Linguistics; 2022. p. 5558-5570. https://doi.org/10.18653/v1/2022.emnlp-main.375

Jin Q, Kim W, Chen Q, et al. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics. 2023;39(11):btad651. https://doi.org/10.1093/bioinformatics/btad651

Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences. 2021;11(14):6421. https://doi.org/10.3390/app11146421

Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. Proceedings of the Conference on Health, Inference, and Learning. Proceedings of Machine Learning Research. 2022;174:248-260. https://proceedings.mlr.press/v174/pal22a.html

Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. npj Digital Medicine. 2025;8:187. https://doi.org/10.1038/s41746-025-01519-z

Zhang G, Xu Z, Jin Q, et al. Leveraging long context in retrieval augmented language models for medical question answering. npj Digital Medicine. 2025;8:239. https://doi.org/10.1038/s41746-025-01651-w

Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. npj Digital Medicine. 2023;6:195. https://doi.org/10.1038/s41746-023-00939-z

Yasunaga M, Ren H, Bosselut A, Liang P, Leskovec J. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In: Proceedings of NAACL-HLT 2021. Association for Computational Linguistics; 2021. p. 535-546. https://doi.org/10.18653/v1/2021.naacl-main.45

Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. In: Proceedings of NAACL-HLT 2021. Association for Computational Linguistics; 2021. p. 4228-4238. https://doi.org/10.18653/v1/2021.naacl-main.334

Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. 2021. https://arxiv.org/abs/2010.11929

Gao T, Yen H, Yu J, Chen D. Enabling Large Language Models to Generate Text with Citations. In: Proceedings of EMNLP 2023. Association for Computational Linguistics; 2023. p. 6465-6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health. 2025;4(6):e0000877. https://doi.org/10.1371/journal.pdig.0000877

Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2022;3(1):2:1-2:23. https://doi.org/10.1145/3458754

Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data. 2019;6:317. https://doi.org/10.1038/s41597-019-0322-0

Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning. 1992;8:229-256. https://doi.org/10.1007/BF00992696

Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research. 2017;70:1321-1330. https://proceedings.mlr.press/v70/guo17a.html

Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022;35:27730-27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html

Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 Herd of Models. arXiv [preprint]. 2024;2407.21783. https://doi.org/10.48550/arXiv.2407.21783

International Journal of Advanced Research