

Cost-aware evidence-triage routing for biomedical RAG: a leakage-controlled evaluation framework

Harish Kumar Mittal

Department of Computer Science and Engineering, University Institute of Engineering (UIE), Chandigarh University, Mohali, Punjab, India

Abstract

Background: Adaptive routing among retrieval-augmented generation (RAG) pipelines is widely presented as a way of improving biomedical question answering, but the protocols used to establish those gains seldom isolate the evaluation data from the modelling and reporting decisions that generate the estimate. Objective: To specify a leakage-controlled evaluation framework for cost-aware evidence-triage routing, to price each control, and to establish what survives once triage is restricted to information a deployed system can observe. Methods: We analysed a fully crossed outcome matrix of 200 biomedical questions and 11 RAG actions (2,200 outcomes) over a 1,000-abstract PubMed corpus. Nine controls governed the estimate; five nested question-level folds used a 120/40/40 partition; inference used 5,000-resample bootstrap intervals and Holm-adjusted paired tests. A leakage ladder re-estimated the same contrast with one control removed at a time. Results: Under all controls, routing did not improve mean answer correctness over the training-locked strongest fixed action (0.806 versus 0.817; difference -0.010 (-0.030 to 0.008); Holm-adjusted $p = 0.28$), and its discretionary decisions were net harmful (5 beneficial, 10 harmful, 7 tied). Removing one control at a time raised the apparent advantage by as much as +0.148; removing three together produced 0.228 (0.190 to 0.269), an artefact more than twenty times the magnitude of the controlled estimate and of the opposite sign. Restricting triage to retrieval-observable features left correctness unchanged at 0.811 (difference -0.006, 95% CI -0.016 to 0.000) while removing the 53% token penalty of the generation-based probe and reducing expected calibration error from 0.040 to 0.023. Triage confidence supported selective answering (risk-coverage area 0.115 versus 0.143 for a question-only router), with overlapping intervals. Conclusions: On this benchmark, cost-aware evidence triage buys efficiency and calibrated abstention rather than accuracy, and much of the accuracy advantage attributed to such routers is recoverable from protocol choices alone. The ladder is released as a reusable audit for adaptive-RAG evaluations.

Keywords: Retrieval-augmented generation; Biomedical question answering; Adaptive routing; Data leakage; Selective prediction; Evaluation methodology

Highlights

- Nine explicit controls define a leakage-controlled protocol for adaptive RAG routing.
- Fully controlled, routing does not beat the strongest fixed biomedical RAG pipeline.
- Dropping three common shortcuts manufactures a large advantage on identical data.
- Retrieval-only triage removes the token penalty and halves calibration error.
- Gains from evidence triage lie in cost and abstention, not in answer correctness.

1. Introduction

Retrieval-augmented generation grounds a language model's output in documents fetched at query time, and has become the default architecture for biomedical question answering because it allows a general model to cite current, verifiable literature rather than recall it [1,3]. The design space is unusually wide: a biomedical RAG pipeline must choose a retrieval depth, a fusion or reranking strategy, whether to rewrite or decompose the question, whether to grade evidence before generating, and whether to run a critic over the draft. Systematic comparisons show that these choices matter and that no single configuration dominates across question types [4,5,6]. Sensitivity to the position and quantity of retrieved evidence compounds the problem, since adding passages can displace the one that carries the answer [36].

The natural response is to route: inspect the question and its retrieved evidence, then select the pipeline expected to answer it best. Adaptive and corrective RAG systems do exactly this, using question complexity, self-reflection or a retrieval-quality critic to decide how much machinery to spend [10,11,12], and recent work casts the decision as a bandit problem or drives it from the question alone [13,14]. The appeal for clinical settings is obvious: most questions are easy and should be answered cheaply; a minority are hard and deserve the expensive pipeline.

That appeal creates a measurement problem which this paper is about. Given a fully crossed matrix of 200 questions and 11 candidate pipelines, an oracle that knows each answer's score in advance reaches 0.940 mean correctness against 0.817 for the best single pipeline. Roughly 12 percentage points of headroom sit on the table, and a router is scored by how much of it the analyst's protocol appears to recover. Almost every step in that protocol admits a shortcut that moves the estimate upward: fitting the scaler or vectoriser before splitting, tuning a decision threshold on the questions used to report it, comparing against a baseline chosen a priori rather than the strongest available alternative, or reporting the most favourable of several folds or variants. These are the classical leakage and selection-bias failures catalogued for data mining and, more recently, for machine-learning-based science as a whole [22,23,24]; the language-

processing literature documents the same problem as split design and train-test overlap [25,26,27]. What has not been done, to our knowledge, is to price these shortcuts inside a single adaptive-RAG study, on one dataset, with everything else held fixed.

We do that here. Our contribution is fourfold. First, we specify a leakage-controlled evaluation framework for cost-aware evidence-triage routing: nine named controls, each attached to the stage of the pipeline it governs (Fig. 1). Second, we build a leakage ladder that re-estimates the router-versus-fixed-pipeline contrast with exactly one control removed at a time, so the inflation attributable to each shortcut is measured rather than asserted. Third, we stratify triage features by provenance — question text, retrieved passages, probe generation, judge scores and gold annotation — and show that a router restricted to the first two strata is generation-free, judge-independent, token-neutral and better calibrated than one using all five. Fourth, we report the result the controlled protocol supports: routing does not improve correctness, its discretionary decisions are net harmful, and its value lies in cost and calibrated selective answering.

2. Related work

Biomedical RAG benchmarks. Systematic benchmarking established that retrieval configuration, corpus and reranking interact strongly on medical question answering, and that gains do not transfer uniformly across datasets [4]. Domain-adapted pipelines report improvements on expert-labelled sets [7,8,9]. A systematic review and meta-analysis found wide heterogeneity in evaluation practice and called for clinical development guidelines [5]; a large expert evaluation reached a similar conclusion, reporting that automatic benchmark gains frequently fail to survive systematic clinician review [6]. Our study sits in that tradition and takes its questions from PubMedQA [2] over an abstract corpus assembled from PubMed [32].

Adaptive and corrective routing. Adaptive-RAG trains a complexity classifier to choose between no retrieval, single-step and multi-step pipelines [10]. Self-RAG learns to retrieve and critique through reflection tokens [11], and corrective RAG adds a retrieval evaluator that triggers re-search when the evidence looks weak [12]. MBA-RAG frames the same selection as a multi-armed bandit [13], and question-only adaptive routing removes the language model from the decision entirely [14]. All report accuracy or efficiency gains over fixed pipelines. Our framework does not dispute those results; it asks what part of a gain of that shape survives once the protocol is fully specified, and supplies an instrument for asking that question of any comparable outcome matrix.

Automated evaluation and its circularity. Reference-free frameworks such as RAGAS and ARES score faithfulness, relevancy and context quality with model judges [15,16], and diagnostic frameworks decompose errors to retriever and generator [18]. Judge reliability is itself an active

concern: model judges exhibit position and verbosity biases [17], and hallucination corpora and multilingual meta-evaluations show that agreement with human raters varies by domain [19,21]. This matters here in a specific way: two candidate triage features in the published router are relevancy and faithfulness scores produced by the same judge that supplies the training labels [20], so the router could learn the judge rather than the task. We treat that as a leakage channel and test it directly.

Selective prediction. The option to abstain rather than answer is the classical reject-option setting [28,29], evaluated through risk-coverage curves and calibration error [30], and extended to distribution-free guarantees by conformal risk control [31]. For clinical deployment this is arguably the more useful capability: a system that knows when its answer is unreliable is safer than one that is marginally more accurate on average. We therefore report risk-coverage and calibration alongside correctness, and find that this is where evidence triage earns its keep.

3. Materials and methods

3.1 Corpus, questions and the action matrix

The knowledge base comprises 1,000 PubMed abstracts retrieved through NCBI E-utilities [32] and indexed with MiniLM sentence embeddings [33,34]. The evaluation set contains 200 questions: 150 derived from PubMedQA [2] and 50 constructed to stress multi-hop, comparative, causal and specificity reasoning over the same corpus. Questions carry a query-type label (150 single-hop, 20 multi-hop, 14 specificity, 11 comparative and 5 causal) and a three-level clinical-severity tier assigned from the question text by a keyword classifier with a language-model fallback (122 medium, 56 high, 22 low).

Every question was answered by all 11 candidate pipelines, giving a fully crossed matrix of 2,200 question-action outcomes with no missing cells for the primary outcome. The actions span retrieval depth (naive retrieval at $k = 3, 5$ and 8), query transformation (expansion, multi-query expansion, decomposition), fusion and reranking (reciprocal rank fusion [35], cross-encoder reranking), domain structure (MeSH-guided retrieval), evidence grading before generation, and a critic pass over the draft answer. Two closed-book configurations were run as a floor but were excluded from the action set because they perform no retrieval. Each cell records an answer-correctness score from an LLM-judge harness [20], end-to-end latency and total token consumption. Table 1 gives the marginal profile of each action.

Table 1. Marginal profile of the candidate action set. Each row is the mean over all 200 questions; every action was run on every question, so the columns are directly comparable. Closed-book configurations are reported as a no-retrieval reference and were not available to the router. Latency is end to end and includes retrieval, transformation and generation.

Action	Answer correctness	Latency (s)	Tokens
Evidence-graded RAG	0.817	14.11	2,156
Query decomposition	0.748	3.76	2,089
Multi-query expansion	0.738	5.43	2,049
Hybrid cross-encoder	0.735	7.56	1,805
Naive RAG k=5	0.715	2.53	1,844
Hybrid RRF	0.696	4.24	1,805
Query expansion	0.679	6.00	1,853
MeSH-guided RAG	0.670	2.47	1,691
Naive RAG k=3	0.668	0.62	1,172
Naive RAG k=8	0.668	7.31	2,800
Naive RAG with critic	0.660	2.86	1,900
Closed-book LLM (reference only)	0.342	—	—
Closed-book LLM with critic (reference only)	0.321	—	—

Two features of Table 1 shape everything that follows. Evidence-graded RAG is the strongest single action at 0.816, and it is also among the most expensive at 14.11 s. The cheapest action, naive retrieval at $k = 3$, answers in well under a second but is 0.148 lower in correctness. Any router therefore faces a genuine trade-off rather than a free lunch, and a closed-book model reaches only 0.342, confirming that retrieval, not parametric recall, is doing the work.

3.2 Routing formulation

Let a question q have a triage state $\phi(q)$ and let each action a in the action set A have an observed outcome $y(q, a)$. The router fits a multi-output action-value model $\hat{y}(\phi(q), \cdot)$ to the training partition and selects an action at test time. We evaluate two decision rules. The accuracy-first rule takes the arg-max of the predicted outcome. The cost-aware rule admits every action whose predicted outcome lies within a tolerance δ of the maximum and, among those, takes the cheapest by expected latency; δ is chosen on a calibration partition as the smallest-cost value whose calibration-set correctness is not below the fixed comparator. The action-value model is an extremely randomised trees ensemble (600 trees, minimum leaf size 4, feature subsample 0.7)

fitted on standardised tabular features concatenated with a 25-component truncated SVD projection of a TF-IDF encoding of the question. The selected action's predicted outcome, clipped to the unit interval, is retained as the router's confidence and is what drives selective answering.

Three reference policies bound the result. Release-only always runs the cheapest action. The best fixed action is the single pipeline with the highest mean training-fold outcome, re-selected inside each fold. The per-question oracle chooses the best action using the observed outcome and is not attainable; it is reported only to size the headroom.

3.3 Triage features and provenance strata

Features fall into five strata according to what a deployed system would have to do to obtain them (Fig. 1A). Question-text features (length in tokens and characters, presence of numerals, TF-IDF projection) are free. Retrieval features (passage count, retrieved token volume, question-evidence lexical overlap) require the retrieval step that every action performs anyway and consume no generation tokens. Probe-generation features (answer length, answer-evidence and answer-question overlap, probe latency and probe tokens) require running a complete generation before routing. Judge features (answer relevancy and faithfulness) are produced by the same evaluator that supplies the training labels. Annotation features (the number of source PMIDs required, the query-type label) are gold metadata that no deployed system observes.

We evaluate four routers over these strata. CET-A uses all five and reproduces the published configuration. CET-J removes the judge stratum. CET-D additionally removes the annotation stratum, leaving only features a deployed system could compute. CET-R keeps only the question-text and retrieval strata, so the triage probe issues no generation call at all; CET-R-E is the same feature set under the cost-aware decision rule. The severity tier is retained throughout because it is derived from the question text by a classifier and is therefore observable at inference. Two published ablations are carried forward for comparability: a question-only router that drops all evidence-derived features, and a probe-only router that drops the question-text projection.

3.4 The leakage-control framework

We define leakage broadly, following the data-mining formulation [22], as any flow of information from the evaluation partition into a modelling or reporting decision. Nine controls close the channels available in this study (Table 2). Seven of them can be removed one at a time while holding data, learner and seed fixed, which is what makes the ladder possible; the remaining two — feature provenance and cluster grouping — are structural and are evaluated by refitting rather than by re-reporting.

Table 2. The nine controls. The final column is the increase in the apparent routing advantage, in units of mean answer correctness, observed when that control alone is removed and everything else is held fixed; the full ladder with

intervals is given in Table 4. Controls C8 and C9 are structural and are examined by refitting the router rather than by re-reporting the same fit.

Control	Channel it closes	Implementation in this study	Inflation if removed
C1 Split isolation	Model fitted on all questions, then evaluated on the same questions	Five outer folds; the test partition is never seen by the fitting code	+0.091
C2 Preprocessing isolation	Scaler and text vectoriser see the test questions before the split	Standardisation and TF-IDF/SVD fitted on the training partition only	+0.002
C3 Comparator isolation	Router compared with a weak baseline fixed before the data are seen	The comparator is the strongest action on each training fold, re-selected per fold	+0.148
C4 Tolerance isolation	Cost tolerance tuned on the questions used to report the result	A 40-question calibration partition, disjoint from training and test, sets δ	+0.008
C5 Variant isolation	The best of several router variants chosen on test outcomes	All router variants pre-specified; every variant is reported	+0.003
C6 Multiplicity control	The most favourable fold or subgroup reported as the result	All five folds pooled; families of tests Holm-adjusted [39]	+0.036
C7 Outcome isolation	Action chosen with knowledge of the outcome it produces	Selection uses predictions only; the oracle is reported as a bound, not a system	+0.134
C8 Feature provenance	Features that require gold annotation or the label judge	Strata declared and ablated; CET-R uses only inference-observable signals	See §4.4
C9 Cluster grouping	Questions sharing a source article split across training and test	Sensitivity analysis with folds formed from PMID-connected components	See §4.6

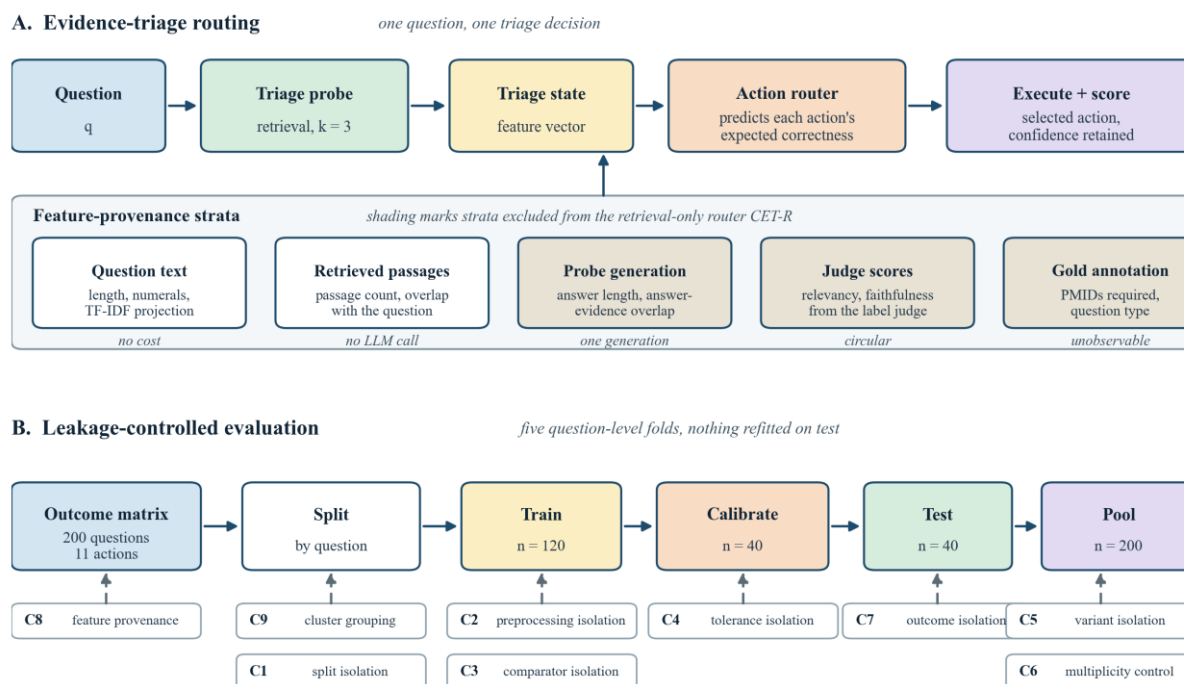


Fig. 1. The framework. (A) Routing pipeline: a question passes through a triage probe, a triage state and an action router that predicts the expected correctness of each of the 11 actions, one of which is executed and scored. The five feature-provenance strata beneath differ in what they cost and in what they presuppose; shaded strata are excluded from the retrieval-only router CET-R. (B) Evaluation pipeline: the 200×11 outcome matrix is split by question into five folds of 120 training, 40 calibration and 40 test items, and out-of-fold predictions are pooled to $n = 200$. Each control is tagged at the stage it governs.

3.5 Cost accounting

Cost is charged honestly against the probe each router requires. A router whose triage state depends on probe-generation or judge features must run the probe pipeline before it can decide, so when it selects a different action the question is charged for both: probe tokens plus selected-action tokens, and probe latency plus selected-action latency. A router restricted to question-text and retrieval features issues no generation call during triage, so it is charged for the selected action alone. Retrieval itself consumes no language-model tokens, making the token accounting for CET-R exact; its reported latency is a lower bound, because the probe's retrieval step is real but was not timed separately in the original runs. We state this asymmetry rather than hide it, and the upper bound for CET-R latency is the accounting used for CET-A.

3.6 Statistical analysis

All estimates are pooled out-of-fold over the full set of $n = 200$ questions. Means carry percentile bootstrap 95% confidence intervals from 5,000 resamples [37]. Paired comparisons between a router and its comparator use the per-question difference, with bootstrap intervals on the mean

difference and two-sided Wilcoxon signed-rank tests [38]; families of comparisons are Holm-adjusted [39]. Selective performance is summarised by the area under the risk-coverage curve, by selective risk at fixed coverage, and by expected calibration error over eight equal-width confidence bins [30]. A cluster bootstrap over the 187 connected components of the question-PMID graph is reported as a sensitivity analysis for non-independence. The learner seed is 20260719 and the analysis is deterministic; the leakage ladder reproduces the published primary contrast exactly at its fully controlled rung, which is the check that the ladder differs from the original analysis only in the controls it removes. No non-inferiority margin was prespecified, so we do not claim equivalence anywhere.

4. Results

4.1 Routing under full control

The fold-locked comparator was Evidence-graded RAG in all five folds, so the router was measured against the strongest available fixed pipeline rather than a convenient one. Under the complete control set, CET-A reached 0.806 mean answer correctness against 0.817 for that comparator, a paired difference of -0.010 (-0.030 to 0.008) with Holm-adjusted $p = 0.28$ (Table 3). The interval spans zero, and even at its most favourable end the advantage would be less than a fifth of the gap between the best and second-best fixed action in Table 1 (0.069). Routing did not improve correctness on this benchmark.

The cost picture is mixed and, for the published configuration, unfavourable. CET-A reduced mean latency from 14.11 s to 13.51 s, a 4% saving that survives adjustment (difference -0.60 s, 95% CI -1.18 to -0.09; Holm $p < 0.001$), but it raised mean token consumption from 2,156 to 3,307, an increase of 1,151 tokens or 53% (Holm $p < 0.001$). The cause is structural rather than incidental: the triage probe is a complete generation that is paid on every question, and on the 89% of questions where the router then selects the comparator action anyway, the question is billed twice. Section 4.3 shows what happens when that probe is removed.

Table 3. Primary out-of-fold operating points, pooled over $n = 200$ questions. Confidence intervals are percentile bootstrap over 5,000 resamples; the final column is the paired per-question difference against the fold-locked best fixed action, which was Evidence-graded RAG in every fold. CET-A and the fixed comparator are charged for the triage probe; CET-R and CET-R-E require no generation during triage and are charged for the selected action alone, with latency reported as a lower bound (§3.5). The oracle is shown to size the headroom and is not a deployable policy.

Policy	Answer correctness (95% CI)	Latency (s)	Tokens	Difference vs fixed action (95% CI)
Release-only RAG (cheapest action)	0.668 (0.623 to 0.713)	0.62	1,172	—

Policy	Answer correctness (95% CI)	Latency (s)	Tokens	Difference vs fixed action (95% CI)
Best fixed action (fold-locked)	0.817 (0.777 to 0.854)	14.11	2,156	reference
CET-A (all five feature strata)	0.806 (0.766 to 0.845)	13.51	3,307	-0.010 (-0.030 to 0.008)
CET-R (retrieval-only triage)	0.811 (0.770 to 0.850)	13.83	2,152	-0.006 (-0.016 to 0.000)
CET-R-E (retrieval-only, cost-aware)	0.770 (0.726 to 0.812)	8.70	2,061	-0.046 (-0.078 to -0.014)
Per-question oracle (not attainable)	0.940 (0.916 to 0.960)	—	—	+0.123

4.2 The leakage ladder

We then re-estimated the same contrast eight more times, each time removing a single control and changing nothing else (Table 4, Fig. 2). The fully controlled rung reproduces the published estimate to four decimal places, which confirms that differences further down the ladder are attributable to the protocol and not to a re-implementation.

Table 4. The leakage ladder. Each rung re-estimates the difference in mean answer correctness between the router and its comparator on the same 200 questions, the same action matrix, the same learner and the same seed, with one control removed. Inflation is the shift relative to the fully controlled estimate C0. Positive values favour the router. L8 applies the three fitting-stage shortcuts L1, L2 and L4 together.

Protocol	What changes	Apparent advantage (95% CI)	Inflation
C0. All seven controls applied (primary estimate)	nested folds; fold-wise preprocessing; comparator locked on training data	-0.011 (-0.030 to 0.008)	—
L1. Preprocessing isolation removed	scaler and TF-IDF/SVD fitted on all 200 questions before splitting	-0.008 (-0.026 to 0.009)	+0.003
L2. Split isolation removed	router fitted on all 200 questions and evaluated on the same questions	0.081 (0.051 to 0.114)	+0.091
L3. Tolerance isolation removed	efficiency tolerance selected using test-fold outcomes	-0.002 (-0.028 to 0.025)	+0.008
L4. Comparator isolation removed	router compared with Naive RAG k=3 rather than the training-optimal action	0.138 (0.099 to 0.177)	+0.148
L5. Multiplicity control removed	only fold 0, the most favourable of five, reported	0.025 (0.000 to 0.070)	+0.035
L6. Variant isolation removed	the best-scoring of four router variants selected per fold using test outcomes	-0.007 (-0.028 to 0.011)	+0.003
L7. Outcome isolation removed	action chosen per question using the observed outcome	0.123 (0.092 to 0.156)	+0.134

Protocol	What changes	Apparent advantage (95% CI)	Inflation
L8. Composite of L1, L2 and L4	the three fitting-stage shortcuts applied together	0.229 (0.190 to 0.269)	+0.239

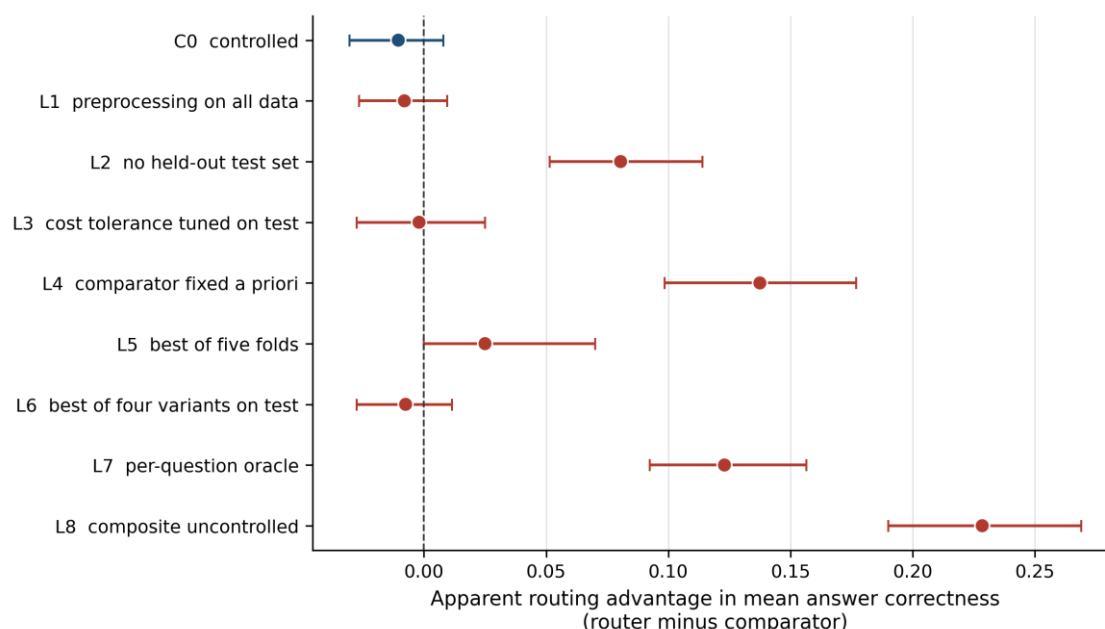


Fig. 2. Apparent routing advantage under each protocol, with bootstrap 95% confidence intervals. The fully controlled estimate C0 is shown in dark blue; every other rung removes one control (L1-L7) or three at once (L8). The dashed line marks no advantage. The vertical spread of this figure is the measurement problem the framework is designed to close: the same data support conclusions ranging from a small deficit to a large advantage depending only on the protocol.

Three shortcuts dominate. Comparing against a baseline fixed a priori rather than the strongest available action (L4) moves the estimate to 0.138 (0.098 to 0.177); this is not leakage in the strict sense, but it is the single largest source of apparent advantage and the easiest to commit unintentionally by inheriting a baseline from an earlier paper. Fitting the router on all 200 questions and evaluating in sample (L2) yields 0.080 (0.052 to 0.114), and choosing each action with knowledge of its outcome (L7), which is what an oracle-style report describes, gives 0.123 (0.092 to 0.156).

The reporting-stage shortcuts are individually smaller but not negligible at this sample size. Presenting the most favourable of five folds (L5) gives 0.025 (0.000 to 0.070) — a positive result from a study whose pooled estimate is negative, obtained with no modelling change at all. Preprocessing on the full dataset (L1, +0.002), tuning the cost tolerance on test (L3, +0.008) and selecting the best of four variants on test (L6, +0.003) each contribute little in isolation, which is

itself worth recording: the familiar advice to guard preprocessing is sound in principle but was not where the damage lay here.

Applying the three fitting-stage shortcuts together (L8) produces 0.228 (0.190 to 0.269) against a controlled estimate of -0.010. The same router, questions and outcome matrix therefore support either a clear success or a small failure, separated by +0.239 in mean correctness, purely as a function of how the evaluation is organised. This is the paper's central quantitative claim, and it generalises beyond our router: a study reporting an adaptive-RAG gain of this magnitude without specifying these controls has not distinguished the two.

4.3 Feature provenance and the cost of triage

Removing the judge stratum changed almost nothing: CET-J reached 0.807 against 0.806 for CET-A, with a risk-coverage area of 0.117 against 0.115 (Table 5). The circularity between the label judge and the triage features is therefore a real design defect but not, on this data, a load-bearing one — a useful negative finding, since the alternative explanation for the router's calibration advantage would have been that it had learned the judge. Removing the gold-annotation stratum as well (CET-D) also left correctness intact at 0.808 while making the policy more conservative: it departed from its modal action on 9 questions against 22 for CET-A.

The retrieval-only router is the substantive result of this section. CET-R reached 0.811 mean correctness, the highest of any deployable policy we evaluated and indistinguishable from the fixed comparator (difference -0.006, 95% CI -0.016 to 0.000), while consuming 2,152 tokens against 2,156 for that comparator and 3,307 for CET-A. The 53% token penalty that made the published router unattractive is therefore an artefact of the generative probe, not of routing: triage computed from retrieved passages costs no language-model tokens, so a router built on it is cost-neutral by construction (Fig. 3). Expected calibration error also fell from 0.040 to 0.023, the lowest of any variant.

The cost-aware rule buys latency at a measurable price in accuracy. CET-R-E ran in 8.70 s against 14.11 s for the fixed comparator, a 38% reduction, and used 2,061 tokens (4.4% fewer), but its correctness fell to 0.770 with a paired difference of -0.046 (-0.078 to -0.014) whose interval excludes zero. It is the only policy here that is reliably worse than the fixed pipeline, and also the only one that routes substantively, departing from its modal action on 98 of 200 questions and using 9 distinct actions. That coupling — real routing behaviour arriving with a real accuracy loss — is the mechanism behind the null result (§4.5).

Table 5. Effectiveness, cost and selective-prediction quality by feature provenance. CET-A uses all five strata; CET-J drops judge scores; CET-D additionally drops gold annotation; CET-R keeps only question-text and retrieval features and therefore needs no generation to route; CET-R-E applies the cost-aware decision rule to the same features. The last two rows are the published ablations. AUROC is the area under the risk-coverage curve and ECE the expected

calibration error over eight bins; lower is better for both. Token counts for CET-R and CET-R-E exclude probe generation because none occurs.

Router variant	Correctness	Difference vs fixed (95% CI)	Tokens	AURC (95% CI)	ECE
CET-A (published triage)	0.806	-0.011 (-0.030 to 0.008)	3,307	0.115 (0.083 to 0.151)	0.040
CET-J (judge-independent)	0.807	-0.009 (-0.029 to 0.009)	3,308	0.117 (0.084 to 0.153)	0.039
CET-D (deployable features)	0.808	-0.009 (-0.022 to 0.002)	3,324	0.116 (0.084 to 0.153)	0.035
CET-R (retrieval-only triage)	0.811	-0.006 (-0.016 to 0.000)	2,152	0.127 (0.094 to 0.166)	0.023
CET-R-E (retrieval-only, cost-aware)	0.770	-0.046 (-0.077 to -0.015)	2,061	0.173 (0.135 to 0.213)	0.031
Question-only router	0.800	-0.017 (-0.035 to -0.001)	3,308	0.143 (0.104 to 0.185)	0.039
Probe-only router	0.789	-0.028 (-0.051 to -0.005)	3,286	0.140 (0.103 to 0.184)	0.054

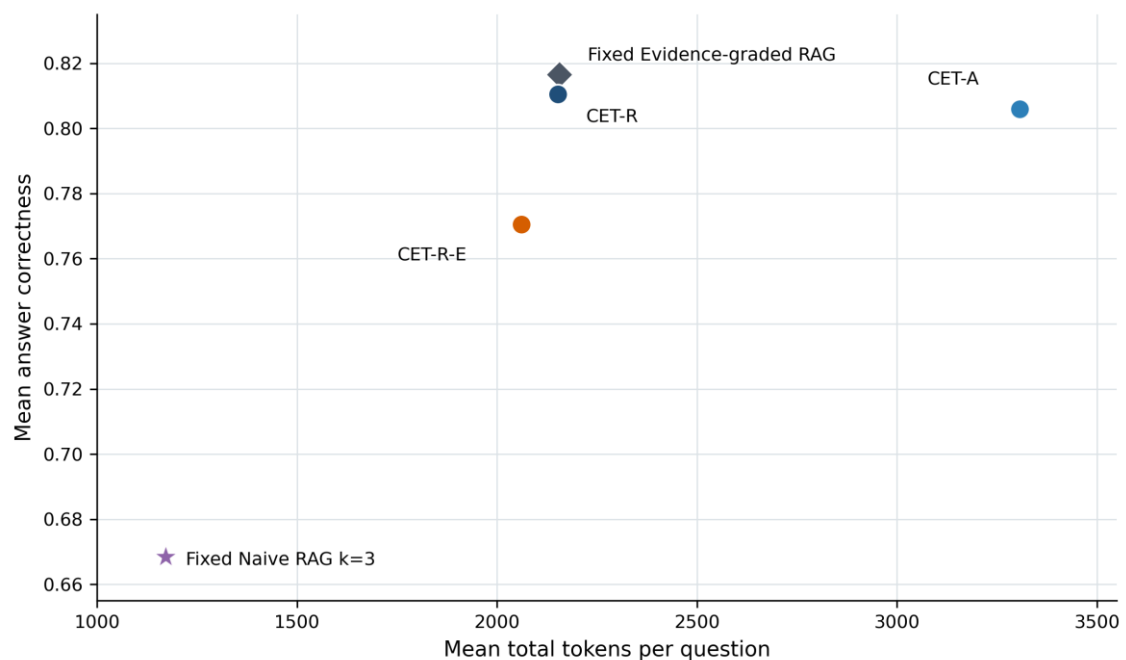


Fig. 3. Cost-effectiveness of the routing policies against the two fixed reference pipelines. The horizontal axis is mean total tokens per question under the accounting of §3.5. CET-A sits far to the right of the fixed pipeline it fails to beat, because its triage probe is a full generation charged on every question; CET-R attains the same correctness at the fixed pipeline's token cost, and CET-R-E trades correctness for a further reduction in both tokens and latency.

4.4 Selective answering and calibration

The signal that fails to improve action selection nonetheless carries usable information about whether an answer should be released. Ranking the 200 out-of-fold answers by the router's own confidence and sweeping the release threshold gives the risk-coverage curves in Fig. 4. CET-A attains an area under that curve of 0.115 (0.083 to 0.151) against 0.143 (0.104 to 0.185) for a

router that sees only the question and 0.140 (0.103 to 0.184) for one that sees only the probe. The ordering is consistent across coverage levels — at 50% coverage selective risk is 0.101 for CET-A against 0.138 and 0.136 for the two ablations — but the intervals overlap substantially at $n = 200$, and we therefore describe this as supportive rather than established.

Calibration behaves the same way and is measured more precisely. Expected calibration error was 0.040 for CET-A, 0.039 for the question-only router and 0.054 for the probe-only router; the retrieval-only router CET-R improved on all of them at 0.023. A router whose confidence is well calibrated is directly useful in a clinical workflow even when its action selection adds nothing, because it supports a defensible abstention or escalation rule [28,29,30]. This is the capability we would carry forward, and it is worth noting that the best-calibrated variant is also the cheapest and the one with the fewest assumptions about what the deployed system can observe.

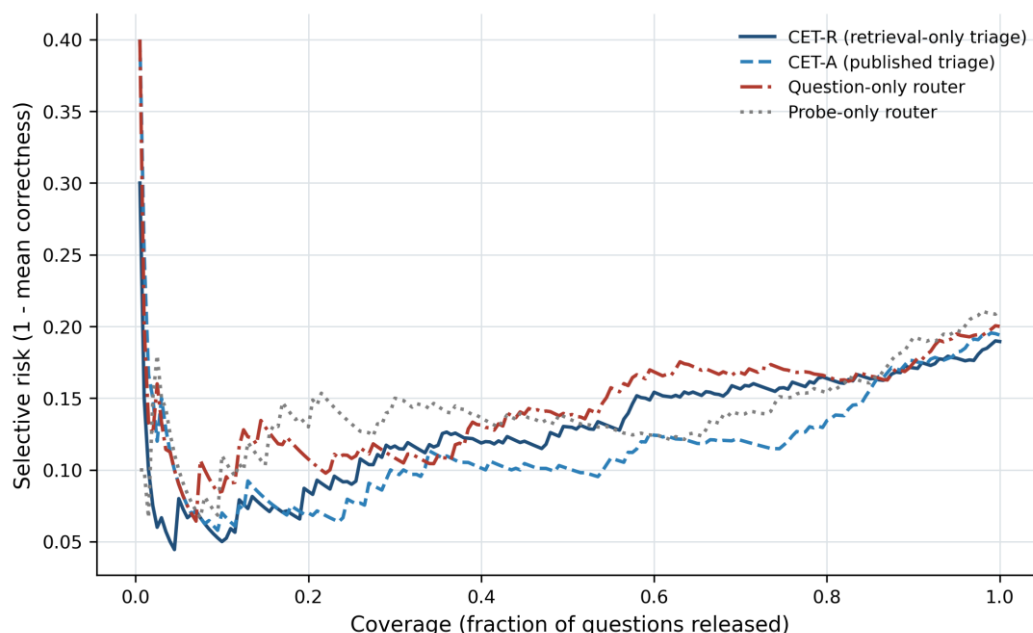


Fig. 4. Out-of-fold risk-coverage curves. Answers are ranked by the router's confidence in its selected action and released from most to least confident; the curve traces mean selective risk against the fraction released. Lower is better. Both evidence-aware routers dominate the question-only and probe-only ablations across most of the coverage range, though the curves converge at full coverage where every policy answers everything. The left-hand instability reflects the small number of answers released at low coverage.

4.5 Where the router acts, and what happens when it does

CET-A selected the fixed comparator on 89% of questions and departed from it on 22. Those departures were net harmful: 5 improved the answer, 10 degraded it and 7 left it unchanged, for a mean correctness change of -0.095 alongside a mean latency saving of 10.41 s. The router is thus trading accuracy for speed on exactly the questions where it exercises discretion, without having been asked to. Reporting this decomposition matters more than reporting the aggregate: a policy

that is null on average because it is inert is a different object from one that is null because its gains and losses cancel, and only the second carries a safety implication.

The pattern by question type is consistent. Against the fixed comparator, CET-A was level on the 150 single-hop questions (0.829 versus 0.833) but lost ground on the small, hard subgroups where routing was supposed to help: multi-hop (0.615 versus 0.625, $n = 20$) and comparative (0.718 versus 0.745, $n = 11$). These subgroups are too small to support inference on their own; we report them because the direction is uniform and because C6 forbids presenting the favourable subgroups without the unfavourable ones. The same holds by question provenance, where the deficit concentrates in the 50 constructed questions (0.736 versus 0.768) rather than the 150 PubMedQA-derived ones.

Table 6. Routing behaviour and selective risk. Departures counts the questions on which the router chose something other than its own modal action; the following column classifies those departures by whether they improved, degraded or left unchanged the correctness obtained by the fixed comparator. Selective risk is one minus mean correctness among the most confident 50% and 75% of released answers. Variants that route more do not answer better; CET-R-E departs on roughly half of all questions and carries the highest selective risk.

Router variant	Actions used	Modal share	Departures	Better / worse / tied	Risk at 50%	Risk at 75%
CET-A	4	89.0%	22	5 / 10 / 7	0.101	0.117
CET-J	4	89.0%	22	5 / 10 / 7	0.116	0.120
CET-D	4	95.5%	9	1 / 3 / 5	0.100	0.138
CET-R	4	97.5%	5	0 / 3 / 2	0.130	0.157
CET-R-E	9	51.0%	98	15 / 45 / 38	0.169	0.191
Question-only router	5	89.5%	21	3 / 10 / 8	0.138	0.170
Probe-only router	6	77.0%	46	6 / 24 / 16	0.136	0.151

4.6 Cluster grouping and fold stability

Questions are not fully independent: the 200 questions resolve into 187 connected components of the question-PMID graph, with the largest component containing 2 questions. Re-forming the outer folds so that a component never straddles training and test (control C9) leaves the conclusion unchanged and slightly strengthens it: CET-A's advantage becomes 0.000 (-0.013 to 0.015) and the policy collapses onto the fixed action for every one of the 200 questions, while CET-R gives -0.006 (-0.018 to 0.003). Grouping is cheap insurance here rather than a correction, because the clusters are small; in a benchmark built from fewer source articles it would not be.

Fold-level variability is the reason C6 is not optional. Across the five folds CET-A ranged from 0.748 to 0.847 and the fixed comparator from 0.762 to 0.847; the sign of the difference changes between folds. With 40 test questions per fold, a single fold is simply not an estimate, and the practice of reporting a favourable split has been criticised in the language-processing literature for exactly this reason [25,26].

4.7 Retrieval completeness and a mechanistic caveat

One further reporting defect is invisible in aggregate tables. Archived context-recall exports covered unequal numbers of questions per architecture, ranging from 73 to 200, with a single export accounting for 127 of the 136 missing scores. Comparing means computed on different question subsets is not valid, and the induced bias here is differential rather than uniform, spanning -0.047 to +0.015 across architectures, so available-case reporting is not salvageable. We therefore report a complete-case tier in which nine architectures are compared on all 200 questions with no missing scores (Table 7); the two naive-retrieval configurations responsible for the missingness are excluded and reported separately in the released artefacts.

Table 7. Complete-case context recall for the nine architectures with no missing scores (n = 200 each). Ranks run from best to worst within this set. The dissociation in the two rank columns is the point: the architecture that answers best is mid-table on how much of the relevant evidence it retrieves.

Architecture	Recall rank	Context recall (95% CI)	Correctness rank	Answer correctness
Naive RAG with critic	1	0.941 (0.913 to 0.965)	9	0.660
Query decomposition	2	0.934 (0.903 to 0.960)	2	0.748
Multi-query expansion	3	0.931 (0.900 to 0.958)	3	0.738
Naive RAG k=5	4	0.928 (0.897 to 0.956)	5	0.715
Hybrid RRF	5	0.918 (0.885 to 0.949)	6	0.696
Hybrid cross-encoder	6	0.911 (0.877 to 0.941)	4	0.735
Evidence-graded RAG	7	0.905 (0.867 to 0.938)	1	0.817
Query expansion	8	0.875 (0.835 to 0.912)	7	0.679
MeSH-guided RAG	9	0.846 (0.799 to 0.889)	8	0.670

Evidence-graded RAG ranks 7th of nine on context recall at 0.905 (95% CI 0.867 to 0.938) while ranking first on answer correctness. Its advantage is therefore not attributable to retrieving more of the relevant evidence; it must arise in how the retrieved evidence is graded and used at generation time. This is a mechanistic claim with a direct bearing on the routing result: if the winning action wins through evidence utilisation rather than evidence coverage, then triage

features computed from the retrieval step alone are unlikely to identify the questions on which a different action would do better, which is precisely the failure mode observed in §4.5.

5. Discussion

5.1 Principal findings

Three findings stand. First, cost-aware evidence-triage routing did not improve answer correctness over the strongest fixed biomedical RAG pipeline under a fully specified protocol (difference -0.010, 95% CI -0.030 to 0.008), and its discretionary decisions were net harmful. Second, the size of that conclusion is dwarfed by the size of the protocol effect: three individually documented shortcuts, applied together, turn a difference of -0.010 into 0.228 on identical data. Third, the useful capability is elsewhere. Restricting triage to retrieval-observable features produced a router that matches the fixed pipeline's correctness at its token cost and improves calibration, and the triage confidence supports selective release better than question-only or probe-only alternatives.

5.2 Why routing fails on this benchmark

The mechanism is visible in the data rather than inferred. One action dominates the marginal table by 0.069 over its nearest rival (Table 1), so a learner minimising expected loss over a 11-action space with 120 training questions per fold has little incentive to deviate, and the fitted policy collapses towards a constant. When it does deviate, the deviations concentrate on multi-hop and comparative questions where the outcome matrix is noisiest, and they lose more than they gain. The complete-case retrieval analysis supplies the reason the available triage signal cannot fix this: the winning action wins through evidence utilisation, not evidence coverage, so features describing the retrieved set do not separate the questions on which an alternative action would prevail. This is consistent with the broader finding that medical queries are among the hardest to route and with expert evaluations reporting that automatic gains in biomedical RAG often do not reproduce under systematic review [6].

It also clarifies what the null does and does not license. We did not prespecify a non-inferiority margin, so we cannot claim the router is equivalent to the fixed pipeline; we can only say that this study, at this sample size, provides no evidence of an advantage while providing clear evidence of a cost difference in both directions depending on how the probe is built. A study designed to establish equivalence would need a prespecified margin and a substantially larger question set.

5.3 Implications for reporting adaptive-RAG results

The ladder suggests a short, checkable reporting standard, and its ordering is informative. The two largest effects — comparator choice and split isolation — are conceptually elementary and cost nothing to implement, yet they are the ones most often left implicit: adaptive systems are routinely

compared against a fixed pipeline named in advance, without demonstrating that this pipeline is the strongest available alternative on the same data. We would ask any adaptive-RAG report to state which fixed comparator was used and how it was chosen, whether that selection used training data only, how thresholds or tolerances were tuned, how many folds and variants were run and whether all are reported, and what the triage features presuppose about deployment. The last item is what this study adds: a router whose features require gold annotation is not deployable, and one whose features require a full generation must be charged for it. Neither property is visible in an accuracy table. The framework itself is not specific to our router — it needs only a crossed matrix of questions and pipelines with per-question outcomes and costs, which any pipeline comparison already produces.

5.4 Limitations

Four limitations bound these conclusions. The study is offline: it learns and evaluates a policy over a fixed, pre-computed outcome matrix and runs no new inference, so the candidate actions cannot adapt to the router and interaction effects between routing and generation cannot be observed. The sample is small — 200 questions over 1,000 abstracts, with only 40 test questions per fold — which is why several intervals in Table 5 overlap and why we describe the selective-prediction advantage as supportive rather than demonstrated; it is also, as §5.2 argues, part of why the policy degenerates. Outcomes come from a single automated judge in a single pass, so judge variance is unmeasured and no clinician adjudication was performed; the judge-independence ablation addresses the circularity between features and labels but not the validity of the labels themselves [17,21]. Finally, the corpus is a single-source abstract collection, so we cannot speak to behaviour under conflicting evidence, full-text retrieval or cross-benchmark transfer; resources built for those settings [40,41,42] are the natural next targets.

5.5 Future work

Three extensions follow directly. A generation-free triage probe of the kind CET-R approximates should be built out with entailment and inter-passage conflict features computed by a small encoder, which would keep the token accounting at zero while adding the semantic signal the lexical features lack. The release gate should be replaced by conformal risk control [31], which supplies a distribution-free finite-sample guarantee and is therefore the right instrument at this sample size. And abstention should become a first-class action with an explicit cost rather than a post-hoc threshold, so that the decision to withhold an answer is learned and evaluated on the same footing as the decision to answer.

6. Conclusions

On a fully crossed benchmark of 200 biomedical questions and 11 retrieval-augmented pipelines, cost-aware evidence-triage routing did not improve answer correctness over the strongest fixed pipeline, and the questions on which it exercised discretion were more often harmed than helped. The same data, evaluated under three individually documented shortcuts, would have supported the opposite conclusion by a margin more than twenty times the size of the controlled estimate. What evidence triage does deliver is cost and confidence: a router restricted to features observable from retrieval alone matched the fixed pipeline's accuracy at its token cost and produced the best-calibrated confidence of any variant tested. We recommend that adaptive-RAG evaluations in the biomedical domain report against a comparator selected on training data, declare the provenance and cost of their triage features, and publish the sensitivity of their headline estimate to the controls they applied.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The question set, the complete question-by-action outcome matrix, the analysis scripts that generate every table and figure in this paper, and the run manifests recording the seed and configuration are available in the study repository. The leakage ladder is implemented as a standalone script that consumes any outcome matrix in the same format. Underlying abstracts are public records retrievable from PubMed [32]; no patient-level data were used.

Ethics statement

This study used publicly available literature abstracts and derived question sets, and involved no human participants, human tissue or identifiable personal data; ethical approval was therefore not required. The systems evaluated here are research prototypes. No result in this paper should be read as evidence of clinical readiness, and none of the pipelines examined is suitable for autonomous use in patient care.

Declaration of generative AI in the writing process

Generative AI tools were used to assist with drafting and language editing of this manuscript and with the implementation of analysis scripts. All experimental design, statistical interpretation and conclusions are the authors' own; the authors reviewed and verified every reported quantity against the released analysis artefacts and take full responsibility for the content of the publication.

References

- [1] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*. 2020;33:9459-74.
- [2] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*; 2019. p. 2567-77. doi:10.18653/v1/D19-1259.
- [3] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2.
- [4] Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. In: *Findings of the Association for Computational Linguistics: ACL 2024*; 2024. p. 6233-51. doi:10.18653/v1/2024.findings-acl.372.
- [5] Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inform Assoc*. 2025;32(4):605-15. doi:10.1093/jamia/ocaf008.
- [6] Kim H, Sohn J, Gilson A, Cochran-Caggiano N, Applebaum S, Jin H, et al. Rethinking retrieval-augmented generation for medicine: a large-scale, systematic expert evaluation and practical insights [preprint]. *arXiv*. 2025 Nov 10. doi:10.48550/arXiv.2511.06738.
- [7] Li M, Kilicoglu H, Xu H, Zhang R. BiomedRAG: a retrieval augmented large language model for biomedicine. *J Biomed Inform*. 2025;162:104769. doi:10.1016/j.jbi.2024.104769.
- [8] Sohn J, Park Y, Yoon C, Park S, Hwang H, Sung M, et al. Rationale-guided retrieval augmented generation for medical question answering. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2025. p. 12739-53. doi:10.18653/v1/2025.naacl-long.635.
- [9] Liang S, Zhang L, Zhu H, Wang W, He Y, Zhou D. RGAR: recurrence generation-augmented retrieval for factual-aware medical question answering. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*; 2025. p. 4006-33. doi:10.18653/v1/2025.findings-emnlp.214.
- [10] Jeong S, Baek J, Cho S, Hwang SJ, Park J. Adaptive-RAG: learning to adapt retrieval-augmented large language models through question complexity. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2024. p. 7036-50. doi:10.18653/v1/2024.naacl-long.389.
- [11] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: learning to retrieve, generate, and critique through self-reflection. In: *International Conference on Learning Representations*; 2024.
- [12] Yan SQ, Gu JC, Zhu Y, Ling ZH. Corrective retrieval augmented generation [preprint]. *arXiv*. 2024 Jan 29. doi:10.48550/arXiv.2401.15884.
- [13] Tang X, Gao Q, Li J, Du N, Li Q, Xie S. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In: *Proceedings of the 31st International Conference on Computational Linguistics*; 2025. p. 3248-54.
- [14] Marina M, Ivanov N, Pletenev S, Salnikov M, Galimzianova D, Krayko N, et al. LLM-independent adaptive RAG: let the question speak for itself. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; 2025. p. 8697-709. doi:10.18653/v1/2025.emnlp-main.439.

- [15] Es S, James J, Espinosa-Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; 2024. p. 150-8. doi:10.18653/v1/2024.eacl-demo.16.
- [16] Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024. p. 338-54. doi:10.18653/v1/2024.naacl-long.20.
- [17] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Adv Neural Inf Process Syst*. 2023;36:46595-623.
- [18] Ru D, Qiu L, Hu X, Zhang T, Shi P, Chang S, et al. RAGChecker: a fine-grained framework for diagnosing retrieval-augmented generation. *Adv Neural Inf Process Syst*. 2024;37. doi:10.52202/079017-0692.
- [19] Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, et al. RAGTruth: a hallucination corpus for developing trustworthy retrieval-augmented language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024. p. 10862-78. doi:10.18653/v1/2024.acl-long.585.
- [20] Confident AI. DeepEval: the LLM evaluation framework [computer software on the Internet]. [cited 2026 Jul 31]. Available from: <https://github.com/confident-ai/deepeval>.
- [21] Cruz Blandon MA, Talur J, Charron B, Liu D, Mansour S, Federico M. MEMERAG: a multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025. p. 22577-95. doi:10.18653/v1/2025.acl-long.1101.
- [22] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data*. 2012;6(4):15. doi:10.1145/2382577.2382579.
- [23] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.
- [24] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-107.
- [25] Gorman K, Bedrick S. We need to talk about standard splits. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 2786-91. doi:10.18653/v1/P19-1267.
- [26] Sogaard A, Ebert S, Bastings J, Filippova K. We need to talk about random splits. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics; 2021. p. 1823-32. doi:10.18653/v1/2021.eacl-main.156.
- [27] Elangovan A, He J, Verspoor K. Memorization vs. generalization: quantifying data leakage in NLP performance evaluation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics; 2021. p. 1325-35. doi:10.18653/v1/2021.eacl-main.113.
- [28] El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. *J Mach Learn Res*. 2010;11:1605-41.
- [29] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. *Adv Neural Inf Process Syst*. 2017;30.
- [30] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017. p. 1321-30.
- [31] Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. In: International Conference on Learning Representations; 2024.
- [32] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2022;50(D1):D20-6. doi:10.1093/nar/gkab1112.
- [33] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019. p. 3982-92. doi:10.18653/v1/D19-1410.
- [34] Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv Neural Inf Process Syst*. 2020;33:5776-88.
- [35] Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2009. p. 758-9. doi:10.1145/1571941.1572114.

- [36] Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the middle: how language models use long contexts. *Trans Assoc Comput Linguist.* 2024;12:157-73. doi:10.1162/tacl_a_00638.
- [37] Efron B, Tibshirani RJ. *An introduction to the bootstrap.* New York: Chapman & Hall; 1993.
- [38] Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull.* 1945;1(6):80-3. doi:10.2307/3001968.
- [39] Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65-70.
- [40] Zhang B, Bornet A, Yang R, Liu N, Teodoro D. HealthContradict: evaluating biomedical knowledge conflicts in language models. *npj Digit Med.* 2026;9:152. doi:10.1038/s41746-025-02336-0.
- [41] Dobrev J, Karasmanakis I, Ivanisevic F, Horvat T, Kitanovski D, Gams M, et al. RAGCare-QA: a benchmark dataset for evaluating retrieval-augmented generation pipelines in theoretical medical knowledge. *Data Brief.* 2025;63:112146. doi:10.1016/j.dib.2025.112146.
- [42] Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics.* 2015;16:138. doi:10.1186/s12859-015-0564-6.