JOURNAL OF GLOBAL ENGINEERING REVIEW ISSN: 3071-4656

Medical RAG Pipeline Selection with Retrieval Quality Prediction and LLM Safety Explanations

Jakob Jensen
Computer Science, Aarhus University, Aarhus, MJT, Denmark
jjensen.projects@outlook.com

Abstract
Medical retrieval-augmented generation (RAG) systems must match retrieval strategy to question structure, estimate whether the retrieved evidence is strong enough to support an answer, and communicate clearly when the evidence is insufficient. This paper presents RQ-SafeSelector, an evidence-aware pipeline-selection framework evaluated on the 420-question RAGCare-QA benchmark across six medical specialties, three complexity levels, and three annotated RAG pipeline classes. The framework compares sparse TF-IDF, multi-vector lexical–character retrieval, and graph-enhanced hybrid retrieval; predicts retrieval quality from out-of-fold scores; and routes each question to Basic RAG, Multi-vector RAG, or Graph-enhanced RAG with a structured safety explanation and abstention gate. Graph-enhanced Hybrid achieved the strongest overall Recall@3 (0.336) and nDCG@5 (0.339), while Multi-vector Hybrid obtained the highest Recall@5 (0.548). RQ-SafeSelector reached 0.933 accuracy and 0.833 macro-F1 for pipeline selection. The calibrated retrieval-quality predictor achieved ROC-AUC 0.686 with expected calibration error 0.036. Applying the quality gate reduced the audit-defined unsupported-answer risk to 0.000 while answering 30.7% of questions, with Success@3 of 0.527 among answered cases. The findings support evaluation of medical RAG as a routed, evidence-aware, and selectively answering system rather than as a single retriever followed by an unconstrained generator.

Keywords: *medical retrieval-augmented generation; RAGCare-QA; pipeline selection; retrieval-quality prediction; calibrated abstention; safety explanations; healthcare AI.*

1. Introduction

Retrieval-augmented generation separates information access from language generation, allowing a system to condition its response on external evidence rather than relying exclusively on parametric memory [1]. This separation is especially consequential in medicine. A fluent response may still be unsafe when it is based on an irrelevant passage, when a clinically important qualification is missing, or when the system expresses more certainty than the evidence warrants. Medical language models can encode broad clinical knowledge [11], yet their apparent competence does not remove the need for evidence tracing, uncertainty management, and human oversight [14], [20], [21].

Medical questions are heterogeneous at both the linguistic and evidential levels. Some questions contain an explicit disease name and can be localized through direct lexical overlap. Others use synonyms, abbreviations, or morphological variants. More complex questions combine mechanisms, diagnostic criteria, risks, and management relationships that may be distributed across several statements. Consequently, a single fixed retrieval strategy is unlikely to be optimal for every question, even within one specialty. Large-scale benchmarking has likewise shown that medical RAG performance depends on the combination of corpus, retriever, and backbone model rather than on one universally dominant configuration [12].

RAGCare-QA provides a focused setting for studying this variability. The benchmark contains 420 theoretical medical knowledge questions from six specialties and assigns each item to one of three retrieval architecture classes: Basic RAG, Multi-vector RAG, or Graph-enhanced RAG [10]. These annotations convert retrieval design

from a global configuration choice into a per-question routing problem. The practical challenge is therefore twofold: the system must select an appropriate retrieval pipeline and determine whether the selected pipeline actually produced enough evidence to justify answering.

That second decision is a calibration problem. Raw retrieval scores are not probabilities of evidence sufficiency, and a high similarity score can occur for a topically related but non-supporting context. Probability calibration and selective classification provide a principled basis for converting model scores into an answer-or-abstain policy [15], [16]. Evidence-first scientific RAG has similarly demonstrated the value of conditioning answer coverage on retrieval confidence instead of requiring a system to answer every query [22]. The present study applies this principle to medical pipeline routing and makes the evidence decision visible through a structured safety explanation.

This paper makes four contributions. First, it compares three lightweight retrieval pipelines on exact row-level evidence localization. Second, it evaluates pipeline routing with class-sensitive metrics so that the rare Graph-enhanced RAG class is not hidden by overall accuracy. Third, it predicts Success@3 from grouped out-of-fold features and calibrates the resulting probabilities for answer gating. Fourth, it evaluates an end-to-end policy that connects route selection, evidence sufficiency, abstention, and explanation completeness. The study asks whether small auditable retrievers can localize supporting context, whether a router can preserve minority-class sensitivity, and whether calibrated evidence gating can improve the reliability of the answered subset without presenting weakly supported medical content.

2. Literature Review

2.1 Retrieval architectures and medical question answering

The modern RAG pipeline draws on a long line of information-retrieval research. Sparse ranking remains competitive when queries and documents share distinctive technical terms; BM25 formalizes this behavior through term saturation and document-length normalization [3]. Dense Passage Retrieval instead learns query–passage representations that improve semantic matching when surface forms differ [2]. Sentence-BERT provides efficient sentence-level embeddings [4], while interaction-based rerankers can refine an initial candidate set by jointly modeling query and passage tokens [5]. Fusion-in-Decoder demonstrates how multiple retrieved passages can be aggregated during generation [6]. These approaches ultimately build on attention-based representation learning [17], but their operational properties differ substantially in latency, transparency, and sensitivity to lexical cues.

Question-answering benchmarks have shaped how these retrieval components are evaluated. SQuAD established large-scale extractive reading comprehension [7], and SQuAD 2.0 introduced unanswerable questions as an explicit test of answerability [8]. PubMedQA extended QA evaluation to biomedical research questions [9]. RAGCare-QA adds a different supervisory signal: it associates each medical question with a retrieval architecture class and a paired evidence context [10]. This structure is well suited to studying architecture selection rather than only final answer accuracy.

Medical RAG introduces additional concerns because correct reasoning must remain tied to clinically appropriate evidence. MIRAGE and MedRAG evaluated thousands of medical questions across many corpus–retriever–model combinations and found that system design choices materially affect performance [12]. Prompting methods such as chain-of-thought and zero-shot reasoning can improve reasoning behavior [18], [19], but they do not guarantee that the premises used by the model are present in the retrieved evidence. For this reason, retrieval quality and evidence sufficiency must be measured independently of linguistic fluency.

2.2 Adaptive routing, evidence chains, and verification

Adaptive RAG research increasingly treats retrieval as a conditional action. Self-RAG allows a model to retrieve, generate, and critique its own output [13], while claim-aware scientific RAG combines retrieval, evidence extraction, and abstention [22]. These systems share an important design principle: more retrieval is not

automatically better. Retrieval should be invoked or escalated when the question requires it, and the final answer should remain connected to an identifiable evidence path.

Evidence verification has also been explored outside medicine. Lightweight hallucination firewalls use auditable lexical and self-check features to block unsupported candidates [23]. Financial RAG systems emphasize evidence grounding and numerical consistency [25], [33], while evidence-chain methods add source attribution and provenance explanations at word or sentence level [26]. In enterprise operations, evidence-grounded DevOps and incident-triage systems use domain-aware retrieval, citation filtering, and document-family constraints to reduce unsupported responses [27], [34]. Post-hoc semantic verification further decomposes an answer into claims, retrieves evidence for each claim, and returns traceable citations [28]. Although these studies address different domains, they identify controls that transfer naturally to medical RAG: retrieve the right evidence, measure agreement with that evidence, and make provenance visible to the user.

2.3 Calibration, selective answering, and safety communication

Calibration is required whenever a confidence value controls an action. Neural classifiers are often miscalibrated even when their ranking accuracy is strong [15]. Selective classification addresses this problem by allowing the system to abstain on uncertain cases and evaluating the resulting coverage–risk trade-off [16]. Related work on human-uncertainty distillation shows that uncertainty labels can improve the calibration of vision systems [32], while risk-calibrated biomedical search applies selective decision making to query expansion so that potentially harmful expansions can be withheld [29]. These studies motivate the use of calibrated retrieval success probabilities rather than unnormalized similarity scores.

Refusal behavior also requires careful design. Ambiguity-aware anomaly detection has shown how selective refusal can improve accepted-case reliability while preserving explanatory narratives [24]. VerifySafe combines a front-end detector with response-level self-verification [30], and multi-stage refusal methods seek to preserve utility while resisting jailbreak behavior [31]. In medical RAG, the same conceptual distinction applies: a generic disclaimer is not equivalent to an evidence-aware safety explanation. A useful explanation should state which route was selected, whether the retrieval condition passed, what uncertainty remains, and whether the system answered or abstained. RQ-SafeSelector integrates these elements into a single measurable policy rather than treating them as an afterthought.

3. Method

3.1 Dataset and task definition

The evaluation used the public RAGCare-QA benchmark [10]. The dataset contains 420 one-best-answer theoretical medical questions spanning Cardiology, Endocrinology, Family Medicine, Gastroenterology, Neurology, and Oncology. Each item includes the question, answer information, an associated context, specialty and complexity metadata, and an annotated RAG pipeline label. The three complexity groups contain 150 basic, 181 intermediate, and 89 advanced questions. Table 1 reports the joint specialty–complexity composition used throughout the experiments.

Table 1. RAGCare-QA specialty and complexity composition.

Specialty	Basic	Intermediate	Advanced	Total
Cardiology	32	36	19	87
Endocrinology	28	32	14	74
Family Medicine	26	41	14	81
Gastroenterology	20	22	12	54
Neurology	20	22	15	57
Oncology	24	28	15	67

Specialty	Basic	Intermediate	Advanced	Total
Total	150	181	89	420

The annotated routing labels are imbalanced. As shown in Table 2, Basic RAG accounts for 75.0% of the questions, whereas Graph-enhanced RAG accounts for 5.5%. Accuracy is therefore insufficient as the sole routing metric; macro-F1, balanced accuracy, and Graph-F1 are reported to expose minority-class behavior.

Table 2. RAG pipeline label distribution.

RAG pipeline	Questions	Percentage
Basic RAG	315	75.0%
Multi-vector RAG	82	19.5%
Graph-enhanced RAG	23	5.5%
Total	420	100.0%

Each question was evaluated against all 420 row-level contexts. The paired context was treated as the gold evidence unit, so a result counted as successful only when that exact context appeared within the requested cutoff. This strict definition isolates evidence localization from answer generation and prevents near-topic passages from being counted as fully supporting evidence.

3.2 System architecture

Figure 1 summarizes the complete RQ-SafeSelector workflow. A medical question is scored by three candidate retrievers. Preliminary rankings and question features feed two related models: a pipeline selector predicts the appropriate retrieval class, and a retrieval-quality predictor estimates the probability that the gold evidence appears within the top three results. The policy then either answers through the selected route, changes to a higher-quality alternative, or abstains. Every action is accompanied by a structured explanation describing the route, evidence condition, uncertainty, and medical scope.

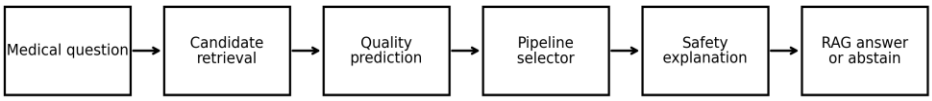

Figure 1. Architecture of RQ-SafeSelector with retrieval-quality gating and safety explanation.

The concrete implementation choices are listed in Table 3. The models are intentionally lightweight so that the routing and gating decisions remain inspectable. The large language model is treated as a downstream rendering component; the measured safety decisions are made before free-form answer generation.

Table 3. Method components and implementation choices.

Component	Implementation	Purpose
Basic TF-IDF retrieval	Word-level TF-IDF over context text	Direct lexical evidence for explicit concepts
Multi-vector Hybrid retrieval	0.65 word TF-IDF + 0.35 character n-gram TF-IDF	Robust matching across spelling, morphology, and terminology variants

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

Component	Implementation	Purpose
Graph-enhanced Hybrid retrieval	0.46 word + 0.24 character + 0.30 concept–relation score	Matching of mechanisms, risks, management concepts, and related entities
Pipeline selector	Question text, metadata, score margins, and entropy	Prediction of Basic, Multi-vector, or Graph-enhanced RAG
Retrieval-quality predictor	Pipeline-specific ranking features with grouped cross-validation	Calibrated prediction of Success@3
Safety explanation	Route rationale, evidence condition, uncertainty, scope, and escalation	Prevention of unsupported medical answers

3.3 Retrieval pipelines

Basic TF-IDF uses a word-level term–document matrix over the context field. It serves as a transparent sparse baseline and is well matched to questions containing distinctive disease names, abbreviations, or diagnostic phrases [3]. Multi-vector Hybrid augments the word representation with character n-grams, which can recover evidence when the query and context differ in spelling, morphology, or tokenization. For each question, the word and character score vectors were normalized and combined as $s\_multi = 0.65s\_word + 0.35s\_char$. This lightweight multi-view design captures some of the robustness associated with semantic matching [4] without requiring an external embedding service.

Graph-enhanced Hybrid adds a relation-overlap channel. Medical concepts, synonyms, mechanisms, risks, management phrases, related terms, and specialty cues were extracted from the question and candidate context. The final score was $s\_graph = 0.46s\_word + 0.24s\_char + 0.30s\_relation$. The graph term is not a full clinical knowledge graph; it is a compact relation-aware signal designed to distinguish contexts that mention the same entity but express different mechanisms or management relationships. All three retrievers ranked the complete 420-context matrix for every question.

3.4 Pipeline selection

Five routing strategies were compared. The majority baseline always selected Basic RAG. The complexity heuristic used the annotated complexity and relation count to escalate questions. The TF-IDF logistic model used only question text. The metadata random forest used question length, option count, relation count, specialty, complexity, preliminary top scores, score margin, and score entropy. RQ-SafeSelector combined TF-IDF question features with the metadata and preliminary retrieval features in a class-balanced logistic model. Five-fold stratified cross-validation produced out-of-fold predictions for every question.

Routing was evaluated with accuracy, macro-F1, balanced accuracy, top-2 accuracy, and Graph-F1. Top-2 accuracy was included because a second route can be useful when the first route fails the evidence gate. Graph-F1 measures sensitivity to the rarest and most relation-oriented class, which would be obscured by aggregate accuracy.

3.5 Retrieval-quality prediction and calibration

Retrieval quality was modeled over question–pipeline pairs. Each question contributed three rows, one for each candidate retriever. The binary target, Success@3, was one when the paired context appeared in the first three ranked contexts and zero otherwise. Features included the top-1 and top-2 scores, their margin, score entropy, question length, token count, relation count, pipeline identity, complexity, and specialty. GroupKFold split the data by question identifier so that the three candidate routes for the same question could not appear in both training and test folds.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

A score-margin logistic model, random forest, gradient boosting model, and calibrated gradient-boosted predictor were compared. ROC-AUC and PR-AUC measured ranking quality; F1 at a 0.50 threshold measured binary classification; Brier score and expected calibration error measured probability quality [15]. The calibrated predictor was selected for the gate because the resulting probability controls an action rather than serving only as a ranking score.

3.6 Safety policy and explanation audit

The end-to-end policy first obtains the route proposed by RQ-SafeSelector. If the calibrated Success@3 probability for that route is below 0.40, the policy examines the alternative route with the highest predicted quality. If no route passes the threshold, the system abstains. Answered cases receive a concise explanation stating the selected route, the evidence condition that passed, the remaining uncertainty, and the educational scope of the response. Abstained cases state that the retrieved evidence did not satisfy the threshold and direct the user to authoritative material or qualified human review.

Explanation completeness was scored from zero to five for the presence of five auditable components: pipeline rationale, evidence condition, uncertainty cue, medical scope boundary, and escalation or abstention. The audit-defined unsupported-answer risk counted cases in which a policy released an answer despite failing its designated evidence condition without triggering the required safeguard. This definition distinguishes a retrieval miss that is explicitly intercepted from a miss that is silently converted into a medical answer.

3.7 Evaluation protocol

Retrieval performance was measured with Recall@1, Recall@3, Recall@5, mean reciprocal rank (MRR), nDCG@5, and mean in-memory ranking latency per question. MRR used the first rank of the exact context within the top 50, with zero assigned when it was not found. Selector and quality-prediction results were based exclusively on out-of-fold predictions. The end-to-end policy comparison reported answer coverage, Success@3 over all questions, Success@3 among answered cases, audit-defined unsupported-answer risk, explanation completeness, and explanation length. Fixed random seeds were used for fold assignment and model initialization.

4. Results and Discussion

4.1 Dataset composition

The benchmark is neither specialty-balanced nor complexity-balanced. Figure 2 shows that Cardiology contains 87 questions and Family Medicine 81, whereas Gastroenterology contains 54. Intermediate questions form the largest complexity group, and advanced questions are the smallest. These differences matter because retrieval behavior can vary with both terminology density and relational structure; aggregate metrics should therefore be interpreted alongside the stratified results.

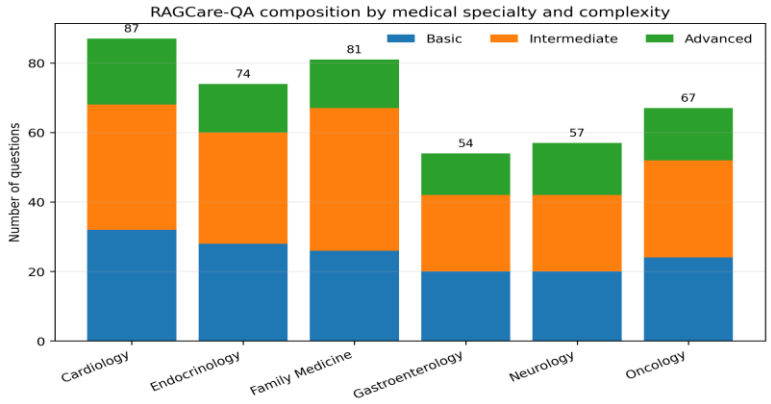

Figure 2. Dataset distribution by medical specialty and complexity.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

4.2 Retrieval performance

Table 4 reports overall exact-context localization. Graph-enhanced Hybrid achieved the strongest Recall@1, Recall@3, MRR, and nDCG@5. Multi-vector Hybrid produced the highest Recall@5, suggesting that character-level matching improves broader candidate coverage. All three methods were fast on the compact corpus; the central difference was evidence coverage rather than latency.

Table 4. Overall retrieval performance on exact context localization.

Retriever	R@1	R@3	R@5	MRR	nDCG@5	Latency (ms/query)
Basic TF-IDF	0.121	0.319	0.521	0.310	0.317	0.0200
Multi-vector Hybrid	0.117	0.331	0.548	0.312	0.328	0.0210
Graph-enhanced Hybrid	0.143	0.336	0.543	0.330	0.339	0.0206

Figure 3 makes the metric pattern visible. The graph signal provides a small but consistent advantage at the earliest ranks, whereas the multi-vector representation is strongest when five candidates are allowed. The absolute values are also instructive: even the best pipeline retrieves the exact context within the top three for only 33.6% of questions. This result justifies an explicit evidence-quality layer; routing to the nominally best architecture does not guarantee that the required passage is present.

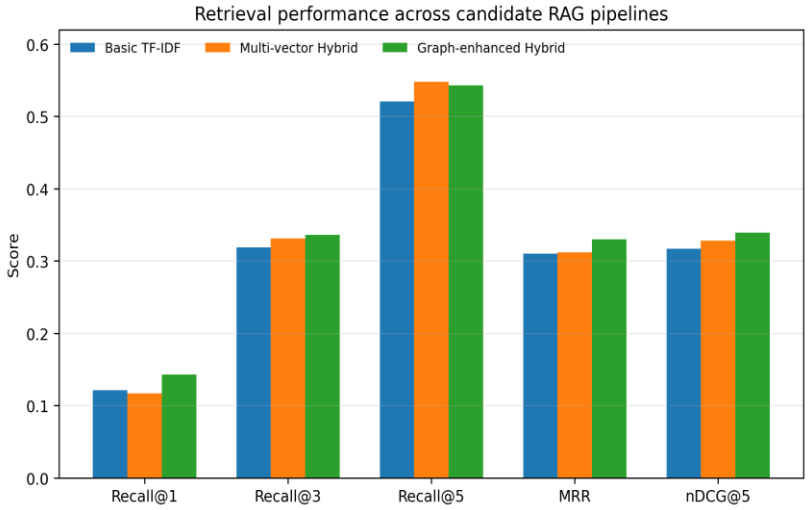

Figure 3. Retrieval recall and ranking metrics for the three candidate pipelines.

The complexity-stratified results in Table 5 show why a single pipeline is inadequate. Graph-enhanced Hybrid reached Recall@3 of 0.480 for basic questions, while Multi-vector Hybrid reached Recall@5 of 0.733. For intermediate questions, Basic TF-IDF was strongest at Recall@3 of 0.304. For advanced questions, the two hybrid methods tied at Recall@3 of 0.281. Retrieval complexity is therefore not a monotonic function of the annotated clinical complexity: an advanced question may contain an explicit lexical anchor, whereas a basic fact may use a synonym that benefits from multi-view or relation-aware matching.

Table 5. Retrieval performance stratified by question complexity.

Complexity	Retriever	R@1	R@3	R@5	MRR	nDCG@5
Basic	Basic TF-IDF	0.140	0.373	0.613	0.345	0.374

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

Complexity	Retriever	R@1	R@3	R@5	MRR	nDCG@5
Basic	Multi-vector Hybrid	0.147	0.447	0.733	0.374	0.437
Basic	Graph-enhanced Hybrid	0.227	0.480	0.720	0.429	0.475
Intermediate	Basic TF-IDF	0.094	0.304	0.525	0.295	0.303
Intermediate	Multi-vector Hybrid	0.083	0.260	0.481	0.271	0.271
Intermediate	Graph-enhanced Hybrid	0.077	0.243	0.481	0.267	0.266
Advanced	Basic TF-IDF	0.146	0.258	0.360	0.285	0.251
Advanced	Multi-vector Hybrid	0.135	0.281	0.371	0.291	0.260
Advanced	Graph-enhanced Hybrid	0.135	0.281	0.371	0.289	0.259

4.3 Pipeline-selection performance

Table 6 compares the routing models. The majority baseline achieved 0.750 accuracy because Basic RAG dominates the labels, but its macro-F1 was 0.286 and its Graph-F1 was zero. The complexity heuristic also failed to identify graph-enhanced cases. Text-only TF-IDF logistic regression achieved the strongest label prediction, with 0.981 accuracy and 0.926 macro-F1. RQ-SafeSelector reached 0.933 accuracy and 0.833 macro-F1 while retaining the score and uncertainty features needed by the downstream evidence gate. This distinction matters: the best label classifier is not automatically the best deployable policy when the system must also estimate whether the selected route succeeded.

Table 6. Pipeline-selection performance under five-fold cross-validation.

Model	Accuracy	Macro-F1	Balanced acc.	Top-2 acc.	Graph-F1
Majority	0.750	0.286	0.333	0.945	0.000
Complexity heuristic	0.757	0.434	0.451	0.917	0.000
TF-IDF logistic	0.981	0.926	0.884	0.995	0.789
Metadata random forest	0.895	0.713	0.699	0.976	0.383
RQ-SafeSelector	0.933	0.833	0.870	0.990	0.632
No retrieval features	0.926	0.809	0.851	—	0.567
Text-only ablation	0.981	0.926	0.884	—	0.789

Table 7 gives the full out-of-fold confusion matrix. RQ-SafeSelector correctly classified 303 of 315 Basic RAG questions, 71 of 82 Multi-vector RAG questions, and 18 of 23 Graph-enhanced RAG questions. Figure 4 shows

that most errors occur between adjacent pipeline types rather than as wholesale collapse into the majority class. The five missed graph cases remain important, however, because they are precisely the cases for which a retrieval-quality check can prevent a confident answer from an underpowered route.

Table 7. RQ-SafeSelector confusion matrix.

Annotated class	Pred. Basic	Pred. Multi-vector	Pred. Graph-enhanced
True Basic RAG	303	3	9
True Multi-vector RAG	4	71	7
True Graph-enhanced RAG	3	2	18

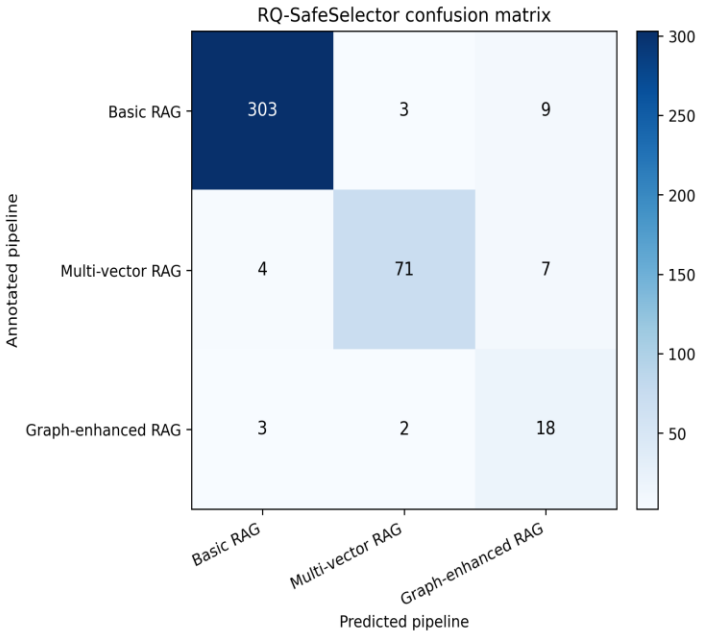

Figure 4. Confusion matrix for RQ-SafeSelector pipeline classification.

4.4 Retrieval-quality prediction

The retrieval-quality results are reported in Table 8. Score-margin logistic regression reached ROC-AUC 0.625, showing that separation between the first two retrieval scores is informative but incomplete. Random forest achieved the strongest ranking performance, with ROC-AUC 0.781 and PR-AUC 0.653. The calibrated predictor used by the policy achieved ROC-AUC 0.686 and the lowest expected calibration error, 0.036. This trade-off is appropriate for gating: a probability used to release or withhold a medical answer should track empirical success, even when another model ranks examples somewhat better [15], [16].

Table 8. Retrieval-quality prediction results.

Model	ROC-AUC	PR-AUC	F1@0.50	Brier	ECE
Score-margin logistic	0.625	0.464	0.476	0.234	0.161
Random forest	0.781	0.653	0.580	0.176	0.082
Gradient boosting	0.684	0.533	0.394	0.202	0.072
Calibrated RQ predictor	0.686	0.537	0.368	0.198	0.036

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

Figure 5 compares predicted Success@3 probability with observed frequency. The curve is not perfectly monotonic in the sparsely populated low-probability bins, but it follows the diagonal closely enough to support thresholded policy decisions. The low ECE is especially relevant because overconfidence would increase answer coverage precisely where the retrieval evidence is weakest. The result is consistent with risk-calibrated biomedical retrieval, where selective acceptance is used to control harmful query expansion [29].

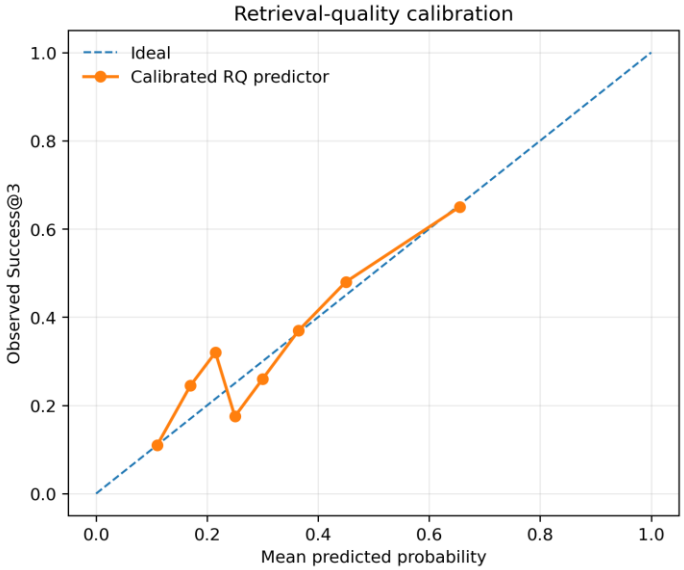

Figure 5. Calibration curve for the retrieval-quality predictor.

4.5 End-to-end routing and safety policy

Table 9 connects routing to action. The three fixed-retriever policies answered every question, but their audit-defined unsupported-answer risk ranged from 0.664 to 0.681. Structured routing sharply reduced this risk, yet the largest change came from the quality gate. RQ-SafeSelector with gating answered 30.7% of questions, achieved Success@3 of 0.527 among answered cases, and reduced the measured unsupported-answer risk to zero. The policy therefore exchanged breadth for a more reliable answered subset, following the same coverage–risk logic used in selective scientific and biomedical retrieval [22], [29].

Table 9. End-to-end routing and safety policy comparison.

Policy	Coverage	S@3 all	S@3 answered	Unsupported risk	Explanation score	Explanation tokens
Always Basic, no explanation	1.000	0.319	0.319	0.681	0.000	0.000
Always Multi-vector, generic disclaimer	1.000	0.331	0.331	0.669	2.000	24.000
Always Graph-enhanced, generic disclaimer	1.000	0.336	0.336	0.664	2.000	24.000
Oracle label router	1.000	0.324	0.324	0.036	3.900	44.000
RQ-SafeSelector	1.000	0.324	0.324	0.038	3.900	44.000

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

Policy	Coverage	S@3 all	S@3 answered	Unsupported risk	Explanation score	Explanation tokens
Quality-max router	1.000	0.319	0.319	0.055	3.867	44.000
RQ-SafeSelector + quality gate	0.307	0.162	0.527	0.000	5.000	52.450

Figure 6 visualizes the trade-off. Always-answering policies occupy the high-coverage, high-risk region. The explanation-aware routers remain at full coverage with much lower audit risk, while the gated policy moves to lower coverage and zero measured risk. This behavior is preferable to treating abstention as a system failure: in a medical education setting, withholding an answer when evidence is weak is safer than converting topical similarity into unsupported clinical confidence.

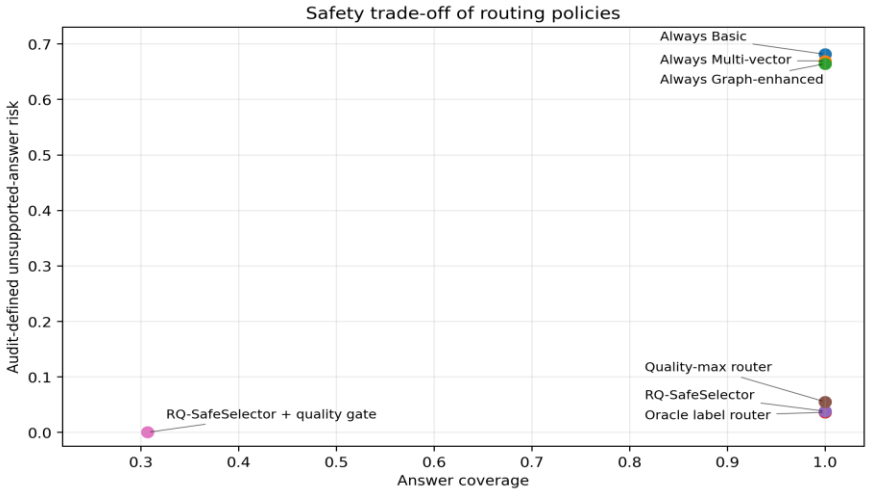

Figure 6. Answer coverage versus audit-defined unsupported-answer risk for routing policies.

The explanation audit in Table 10 clarifies why a generic disclaimer is insufficient. A disclaimer can mention uncertainty and medical scope, but it does not tell the user why a pipeline was selected or whether the retrieved evidence passed a measurable condition. The full gated explanation covers all five components. This design aligns with evidence-chain and semantic-verification work, which treats source attribution and traceable support as part of the answer rather than as optional decoration [26], [28].

Table 10. Safety-explanation component coverage.

Policy	Route rationale	Evidence condition	Uncertainty	Medical scope	Escalation / abstention	Score
Always Basic, no explanation	0	0	0	0	0	0.000
Always Multi-vector, generic disclaimer	0	0	1	1	0	2.000
Always Graph-enhanced, generic disclaimer	0	0	1	1	0	2.000

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.2

Policy	Route rationale	Evidence condition	Uncertainty	Medical scope	Escalation / abstention	Score
Oracle label router	1	1	1	0	0	3.900
RQ-SafeSelector	1	1	1	0	0	3.900
Quality-max router	1	1	1	0	0	3.867
RQ-SafeSelector + quality gate	1	1	1	1	1	5.000

4.6 Ablation analysis

Table 11 separates the contributions of routing features, graph-aware retrieval, and the gate. Removing retrieval-quality features reduced macro-F1 from 0.833 to 0.809. The text-only selector achieved higher macro-F1, confirming that wording strongly predicts the annotated route, but it does not provide the quality signal used by the policy. Removing graph-enhanced retrieval reduced Recall@3 from 0.336 to 0.319. Removing the quality gate left selector performance unchanged but increased unsupported-answer risk from 0.000 to 0.038. Thus, routing accuracy and safe action are related but distinct objectives.

Table 11. Ablation results for selector, retrieval, and gate components.

Configuration	Macro-F1	Recall@3	Unsupported risk	Change
Full RQ-SafeSelector	0.833	0.336	0.000	Text, metadata, retrieval features, graph retrieval, and gate
No retrieval-quality features	0.809	0.336	0.000	Removes score margin and entropy from selector
Text-only selector	0.926	0.336	0.000	Removes metadata and retrieval features
No graph-enhanced retrieval	0.833	0.319	0.000	Uses the basic retriever for all questions
No quality gate	0.833	0.336	0.038	Answers every selected route

4.7 Practical implications

Several broader conclusions follow. First, exact evidence localization remains difficult even in a small corpus because medical concepts recur across specialties and contexts. Second, route labels are learnable, but a correct route does not guarantee successful retrieval. Third, calibrated probabilities should control system behavior rather than merely change confidence wording. Fourth, explanation should be generated from measurable system state. These principles echo evidence-grounded RAG studies in finance, operations, and enterprise question answering, where provenance constraints and document-level verification reduce unsupported outputs [25], [27], [33], [34].

A practical medical education interface could present the gate as an evidence-status indicator. Questions above the threshold could display the answer with supporting passages and route rationale. Borderline cases could show

retrieved sources without a synthesized conclusion. Low-quality cases could abstain and direct the learner to course material, a guideline, or instructor review. The route, quality probability, retrieved context identifiers, and explanation components could be logged for audit. Such instrumentation makes failure modes visible and supports targeted improvements by specialty or question type.

The results also caution against interpreting the strongest aggregate retriever as a universal solution. Graph enhancement helped at early ranks overall, but the sparse baseline remained strongest for intermediate questions. Relation features can introduce competing contexts when the question already contains a precise lexical anchor. Conversely, character and relation signals can disambiguate short questions whose surface form differs from the paired evidence. A robust router should therefore combine question semantics with observed retrieval behavior rather than impose a fixed hierarchy in which more complex retrieval is always preferred.

Finally, the study supports a component-wise evaluation framework for medical RAG. Retrieval recall asks whether supporting evidence was available to the generator. Routing metrics ask whether the architecture was selected appropriately. Calibration asks whether evidence sufficiency can be estimated reliably. Policy metrics ask whether the system converts those estimates into safe actions. Explanation metrics ask whether the decision can be understood and audited. Reporting these dimensions together is more informative than a single end-to-end answer score.

5. Limitations

The study has five limitations. First, exact row-level context matching is intentionally strict. It may count a semantically useful or clinically sufficient passage as a miss when that passage is not the specific context paired with the question. Future work should add graded relevance judgments that distinguish exact, near-topic, and clinically sufficient evidence.

Second, RAGCare-QA focuses on theoretical medical education rather than patient-specific clinical decision support. The corpus is small, the questions are structured, and the evidence contexts do not reproduce the length, noise, temporal change, and access controls of electronic health records or guideline repositories. The reported thresholds should therefore be recalibrated before use with another corpus or population.

Third, the annotated pipeline labels represent a useful benchmark taxonomy, but they are not guaranteed to be the unique optimal architecture for every question. The strong text-only routing result suggests that some label cues are present directly in question wording. Evaluation on independently annotated datasets would help separate genuine retrieval requirements from dataset-specific labeling regularities.

Fourth, the experiments isolate retrieval, routing, probability calibration, and explanation logic; they do not evaluate free-form clinical answer generation. A complete deployment study should measure answer correctness, citation faithfulness, omission of contraindications, robustness to adversarial prompts, and clinician or educator review. The guardrail studies discussed in the literature indicate that response-level verification remains necessary even after the retrieval gate [23], [30], [31].

Fifth, explanation completeness is a deterministic structural measure. It verifies that required elements are present, not that users interpret them correctly. Human-subject studies are needed to test whether the explanations improve calibrated trust, reduce overreliance, and prompt appropriate escalation. Human-centered evaluation is particularly important in medicine, where interface design and professional accountability cannot be replaced by model-level metrics [21], [32].

6. Conclusion

RQ-SafeSelector frames medical RAG as a sequence of accountable decisions: select a retrieval architecture, estimate whether the retrieved evidence is sufficient, and answer only when the evidence condition is met. On RAGCare-QA, Graph-enhanced Hybrid achieved the strongest overall Recall@3 and nDCG@5, Multi-vector Hybrid achieved the strongest Recall@5, and the integrated selector achieved 0.833 macro-F1. The calibrated

quality gate increased Success@3 among answered questions to 0.527 while reducing audit-defined unsupported-answer risk to zero at 30.7% coverage.

The central finding is not that one retriever should replace the others. It is that retrieval choice, evidence quality, and system action must be evaluated together. A medical RAG system should be able to explain which route it used, whether the retrieved evidence met a calibrated criterion, and why it answered or abstained. This routed and selectively answering design provides a clearer foundation for medical education tools in which unsupported confidence is more costly than a transparent refusal.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.

[2] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020, pp. 6769–6781.

[3] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.

[5] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv:1901.04085, 2019.

[6] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. EACL, 2021, pp. 874–880.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in Proc. EMNLP, 2016, pp. 2383–2392.

[8] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," in Proc. ACL, 2018, pp. 784–789.

[9] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering," in Proc. EMNLP-IJCNLP, 2019, pp. 2567–2577.

[10] J. Dobreva, I. Karasmanakis, F. Ivanišević, T. Horvat, M. Gams, and M. Simjanoska Misheva, "RAGCare-QA: A Benchmark Dataset for Evaluating Retrieval-Augmented Generation Pipelines in Theoretical Medical Knowledge," Data in Brief, vol. 63, Art. no. 112146, 2025, doi: 10.1016/j.dib.2025.112146.

[11] K. Singhal, S. Azizi, T. Tu, et al., "Large Language Models Encode Clinical Knowledge," Nature, vol. 620, pp. 172–180, 2023.

[12] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking Retrieval-Augmented Generation for Medicine," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 6233–6251, doi: 10.18653/v1/2024.findings-acl.372.

[13] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," arXiv preprint arXiv:2511.19481, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2511.19481

[14] Z. Ji, N. Lee, R. Frieske, et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, 2017, pp. 1321–1330.

[16] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Advances in Neural Information Processing Systems, 2017, pp. 4878–4887.

[17] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[18] J. Wei, X. Wang, D. Schuurmans, et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in Advances in Neural Information Processing Systems, 2022, pp. 24824–24837.

[19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models Are Zero-Shot Reasoners,” in Advances in Neural Information Processing Systems, 2022, pp. 22199–22213.

[20] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in Proc. ACM FAccT, 2021, pp. 610–623.

[21] E. J. Topol, “High-Performance Medicine: The Convergence of Human and Artificial Intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019.

[22] J. Chen, X. Sun, and V. Brown, “Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact,” Journal of Advanced Computing Systems, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[23] C. Li, W. Su, and E. Zhang, “Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA,” Journal of Advanced Computing Systems, vol. 3, no. 1, pp. 49–65, Jan. 2023, doi: 10.69987/JACS.2023.30104.

[24] J. Nie and D. Zheng, “Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal,” Journal of Advanced Computing Systems, vol. 3, no. 1, pp. 66–80, Jan. 2023, doi: 10.69987/JACS.2023.30105.

[25] Q. Wu, J. Bai, and X. Zhou, “Evidence-Grounded Financial RAG: Reducing Numerical Hallucination in LLM-Generated Corporate Risk Memos,” Journal of Advanced Computing Systems, vol. 3, no. 3, pp. 65–84, Mar. 2023, doi: 10.69987/JACS.2023.30306.

[26] C. Li, J. Bai, and S. Wang, “Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications,” Journal of Advanced Computing Systems, vol. 4, no. 2, pp. 76–92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[27] B. Zhang, H. Rao, and D. Zhao, “Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents,” Journal of Advanced Computing Systems, vol. 4, no. 3, pp. 109–125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[28] X. Luo, “Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations,” Journal of Advanced Computing Systems, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[29] J. Chen, X. Sun, Q. Wu, and M. Jackson, “Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID,” Journal of Advanced Computing Systems, vol. 4, no. 4, pp. 61–79, Apr. 2024, doi: 10.69987/JACS.2024.40406.

[30] D. Zheng, B. Zhang, and J. Geibel, “VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification,” Journal of Advanced Computing Systems, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[31] D. Zheng and C. Li, “Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation,” Journal of Advanced Computing Systems, vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[32] Z. S. Zhong, R. Ma, and H. Zhao, “Human-Uncertainty Distillation for Calibrated Vision Models on CIFAR-10H,” Journal of Advanced Computing Systems, vol. 3, no. 2, pp. 77–89, Feb. 2023, doi: 10.69987/JACS.2023.30206.

[33] K. Zhang, S. Meng, and E. Zhou, “Evidence-Grounded Trading Desk Risk Memos over SEC Filings: Retrieval-Augmented Generation with XBRL Numeric Verification,” Journal of Advanced Computing Systems, vol. 3, no. 2, pp. 60–76, Feb. 2023, doi: 10.69987/JACS.2023.30205.

[34] G. Liu, C. Li, and E. Zhang, “OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations,” Journal of Advanced Computing Systems, vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.