

Supplementary Appendix 1 for Every-successful-replay route admission changes first-token latency rankings in clinical question answering

Rui Li*, Jason Zhao, Shuang Cao, Alexandre Duprey, and Ruihua Liu

Department of AI, Hill Research, Princeton, NJ 08540, USA

***Correspondence:** rui.li@hillresearch.ai

Journal: BMC Medical Informatics and Decision Making

S1 Implementation and Workload Details

All experiments ran on Microsoft Azure. The graph store uses $5 \times$ Standard E32ads v5 instances (32 vCPUs, 256 GiB RAM, AMD EPYC 7763) running Neo4j 5.15.0 Enterprise with 128 GiB page cache. The query service runs on $2 \times$ Standard E16ads v5 instances (16 vCPUs, 128 GiB RAM) using Python 3.11 and PyTorch 2.1.2. The vector index runs on $1 \times$ Standard NC24ads A100 v4 instance (24 vCPUs, 220 GiB RAM, NVIDIA A100 80GB, CUDA 12.2) hosting Faiss 1.7.4 HNSW with $M = 32$ and efSearch=64. Redis 7.2.3 provides a 64 GiB cache with allkeys-lru, and the LLM is Azure OpenAI GPT-4-Turbo (gpt-4-1106-preview, 128K context, temperature 0.2, max_tokens=2048).

The compiler uses a fine-tuned DistilBERT intent classifier, lexical and fuzzy linking over UMLS-style aliases, and dense biomedical retrieval initialized from encoders in the BioBERT/ClinicalBERT family [1–3]. Template selection chooses among 47 parameterized Cypher queries [4]. On the pre-benchmark 2,480-question development evaluation, the intent classifier reaches macro-F1 0.940. The entity-linking cascade reaches 96.8% accuracy on a manually annotated set of 500 questions. Template selection covers 98.7% of the 847-question benchmark.

The study-specific TextRAG comparator uses BM25 retrieval in Elasticsearch 8.11.3, ColBERT reranking, top-10 passage retrieval, and GPT-4-Turbo with passage-level citations. GraphRAG-OD performs query-time graph construction using the Microsoft GraphRAG implementation (v0.3.2) [5], followed by graph traversal and summarization. Heuristic-GF is MedRouteGuard with a fixed graph-first plan. OfflineGraph-Tuned uses the

same evidence graph, the same 847-query benchmark, the same Azure worker classes, the same Redis cache budget, the same admission contract, and the same latency protocol as MedRouteGuard; admitted-answer counts are reported separately. It keeps a fixed graph-first serving policy with hand-tuned Neo4j query hints and ColBERT-based entity linking. OfflineGraph-Tuned is therefore the matched fixed graph-serving comparator for the Q2 latency claim; TextRAG supports the specified Q1 output audit, while GraphRAG-OD is descriptive. Table S14 supplies the paired quality bridge for Q2 by auditing MedRouteGuard and OfflineGraph-Tuned on the same sampled questions and admission contract.

Table S1 maps the maintained-service experiment blocks with retained accounting to fixed worker pools, wall-clock duration, and aggregate compute.

Table S1 Retained compute accounting for the maintained-service experiment blocks. CPU node-hours include the five graph-store and two query-service workers; GPU hours refer to the NC24ads A100 v4 worker.

Experiment block	Primary workers / measured unit	Wall-clock and aggregate compute
Matched latency, throughput, and planner analyses	Full online stack: 5 × E32ads v5 Neo4j workers, 2 × E16ads v5 query-service workers, 1 × NC24ads A100 v4 vector-index worker, and Redis; measured on the 847-query benchmark.	18.4 h; 128.8 CPU node-hours; 18.4 A100 hours.
Cache, warm/cold, and update analyses	Same serving stack plus injected update replays on the maintained graph; measured on the no-update run and the 5/20/100 upd/s stress settings.	11.7 h; 81.9 CPU node-hours; 11.7 A100 hours.
Retrospective answer and citation audits	Blinded answer and citation packets exported from the same benchmark; system inference used the same serving stack, while reviewer time is outside system compute.	6.3 h; 44.1 CPU node-hours; 6.3 A100 hours.
Total retained compute for these blocks	Three experiment blocks above; excludes clinician review time and the separately executed medication-safety workload.	36.4 h; 254.8 CPU node-hours; 36.4 A100 hours.

The first row anchors online serving results to the exact worker classes used in the main paper. The cache and update analyses reused that serving stack rather than a separate simulator. Reviewer time is excluded from system compute because it measures human evaluation rather than inference. The total is the sum of the three non-overlapping measured blocks.

S2 Operational Support for Plan Selection, Cache Reuse, and Ablations

Table S2 Plan-selection quality under the admissible oracle on the 847-query benchmark. Higher oracle match and lower regret are better.

Strategy	Oracle Match	P95 (s)	Regret (s)
Cost-based (MedRouteGuard)	94.2%	3.24	0.18
Always graph-first	67.4%	5.49	1.82
Always hybrid	58.1%	4.73	1.24
Random	33.3%	7.21	3.47

The oracle is the fastest admissible plan family measured under the same graph snapshot and cache state, and regret is averaged over all 847 questions. The table shows that this workload requires routing inside the admitted set: the fixed graph-first policy reaches 67.4% oracle match and 5.49 s planner-diagnostic P95, compared with 94.2% and 3.24 s for the cost-based planner. When the planner misses, the penalty appears in tail delay. The cost-based planner therefore improves search within a fixed admission denominator, while Section S6 reports additional fit diagnostics.

Table S3 Cold versus warm cache latency. Lower latency and higher hit rate are better.

State	Hit Rate	P50 (s)	P95 (s)	P99 (s)
Cold	0.0%	1.87	6.78	9.42
Warm	67.3%	1.42	3.24	4.87

Table S3 shows that reuse is operationally material: priming raises hit rate to 67.3% and cuts deployed MedRouteGuard first-generated-token P95 from 6.78 s to 3.24 s. The gain is concentrated in the tail rather than only at the median, so cache state must be fixed when comparing tail latency in this deployment. The warm row anchors the Q2 deployed-system operating point and the online planner and ablation diagnostics. Q3’s gate-on retrospective-oracle P95 happens also to be 3.24 s, but it remains a separate fixed-candidate estimand. The table therefore turns cache state from a background implementation detail into a controlled part of the benchmark contract.

Table S4 Cache behavior under injected updates across 1,968 auditable reuse events in the archived replays. One auditable reuse event is one attempted cache or derived-summary reuse checked against the current watermark. P95 is the deployed-stack first-generated-token diagnostic. Stale-hit entries report whether any stale reuse was observed within each setting; aggregated across all rows, the archived replays contain 0/1,968 observed stale-hit events. Lower false-invalidation rates and lower latency are better.

Upd/s	Hit Rate	False Inv.	Stale Hit	P95 (s)
0	67.3%	0.0%	0 obs.	3.24
5	64.2%	1.7%	0 obs.	3.41
20	58.7%	4.3%	0 obs.	3.82
100	42.1%	12.4%	0 obs.	5.21

Table S4 shows the safety side of that reuse. Across the archived replays, stale-hit monitoring records 0/1,968 observed stale-hit events, while false invalidations rise from 0.0% to 12.4% as update pressure increases from 0 to 100 upd/s. Treating those replays as a zero-count sample yields a one-sided 95% Clopper–Pearson upper bound of 0.15% on the stale-hit rate within the archived replays. The stress test documents conservative invalidation under churn and quantifies its increasing latency cost.

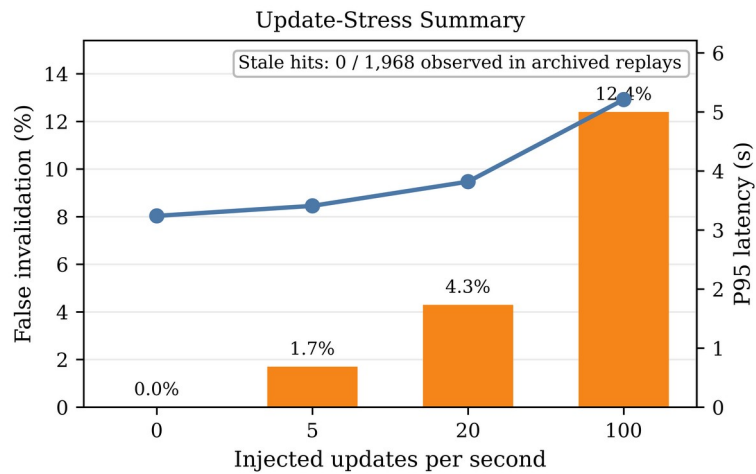


Fig. S1 Update-stress summary backed by Table S4: false invalidation rises with update rate, while stale-hit monitoring records 0/1,968 observed stale-hit events in the archived replays

Figure S1 turns that zero-count table into the intended trade-off view. Under heavier churn, the system gives up hit rate and accepts more false invalidations rather than serve stale state. The x-axis therefore tracks the cost of refusing stale reuse, not the benefit of a permissive cache. The intended asymmetry is simple: the system

gives up speed before it gives up freshness. That asymmetry makes the freshness clause an observed systems cost within the benchmark rather than an untimed metadata check.

Table S5 Runtime configuration diagnostics at the warm-cache, no-update operating point. Lower latency and higher evidence precision are better; rows compare complete configurations using the archived aggregate operating-point summaries.

Configuration	P50 (s)	P95 (s)	Evid. Prec.
Full MedRouteGuard	1.42	3.24	0.91
– Cost-based planner	1.68	5.49	0.91
– Subgraph cache	1.87	6.78	0.91
– Derived summaries	1.58	4.12	0.89
– Fan-out caps	1.52	8.73	0.91
– All caching	1.91	7.21	0.91

At this archived operating point, every shown removal had a higher P95 point estimate than the full configuration. The Full MedRouteGuard row is the warm-cache, no-update point reported in the main-text deployed-system latency table and Table S2; the –Cost-based planner row is the internal fixed-policy diagnostic rather than OfflineGraph-Tuned. The only lower evidence-precision point estimate was observed after removal of derived summaries. These configuration diagnostics show how route selection and dependency-aware reuse operate together; the main-text membership intervention remains the controlled analysis of admission.

S3 Clause-Robustness of the Boundary Audit

We tested clause-level robustness by recomputing the retrospective oracle on the same fixed 847-question executions while relaxing one clause label at a time. The graph snapshot, prompt bundle, Azure worker classes, candidate plan families, cache budget, gate computations, and latency estimator remained fixed. The full-gate baseline is the admitted benchmark reported in the main-text Results; the comparison variants ignore provenance only, temporal only, conflict only, freshness only, or all four labels. For each variant we report apparent P95 latency, excluded-candidate speed events, median-route winner changes, candidate-pool bucket-set changes, dominant affected question families, and 95% bootstrap intervals over question IDs.

Table S6 Clause-robustness audit for the same-stack boundary effect. Every row uses the fixed 847-question benchmark, graph snapshot, hardware envelope, prompt bundle, candidate plan families, and latency estimator; only the named candidate-level admissibility label is relaxed. A qualifying fast replay comes from a candidate classified analysis-inadmissible; the replay itself is not assigned the named failed-clause tag.

Ignored label	Apparent P95	Excluded-candidate speed event	Winner flip	95% CI	Bucket-set change	Dominant affected families
Provenance only	3.02 s	62/847 (7.3%)	34/847 (4.0%)	[2.7%, 5.5%]	45/847 (5.3%)	Candidate-level provenance tags concentrate in safety and aggregation questions.
Temporal only	3.07 s	57/847 (6.7%)	31/847 (3.7%)	[2.4%, 5.1%]	42/847 (5.0%)	Candidate-level temporal tags concentrate in temporal and aggregation questions.
Conflict only	3.14 s	29/847 (3.4%)	16/847 (1.9%)	[1.1%, 2.8%]	21/847 (2.5%)	Candidate-level conflict tags concentrate in aggregation and safety questions.

Ignored label	Apparent P95	Excluded-candidate speed event	Winner flip	95% CI	Bucket-set change	Dominant affected families
Freshness only	3.10 s	41/847 (4.8%)	24/847 (2.8%)	[1.9%, 3.9%]	33/847 (3.9%)	Candidate-level freshness tags concentrate in temporal and multi-hop questions.
All clauses	2.89 s	98/847 (11.6%)	61/847 (7.2%)	5.7% to 9.0%	83/847 (9.8%)	Joint relaxation most affects safety, temporal, and aggregation questions.

Table S6 shows a graded boundary effect. Provenance and temporal relaxations produce the largest single-clause winner changes, freshness affects update-sensitive questions, and conflict-only relaxation also changes benchmark membership. Relaxing all four labels produces the largest effect, consistent with interacting clause families.

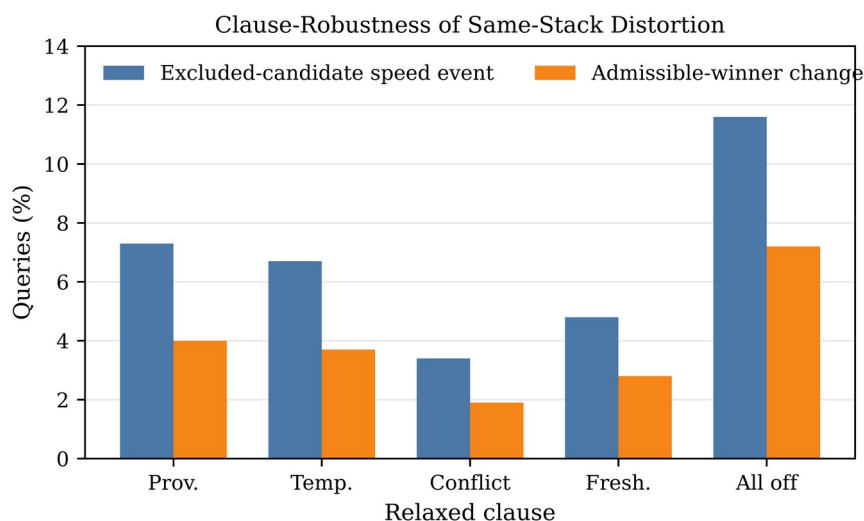


Fig. S2 Clause-robustness view of the same-stack boundary audit. Bars show excluded-candidate speed events and median-latency winner changes for each candidate-level label relaxation on the 847-question benchmark. The figure does not assert that a qualifying replay failed the named clause

Figure S2 makes the robustness pattern visible at a glance. Winner changes are largest when provenance or temporal checks are relaxed, but every single-clause relaxation remains above zero. The same benchmark-membership effect survives milder relaxations, and the clause families differ in magnitude rather than only in wording.

S4 Aggregate Support for the Boundary Audit

Because the main-text route-membership table is the direct audit of whether the timing boundary matters, this section reports the two auxiliary projections needed to interpret the aggregate: overlap among violated clauses and concentration by question family. Both use the same 847-question candidate pool, synchronized replay conditions, and question-level bootstrap described in Methods.

Table S7 Overlap structure among candidate-level failed-clause tags for excluded candidates associated with a speed event. Pair counts are question-level and non-exclusive because one candidate can carry more than two tags. Tags may originate from different replays and do not describe the clause state of the qualifying fast replay.

Violation pattern	Questions	Interpretation
Provenance + temporal	7	Associated excluded candidates carry both provenance and temporal tags across their measured-replay sets.

Violation pattern	Questions	Interpretation
Provenance + freshness	6	Associated excluded candidates carry both provenance and freshness tags across their measured-replay sets.
Temporal + freshness	5	Associated excluded candidates carry both temporal and freshness tags across their measured-replay sets.
Provenance + conflict	4	Associated excluded candidates carry both provenance and conflict tags across their measured-replay sets.
Temporal + conflict	3	Associated excluded candidates carry both temporal and conflict tags across their measured-replay sets.
Conflict + freshness	2	Associated excluded candidates carry both conflict and freshness tags across their measured-replay sets.
Any two or more clauses	18/98 (18.4%)	Multi-clause failures are common enough that the aggregate distortion is not driven by one isolated admissibility clause.

Table S7 shows that the speed-event slice contains structured, overlapping candidate-level failures. Provenance and temporal overlap is the most common pair, followed by combinations involving freshness. The overlap structure supports the interpretation of benchmark membership as a multi-clause evidence boundary.

Table S8 Question-family breakdown for the same-stack distortion diagnostic. Shares use the fixed main-text cohort denominators. Clause descriptors are candidate-level tags and are not attributed to the replay producing a qualifying fast latency.

Question family	Excluded-candidate speed event	Median-route winner flip	Dominant relaxed clause
Entity lookup	12/187 (6.4%)	6/187 (3.2%)	Candidate-level provenance.

Question family	Excluded-candidate speed event	Median-route winner flip	Dominant relaxed clause
Multi-hop	25/241 (10.4%)	14/241 (5.8%)	Candidate-level provenance and freshness.
Aggregation	22/168 (13.1%)	15/168 (8.9%)	Candidate-level conflict.
Safety	21/139 (15.1%)	14/139 (10.1%)	Candidate-level provenance and conflict.
Temporal	18/112 (16.1%)	12/112 (10.7%)	Candidate-level temporal scope and freshness.
95% bootstrap CI over question IDs	9.6% to 13.9%	5.7% to 9.0%	Distortion remains visible after uncertainty is propagated over benchmark questions.

Table S8 shows where the same-stack distortion is concentrated. Entity-lookup questions are least affected, consistent with their shorter evidence chains and constrained route space. Safety, temporal, and aggregation questions are more fragile because they place simultaneous pressure on provenance visibility, contradiction handling, and scope control. The bootstrap rows propagate uncertainty over benchmark questions rather than presenting an unqualified pooled point estimate. The table therefore identifies clinically meaningful workload strata in which denominator distortion is most frequent.

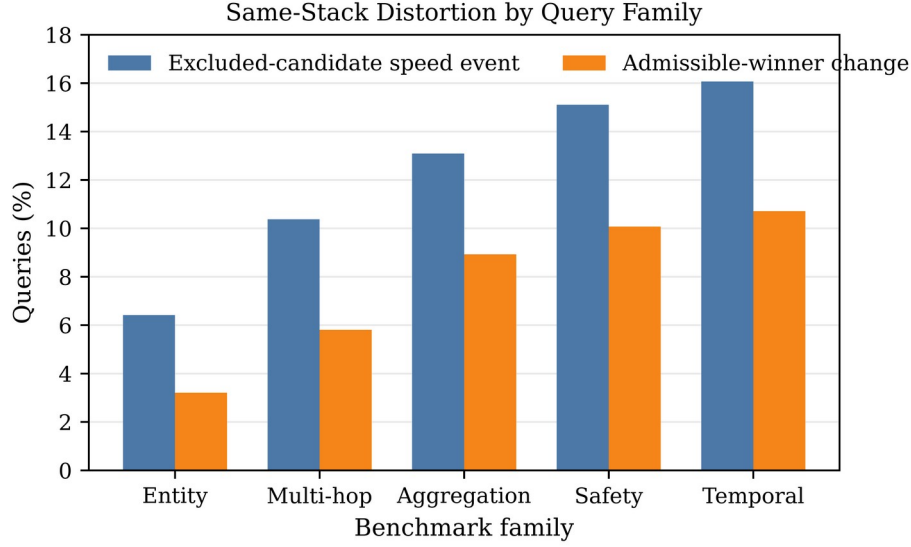


Fig. S3 Distribution view of the same-stack membership diagnostic. Bars show excluded-candidate speed events and median-latency winner changes by question family. A qualifying replay inherits only the candidate-pool classification, not the candidate’s failed-clause tags

Figure S3 makes the family concentration behind the main-text route-membership result visible at a glance. Winner flips are highest in safety (14/139 (10.1%)), temporal (12/112 (10.7%)), and aggregation (15/168 (8.9%)) questions, where provenance, contradiction handling, and time scope most directly affect clinical usefulness. The pattern concentrates membership sensitivity in evidence-intensive question families. Entity lookup has the lowest rate, showing that membership sensitivity increases with evidence-composition demands.

S5 Portable Public-Only Slice

To test whether the timing boundary remains behaviorally active outside the private-backed slice, we define a public-only workload containing 214 benchmark questions whose evidence packets can be satisfied from the 127K public-document subset and the corresponding public graph edges without EMR-derived evidence. The admissibility clauses, prompt bundle, candidate plan families, and hardware envelope are unchanged; only the private evidence slice is removed. This analysis isolates portability of the boundary effect, while the matched Q2 latency comparison remains anchored to the full workload.

Table S9 Portable public-only slice for the gate-on versus gate-off distortion diagnostic.

Metric	Public-only slice summary
Questions in slice	214.

Metric	Public-only slice summary
Admissible P95 latency	3.61 s.
Gate-off P95 latency	3.18 s.
Excluded-candidate speed events	19/214 (8.9%).
Median-latency route-winner changes	11/214 (5.1%).
Dominant relaxed labels	Candidate-level temporal-scope and provenance tags are most frequent among public-only winner changes.

The public-only slice remains boundary-sensitive: the gate-off analysis lowers apparent tail latency from 3.61 s to 3.18 s and changes 11/214 (5.1%) median-route winners. Its direction agrees with the full benchmark and separates the membership effect from the private EMR-backed evidence component. This slice evaluates portability of the route-membership diagnostic while the main Q2 comparison remains anchored to the matched full-workload baseline.

S6 Additional Planner Analyses

Table S10 supplies the planner-fit diagnostics. Prediction error concentrates in a small set of high-fan-out or statistics-stale cases rather than being uniformly diffuse, explaining how the planner attains high oracle match. These diagnostics characterize the evaluated deployment and statistics-refresh policy.

For Table S2, the oracle is the fastest admissible plan family among graph-first, hybrid, and text-fallback under the same graph snapshot and cache state used for the evaluated query. Oracle match counts ties. Reported regret is the latency gap to that oracle averaged over all 847 queries, so the 94.2% tie-or-match cases contribute zero regret. Planner misses are dominated by underestimated fan-out for newly popular entities, overestimated vector-retrieval utility for rare diseases, and temporal-selectivity errors.

Table S10 makes the error scale explicit for each plan family. This matters because the main paper’s regret claim depends on ranking adequacy, not on exact runtime simulation. The overall column is therefore the key summary for the main claim, while the family-specific columns explain where the remaining misses come from. Read against the multi-second main-text latency gaps, the table shows why sub-second estimation noise can still be adequate for plan ranking.

Table S10 Latency-estimation accuracy across the 847-query benchmark.

Metric	Graph-first	Hybrid	Text-fallback	Overall
MAE (ms)	187	241	176	203
RMSE (ms)	312	398	295	347
MAPE (%)	14.2	18.7	13.5	15.8
P95 error (ms)	524	687	512	583
Max error (ms)	1,847	2,312	1,932	2,312

Hybrid plans are hardest because they inherit variance from both retrieval and graph expansion. Even so, the overall error scale remains small relative to the multi-second tails in the main-text latency results. The table therefore supports ranking adequacy inside the admissible set. The family breakdown also explains why a few high-fan-out misses can coexist with a strong oracle-match rate in Table S2. In other words, the estimator only needs to be ranking-adequate at the scale of route choice, not numerically exact on every query.

S7 Additional Cache and Service Analyses

The extended Results report matched latency and complete-answer throughput; Section S2 reports the cold/warm and update-stress tables. This section adds the component-level warm-cache breakdown. The planner’s control overhead remains small relative to retrieval and summarization even in a reuse-active regime, although these measurements remain specific to one deployment stack. For the update-stress setting referenced by Table S4, one auditable reuse event denotes one attempted subgraph-cache or derived-summary reuse checked against the current watermark. Aggregated across the archived replays, the logs record 0/1,968 observed stale-hit events, so the 0.0% cells should be read as rounded zero counts. Treating those replays as a zero-count sample yields a one-sided 95% Clopper–Pearson upper bound of 0.15% on the stale-hit rate within the archived replays.

Table S11 Archived warm-cache component P50 values for MedRouteGuard in milliseconds. Displayed percentages are rounded and sum to 100.2%.

Component	P50 (ms)	% of total
Intent classification	18	1.3%

Component	P50 (ms)	% of total
Entity linking	31	2.2%
Plan selection	8	0.6%
Cache lookup	4	0.3%
Graph traversal	302	21.3%
Subgraph assembly	28	2.0%
LLM summarization	862	60.7%
Network/serialization	167	11.8%
Total	1,420	100%

Table S11 makes the warm-cache operating point explicit. LLM summarization remains the dominant reported component, while graph traversal is the largest non-LLM systems component. Planner and cache bookkeeping are small in the archived component summary; admission checks were not retained as a separate component row. The table also explains why the planner can matter materially even though its own runtime is small: a routing decision influences a much larger downstream latency budget. Read together with Tables S3 and S5, the breakdown locates most reported latency in cache state, traversal, and summarization rather than plan-selection bookkeeping.

S8 Quality Breakdown and Citation-Error Mechanisms

Among the 500 MedRouteGuard output packets in the answer/provenance audit, reviewer-coded recall-error categories occurred for missing derived edges on rare drug–disease pairs (7% of packets), temporal filters excluding recent evidence (5%), and entity-linking failures (4%); categories were non-exclusive. Packet-level precision-error categories were overly broad traversals (5%) and conflict-group edges included when not requested (4%). These rates are distinct from the 23 citation-error instances in the separate mechanism taxonomy below. They localize residual failures to specific retrieval, temporal-filtering, linking, and conflict-group mechanisms. Within this sample, reviewers identified no factually incorrect medical advice with serious safety implications.

Table S12 Unit crosswalk for the separate 200-question blinded citation-mechanism audit.

Analysis view	Unit	Interpretation
Mechanism-audit sample	200 matched questions; 400 blinded query–answer items	Three citation experts inspected both system outputs to preserve blinded review context.
Reviewer exposure	1,200 output–expert assessments	Each expert assessed one blinded output for missing citations, citation mismatches, and unflagged conflicts.
Prespecified reported endpoint	23 MedRouteGuard citation-error instances	The endpoint is a descriptive mechanism taxonomy, not a second comparative provenance effect; one output can contribute multiple instances.

Table S12 separates matched-question, query–answer-item, output–expert, and error-instance denominators. The comparative provenance effect comes from the 500-pair audit in the main paper; this separate endpoint localizes MedRouteGuard error mechanisms. One audited output can contribute multiple instances when it simultaneously misses one citation, mismatches another, and fails to expose a conflicting edge. Within the 23-instance taxonomy, the most common causes are entity-linking failures (8 cases, 35%), temporal-filter over-pruning (6, 26%), conflict-group omission (5, 22%), and generation error despite provided evidence (4, 17%). This breakdown distinguishes retrieval, admission, and generation failures without treating instance percentages as question-level rates.

The 500-pair answer/provenance audit and separate 200-question citation-mechanism audit were sampled from the 847-question benchmark. Each blinded item showed one question, one system answer, and that answer’s returned evidence; the mechanism audit therefore comprised 400 query–answer items and 1,200 output–expert assessments. Five board-certified specialists scored correctness, completeness, safety, and actionability; three citation experts classified support failures. Reviewers penalized unsupported claims, omitted contradictions, temporal mismatches, and clinically risky phrasing even when the prose was fluent. Comparative provenance counts, score distributions, reviewer-aware inference, agreement, and the missing-rating rule come from the 500-pair audit reported in Methods and Box 1.

For the matched Q2 bridge, we draw 240 benchmark queries from the same 847-query workload, stratified by the five main-text query families. Auditors see the original question, one blinded system answer, and the surfaced evidence returned by that system; MedRouteGuard and OfflineGraph-Tuned share the graph snapshot, question sample, admissibility contract, and blinded rubric, and only the serving policy changes. The table reports mean answer-level scores, acceptable-rate percentages, surfaced-provenance precision/recall/F1, and paired deltas with 95% intervals.

Table S13 Family stratification for the matched Q2 paired audit. The sample preserves the fixed five-family workload mix while holding the benchmark source, graph snapshot, question denominator, and admissibility contract fixed.

Query family	Sampled queries	Why this stratum is preserved
Entity lookup	53	Keeps direct evidence-lookup behavior visible instead of letting harder families dominate the packet.
Multi-hop	68	Preserves the route family with the largest graph-expansion heterogeneity.
Aggregation	48	Keeps conflict-handling and evidence-composition questions visible in the matched audit.
Safety	39	Preserves the family where unsupported or contradiction-hiding answers are most consequential.
Temporal	32	Preserves the family most exposed to time-scope and freshness failures.
Total	240	Provides the paired audit sample reported in Table S14.

Table S13 shows that the Q2 audit preserves the benchmark mix across entity lookup, multi-hop, aggregation, safety, and temporal behavior. That balance aligns the paired quality bridge with the same question population and admissibility contract used for the latency claim.

Table S14 Aggregate paired blinded audit for the matched Q2 comparator on 240 sampled benchmark queries. Paired deltas are MedRouteGuard minus OfflineGraph-Tuned; positive is better.

Metric	MedRouteGuard	OfflineGraph-Tuned	Paired delta (95% CI)
Correctness	4.69	4.63	+0.06 [-0.02, +0.14]
Completeness	4.16	4.03	+0.13 [+0.03, +0.23]
Safety	4.80	4.75	+0.05 [-0.01, +0.11]
Actionability	4.01	3.92	+0.09 [+0.00, +0.19]
Acceptable (%)	93.8	91.7	+2.1 pp [-1.0, +5.4]
Evid. precision	0.90	0.86	+0.04 [+0.01, +0.07]
Evid. recall	0.83	0.80	+0.03 [+0.00, +0.06]
Evid. F1	0.86	0.83	+0.03 [+0.01, +0.06]

Table S14 provides the aggregate paired quality bridge for the Q2 comparator. Acceptability was 93.8% for MedRouteGuard and 91.7% for OfflineGraph-Tuned, with a paired difference of +2.1 percentage points [− 1.0, +5.4]; completeness and surfaced-provenance estimates favored MedRouteGuard on the matched subsample. The audit links the latency comparison to the same question population, evidence contract, and blinded quality rubric.

Table S15 Analysis specification for the matched Q2 paired audit.

Analysis component	Specification	Role in interpreting Table S14
Sampling	Unique benchmark question IDs and fixed five-family labels	Establishes that the audit is drawn from the same 847-question population as Q2.
Blinded item	One question, one system answer, and that answer’s surfaced evidence	Fixes the judged unit and prevents side-by-side system cues.
Assignment	Two initial ratings per output with independent adjudication of acceptability or safety disagreements	Separates reviewer replication from the paired question-level analysis.

Analysis component	Specification	Role in interpreting Table S14
Observed rating completeness	960/960 primary ratings complete	Confirms that no sampled output was removed for a missing primary rating.

Table S15 fixes the Q2 audit population, sampled unit, blinded presentation, reviewer assignment, and complete primary-rating denominator before inference is specified.

Table S16 Endpoint and inference specification for the matched Q2 paired audit.

Analysis component	Specification	Role in interpreting Table S14
Adjudication	37/480 outputs (7.7%)	Reports the output-level frequency of independent acceptability or safety disagreement resolution.
Acceptability	Separate overall task-adequacy judgment, with no consequential error or serious safety defect	Makes the binary endpoint operational and clinically interpretable.
Paired inference	Question-level paired deltas with 95% intervals	Preserves pairing between systems rather than treating outputs as independent.
Missing outcomes	No imputation; missing items remain in the reported denominator	Prevents selective removal of difficult or failed outputs.

Tables S15 and S16 fix the audit unit, blinding, assignment, endpoint, pairing, and missing-outcome rule. These design choices attach the quality analysis to the same question population and admissibility contract as Q2. The resulting intervals address compatibility with the observed latency difference.

Table S17 Blinded audit-item template used in human evaluation.

Template field	Visible content and decision role
Benchmark question	The original clinician-authored query, shown verbatim so the reviewer can judge scope, temporality, and requested action.
Blinded system answer	One system output with system identity removed; reviewers judge correctness, completeness, safety, and actionability from this text.

Template field	Visible content and decision role
Surfaced supporting evidence	The edge- or passage-level citations returned by that same system, shown together with the answer so unsupported claims and omitted contradictions can be penalized directly.
Rubric reminder	Reviewers are instructed to penalize unsupported claims, temporal mismatch, contradiction suppression, and clinically risky phrasing even when the prose is fluent.

Table S17 makes the human audit concrete at the level of one reviewed output. The benchmark question anchors scope, the blinded answer defines the judged output, and the supporting evidence permits direct assessment of claim support. The rubric reminder identifies clinically relevant errors that must be penalized even when the prose is fluent.

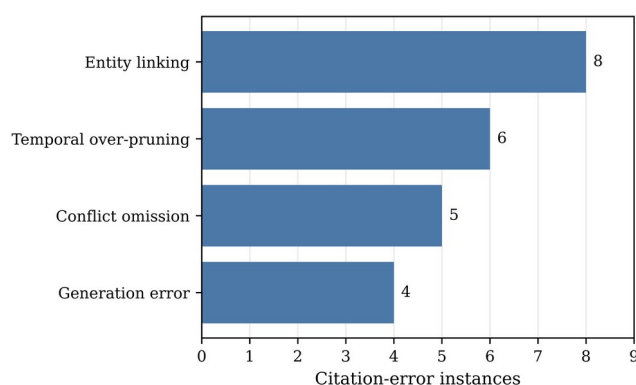


Fig. S4 Instance-level taxonomy of the 23 MedRouteGuard citation-error cases described in this section

Figure S4 presents the 23-case failure taxonomy as a distribution. Entity linking is the largest single source, followed by temporal over-pruning and conflict omission. Most residual audited failures therefore arise before final answer wording rather than from free-form generation alone. The instance-level composition complements the packet-level prevalence estimates reported above.

Tables S17 and S18 summarize the item interface, reviewer instructions, assignments, and analysis units.

Table S18 Protocol summary for the 500-pair answer/provenance audit and separate 200-question citation-mechanism audit.

Protocol element	Specification
Reviewed item content	One benchmark question, one blinded system answer, and the supporting cited evidence returned by that system.

Protocol element	Specification
Item order and blinding	Single-system packets with system identity removed; reviewers inspect one answer at a time rather than a side-by-side ballot.
Answer-quality rubric	Four 1–5 Likert dimensions: correctness, completeness, safety, and actionability. Overall acceptability was rated separately as a binary judgment.
Citation-mechanism focus	Descriptive MedRouteGuard error instances—missing citations, citation mismatches, and unflagged conflicts—in a separate stratified sample of 200 matched questions.
Evaluator pools	Five board-certified specialists for answer quality and three experts for citation-mechanism classification.
Reviewer assignment	Every answer output received three independent ratings; 3,000 ratings total and 600 per reviewer; every mechanism-audit output was rated by all three citation experts.
Instructions	Penalize unsupported claims, omitted contradictions, temporal mismatches, and clinically risky phrasing even when the prose is fluent; escalate any serious safety disagreement.
Interface behavior	Present one answer at a time, require all rubric fields before submission, and conceal the paired system output until the item is locked.
Compensation and independence	Reviewers were uncompensated beyond normal salary, did not author sampled questions, and had no role in system implementation.
Paired analysis	The unique question was the unit for acceptable-rate differences; reviewer and question random intercepts were used for ordinal outcomes.

Together, the tables define what each reviewer saw, how ratings were assigned, which errors were penalized, and how repeated ratings were analyzed. This makes the human-evaluation design interpretable alongside the raw counts and uncertainty in Box 1.

S9 Common-Packet Robustness of the Medication-Safety Review

The independent review copied the same frozen answer-and-evidence packet into both blinded conditions whenever the gate-on and gate-off retrospective oracles selected the same route. This occurred for 166/180 pairs. Independent presentation produced balanced condition-only disagreements on these identical objects: 4 versus 4 for material error and 2 versus 2 for unsafe omission. Constraining each identical packet to one common outcome therefore removes repeated-rating variation without changing either paired risk difference.

Table S19 Exact reconstruction of the common-packet robustness analysis. Condition-only discordances are gate-on-only versus gate-off-only. Subtracting identical-packet discordances from the full discordance table leaves the 14 route- and output-discordant pairs. Bracketed effect intervals are 95% percentile intervals from the enumerated paired-bootstrap distribution over those 14 questions.

Endpoint	Full condition totals	Full condition-only discordance	Identical-packet discordance	Output-discordant sensitivity
Material evidence-validity error	7/180 versus 18/180	5 versus 16	4 versus 4	1 versus 12; active-pair reduction 78.6 pp [42.9, 100.0]; exact $P = 0.0034$.
Unsafe omission	4/180 versus 13/180	3 versus 12	2 versus 2	1 versus 10; active-pair reduction 64.3 pp [28.6, 92.9]; exact $P = 0.0117$.

Within the 14 output-discordant pairs, condition totals were 2 versus 13 for material error and 1 versus 10 for unsafe omission. For material error, one pair was positive in both conditions, one was gate-on-only, 12 were gate-off-only, and none was negative in both; for unsafe omission, one was gate-on-only, ten were gate-off-only, and three were negative in both. The resulting gate-off minus gate-on risk differences were 78.6 percentage points [42.9, 100.0] and 64.3 points [28.6, 92.9], respectively. The paired error differences are therefore generated by changed outputs, while the balanced disagreements on repeated packets contribute uncertainty but no net condition effect.

Table S20 Descriptive request-to-first-generated-token latency for the non-matched implementations. These rows provide implementation context and are not part of the matched Q2 comparison in the main text.

Implementation	P50 (s)	P95 (s)	P99 (s)	Study role
TextRAG	2.14	4.87 [4.61, 5.14]	7.21	Study-specific text-only comparator for Q1 output auditing.
GraphRAG-OD	7.83	18.42 [17.21, 19.68]	27.61	Contextual query-time graph builder.

S10 Governance, Model/Data Documentation, and External Assets

The work uses public biomedical sources, commercial infrastructure, and existing software systems including Neo4j, Redis, Faiss, Elasticsearch, LangChain, ColBERT, BioBERT/ClinicalBERT-family encoders, and the Microsoft GraphRAG codebase. The principal named software and service terms in the reported stack are Neo4j Enterprise 5.15.0 under Neo4j’s self-managed commercial software agreement, Redis 7.2.3 under BSD-3-Clause, Faiss 1.7.4 under MIT, Elasticsearch 8.11.3 under Elastic License 2.0, LangChain under MIT, and Azure OpenAI GPT-4-Turbo under the Microsoft Azure product terms. TextRAG and OfflineGraph-Tuned use ColBERT reranking or linking, and the compiler initializes dense retrieval from the BioBERT/ClinicalBERT family. Table S21 documents the principal software, model-family, vocabulary, and data-source classes used to interpret and reproduce the released aggregate analyses. The primary foreseeable harm is overreliance on evidence that is incomplete, stale, or internally conflicting.

Table S21 converts that compact narrative into a concise manifest organized by experimental role and release status.

Table S21 Principal software and model assets used in the reported experiments.

Asset	Role	Terms / release status
Neo4j Enterprise 5.15.0	Graph store	Neo4j self-managed commercial software agreement.
Redis 7.2.3	Online caching	BSD-3-Clause.
Faiss 1.7.4	Vector retrieval	MIT.

Asset	Role	Terms / release status
Elasticsearch 8.11.3	TextRAG BM25 retrieval	Elastic License 2.0.
LangChain	Orchestration utilities	MIT.
Azure OpenAI GPT-4-Turbo (gpt-4-1106-preview)	Evidence summarization	Microsoft Azure product terms.
ColBERT reranker / linker	TextRAG reranking and OfflineGraph-Tuned entity-linking support	Upstream code and checkpoint terms apply; weights are not redistributed.
BioBERT / ClinicalBERT-family encoders	Dense biomedical retrieval and compiler initialization	Upstream checkpoint terms apply; model weights are not redistributed.
Microsoft GraphRAG implementation (v0.3.2)	GraphRAG-OD baseline	Upstream project terms apply.

Table S22 records the terminology and data-source classes whose redistribution conditions differ from software licenses.

Table S22 Terminology and data-source assets used in the reported experiments.

Asset	Role	Terms / release status
UMLS / SNOMED CT / RxNorm vocabularies	Normalization and entity linking	Upstream terminology licenses and jurisdiction-specific terms apply; terminology content is not redistributed.
Public biomedical document slice	Public evidence source class for the 127K public documents	Upstream publisher or government terms vary by source; the document corpus is not redistributed.
De-identified EMR slice	Private deployment evidence	Subject to institutional review and data-use controls; not redistributed.

Tables S21 and S22 separate software, services, vocabularies, public documents, and private data because those assets carry different licensing and redistribution constraints. Together with the intended-use, analysis-map, and cohort definitions in Tables S28 and S30, they provide a concise model-card and dataset-datasheet view of the evaluated system: asset versions, data provenance, deployment scope, release status, limitations, and monitoring risks associated with incomplete, stale, or conflicting evidence.

Extended Scientific Context, Results, Methods, and Discussion

The following sections provide the detailed technical context, results, methods, and boundary discussion supporting the concise main manuscript. They preserve denominators, estimands, robustness analyses, and implementation details needed for independent scrutiny and reproducibility.

S11 Relation to clinical RAG, evidence verification, and adaptive routing

S11.1 Closest decision boundaries

The closest technical lines of work protect different objects and act at different times (Table S23). Clinical RAG and graph QA improve the context or reasoning used to construct an answer [6–13]. Adaptive question routing chooses whether and how to augment a query, while ontology-guided conversational memory uses structured facts and budget-aware retrieval to improve context selection [14,15]. Groundedness and selective prediction score a completed output or decide whether to answer at a chosen risk–coverage point [16,17]. Temporal and provenance systems protect records, lineage, and maintained views [18–24], while adaptive query optimization predicts or reacts to physical-plan cost [25–27]. MedRouteGuard composes these concerns around a different decision unit: the complete answer-and-evidence route as a member of the latency benchmark.

Table S23 Feature-level contrast with the closest technical boundaries. “Not defined” means that the cited approach family does not use the feature as its primary protected unit or benchmark intervention; it does not imply that individual implementations cannot expose related metadata.

Approach family	Protected decision unit	Evidence decision point	Benchmark membership and failed-attempt timing
Clinical RAG and graph QA	Retrieved context, graph traversal, or executed answer	Retrieval and generation, followed by answer-quality evaluation	Complete-route membership and a shared failed-attempt clock are not defined as the intervention.

Approach family	Protected decision unit	Evidence decision point	Benchmark membership and failed-attempt timing
Adaptive RAG and question routing	Augmentation path, retriever, or retrieval-augmented model	Pre-execution choice optimized for expected answer quality, cost, or latency	Selects a route but does not use a post-execution clinical evidence contract to determine benchmark membership or retain failed-attempt time.
Groundedness and selective prediction	Generated output or answer/abstain decision	Post-execution support score or confidence-based selection	May filter answers, but does not define fixed candidate-route membership for latency credit.
Temporal, provenance, and incremental systems	Record, lineage object, temporal relation, or maintained view	Ingestion, query, or update-time validity and version checks	Does not join answer claims, route execution, and benchmark denominator in one decision.
Adaptive query optimization	Physical execution plan	Pre-execution prediction or runtime adaptation based on cost and cardinality	Optimizes among plans without a four-clause clinical evidence contract as the membership rule.
MedRouteGuard	Complete answer-and-evidence route	Capability and reuse checks before execution; returned-evidence, claim-alignment, conflict, and freshness checks after execution	Assigns fixed-candidate membership explicitly and charges failed attempts and fallbacks to the original request clock.

Established provenance, temporal metadata, abstention, cache validation, and routing mechanisms provide the component primitives. The technical novelty evaluated here is their use as one four-clause membership intervention on the complete question–route candidate, followed by a ranking-sensitivity test on fixed executions. This definition keeps the claim attached to the paper’s protected unit, pre/post decision timing, and same-clock benchmark treatment.

S11.2 Clinical RAG and biomedical graph QA

Long-context clinical RAG enlarges the usable evidence window [28]; rationale-guided retrieval focuses retrieval through an intermediate clinical rationale [29]; recurrent generation-augmented retrieval alternates retrieval and generation [30]; and multi-step radiology systems decompose retrieval and reasoning [31]. Specialty evaluations measure clinical answer quality in medical fitness, evidence-based neurology, and contrast-media consultation [32–34]. These advances strengthen an executed answer. The route-membership estimand instead asks whether that complete execution satisfied the evidence task required to enter the latency distribution.

Graph QA exposes entities and relations for multi-hop and aggregation questions through maintained knowledge graphs or query-time graph construction. MedRouteGuard uses a maintained graph because temporal intervals, conflict groups, and dependency versions must persist across requests, but its protected unit extends beyond graph retrieval to graph traversal, text fallback, generated claims, and returned evidence. The fixed-candidate audit then intervenes on membership of that unit rather than on graph-retrieval score or answer score.

S11.3 Provenance, temporal data management, and maintained state

Clinical knowledge graphs use UMLS, SNOMED CT, and RxNorm to connect drugs, diseases, observations, and outcomes across structured and unstructured sources [35–39]. MedRouteGuard promotes provenance, temporal validity, registered conflict groups, and dependency watermarks from retrieval metadata to a joint eligibility rule. Their conjunction determines whether the complete route receives latency credit, and the controlled gate-on/gate-off analysis measures the ranking consequence of that decision.

S11.4 Selective prediction, evidence verification, and routing

Selective prediction controls answer coverage, and post-hoc groundedness characterizes support for a generated output. Adaptive RAG routers choose augmentation paths, retrievers, or models to optimize expected accuracy, cost, or latency, while adaptive query optimizers choose physical plans using workload and cardinality information. MedRouteGuard separates route selection from benchmark membership: evidence capabilities and current reusable state constrain the candidate set before execution; returned evidence and claims determine final admission after execution; latency prediction orders only the pre-admissible set. Unlike routing objectives that score the selected path, the fixed-candidate intervention asks whether every completed route satisfied one

clinical evidence contract before it could receive latency credit, while retaining failed-route time. The resulting same-stack contrast—fixed executions with membership labels enforced or ignored—is the irreducible Q3 test.

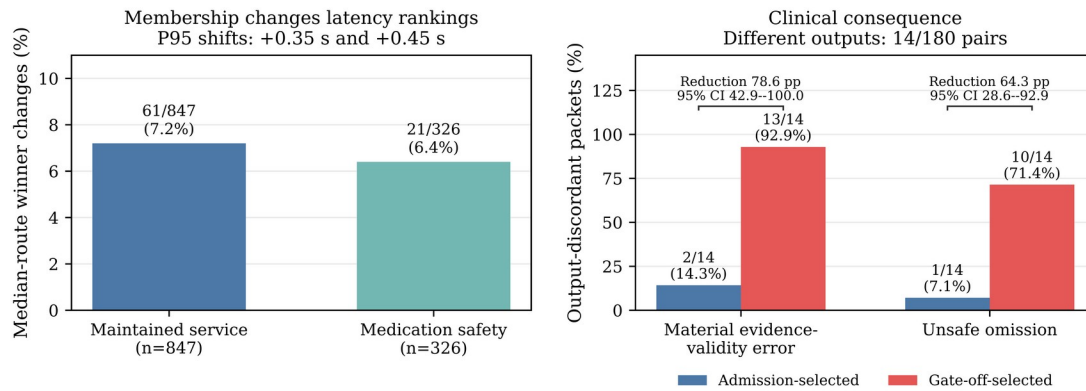
S11.5 Comparator contract

Each evaluated comparator identifies a different part of the paper’s claim. A study-specific text-only RAG comparator (TextRAG) combines BM25 retrieval, ColBERT reranking, passage citations, and the shared generator for the blinded Q1 answer and provenance audit. OfflineGraph-Tuned shares the maintained evidence graph, four-clause admission contract, questions, hardware envelope, cache budget, and endpoint protocol; its fixed graph-first policy makes it the matched deployed-system comparator for Q2. GraphRAG-OD represents query-time graph construction and provides a descriptive systems reference, while the internal fixed-policy configuration isolates route heterogeneity within the MedRouteGuard stack. Q3 identifies its estimand within one exhaustively replayed candidate universe by recomputing the retrospective oracle with membership labels enforced or ignored. This map keeps the study-specific output-quality comparator, matched deployed comparator, contextual implementation, and same-stack membership intervention attached to their respective claims.

S12 Results

The primary result is a denominator effect with a clinical consequence: under the prespecified every-successful-replay rule, route membership changed which completed candidates defined request-to-first-generated-token latency. In the 847-question maintained-service cohort, enforcing membership changed 61/847 (7.2%) median-route winners and increased retrospective-oracle P95 by 0.35 s [0.18, 0.53]. Every question retained at least one admitted candidate (0/847 zero-admissible questions), so both gate-on and gate-off P95 values use the same candidate-median first-token endpoint and denominator; the terminal-response fallback branch was not invoked in this Q3 cohort. The independent medication-safety workload reproduced 21/326 (6.4%) winner changes. Its selection rules produced different frozen packets for 14/180 reviewed questions; within these output-discordant pairs, admission-based selection reduced material evidence-validity errors from 13/14 to 2/14, an absolute reduction of 78.6 percentage points [42.9, 100.0], and unsafe omissions from 10/14 to 1/14, a reduction of 64.3 points [28.6, 92.9]. These 95% paired-bootstrap intervals propagate uncertainty across the 14 active question pairs; the 166 identical packet pairs contributed balanced rating variation.

Figure S5 summarizes the central evidence chain. The same-stack audit establishes ranking non-invariance on a fixed, exhaustively replayed candidate pool; the held-out expert audit validates the admission labels; and the independent workload connects the alternative selection rules to adjudicated packet outcomes. The deployed-system analyses then show that enforcing this evidence boundary remains compatible with strong answer quality, high coverage, and faster recorded processing milestones than the matched graph baseline.



The other 166/180 pairs contained identical frozen outputs and contributed balanced rating discordance.

Fig. S5 Synthesis across the two evaluation regimes. The maintained-service panel reports 61/847 winner changes and the +0.35 s paired P95 shift under the every-successful-replay membership rule; the independent medication-safety panel reports 21/326 winner changes and a +0.45 s paired P95 shift. Clinical bars show raw n / N outcomes among the 14/180 output-discordant packet pairs; brackets show 95% paired-bootstrap intervals for the absolute error reductions (78.6 [42.9, 100.0] percentage points for material error and 64.3 [28.6, 92.9] points for unsafe omission). The remaining 166/180 pairs contained identical frozen packets and contributed balanced repeated-rating variation

All online experiments ran on the fixed Azure deployment detailed in Supplementary Note S1. Request-to-first-generated-token latency was the historical primary systems metric, and Table S26 reports 95% bootstrap confidence intervals from 1,000 resamples. OfflineGraph-Tuned was the matched fixed graph-serving comparator. Always graph-first in Supplementary Table S2 and the –Cost-based planner row in Supplementary Table S5 denote the same internal fixed-policy diagnostic; the corresponding P95 values of 4.18 s and 5.49 s arise from the matched comparator and internal ablation, respectively.

Five board-certified specialists reviewed a blinded 500-pair answer and provenance sample, and three experts reviewed a separate blinded 200-question citation-mechanism sample comprising 400 query–answer items. Items were stratified across the five question families in Table S30. Reviewers scored correctness, completeness, safety, and actionability on five-point scales; provenance outcomes used returned evidence units

as the audit object. Tables S24 and S25 report the primary comparative results, while Box 1 provides raw distributions, counts, paired uncertainty, and reviewer-aware inference.

S12.1 Supporting output-quality validation under the fixed text-retrieval contract

Q1 tests whether the complete-route architecture also returns clinically usable, inspectable answers against the prespecified study-specific text-only RAG comparator. The controlled route-membership claim is identified separately by Q3.

Table S24 Blinded 500-pair answer-quality audit. Rows report the mean $\pm$ sample SD across 500 question-level consensus scores on a 1–5 Likert scale. Higher is better; Overall acceptable is the share receiving an affirmative separate overall judgment in this 500-pair audit; κ denotes Fleiss’ inter-rater agreement.

Metric	MedRouteGuard	TextRAG	κ
Correctness	4.72 $\pm$ 0.50	4.31 $\pm$ 0.74	0.78
Completeness	4.18 $\pm$ 0.72	3.87 $\pm$ 0.84	0.71
Safety	4.81 $\pm$ 0.41	4.52 $\pm$ 0.65	0.82
Actionability	4.04 $\pm$ 0.82	3.62 $\pm$ 0.94	0.69
Overall acceptable	94.2%	87.4%	0.75

Table S24 shows higher mean correctness, completeness, safety, and actionability for MedRouteGuard on this blinded sample. Overall acceptability was 94.2% for MedRouteGuard and 87.4% for the configured BM25–ColBERT–GPT-4-Turbo comparator, a paired difference of +6.8 percentage points [4.2, 9.5]. This comparison supports usability in the sampled Q1 contract. A separate matched 240-query audit against OfflineGraph-Tuned links quality directly to the Q2 latency comparator in Box 2 and the Supplementary Information.

Table S25 Blinded 500-pair provenance audit on the same answer sample as Table S24. Higher is better.

Metric	MedRouteGuard	TextRAG
Evidence precision	0.91	0.78
Evidence recall	0.84	0.71
F1	0.87	0.74

Metric	MedRouteGuard	TextRAG
Fleiss' κ	0.73	0.68

Each provenance metric treats one returned evidence unit as the audit unit: precision counts cited units judged relevant to answer claims, recall counts required support units that were actually returned, and the primary micro-averaged precision, recall, and F1 are computed after pooling TP, FP, and FN across the 500 audited answers. Evidence precision and recall were 0.91 and 0.84 for MedRouteGuard, compared with 0.78 and 0.71 for TextRAG (Table S25); evidence F1 was 0.87 versus 0.74. The agreement values indicate substantial consistency among reviewers. Together with the raw support-unit counts, the audited MedRouteGuard outputs showed more complete, inspectable support under the sampled Q1 protocol.

Box 1 | Statistical companion to the blinded 500-pair audit. Score distributions list counts at Likert levels 1/2/3/4/5 for MedRouteGuard versus TextRAG. Paired effects use the unique question as the analysis unit.

Analysis	Result
Acceptable answers	MedRouteGuard 471/500; TextRAG 437/500.
Discordant acceptable pairs	43 MedRouteGuard-only versus 9 TextRAG-only.
Paired acceptable-rate difference	+6.8 percentage points [4.2, 9.5]; exact McNemar $P = 2.04 \times 10^{-6}$.
Correctness distribution	0/1/10/117/372 versus 2/9/45/220/224.
Completeness distribution	1/6/70/248/175 versus 5/23/112/252/108.
Safety distribution	0/0/4/87/409 versus 1/4/24/176/295.
Actionability distribution	2/14/104/222/158 versus 10/42/164/196/88.
Provenance TP/FP/FN	840/83/160 versus 710/200/290; these counts yield the micro-averaged values in Table S25.
Overall acceptable agreement	Fleiss' $\kappa = 0.75$, 95% CI [0.71, 0.79].
Reviewer-aware pooled ordinal model	common odds ratio 2.31 [1.82, 2.92] for a higher MedRouteGuard rating, with dimension as a fixed effect and reviewer and question random intercepts.

The raw distributions reproduce the mean scores and sample SDs in Table S24. Paired analysis retained the positive acceptable-rate difference, and the reviewer-aware ordinal model yielded a common odds ratio of 2.31 [1.82, 2.92] for a higher MedRouteGuard rating. The separate 200-question mechanism audit identified 23 MedRouteGuard citation-error instances: eight entity-linking failures, six temporal-filter over-pruning errors, five conflict-group omissions, and four generation errors despite supplied evidence. Supplementary Table S12 and Figure S4 report this descriptive taxonomy with its instance-level denominator.

S12.2 Matched deployed-system milestones under shared evidence requirements

MedRouteGuard and OfflineGraph-Tuned shared the graph snapshot, questions, Azure worker classes, Redis cache budget, route-admission requirements, and latency protocol. They differed in serving policy: OfflineGraph-Tuned used a fixed graph-first policy, whereas MedRouteGuard selected among pre-admissible graph-first, hybrid, and text-fallback routes and applied the same post-execution contract.

Table S26 Matched request-to-first-generated-token latency across 847 benchmark questions. Both systems use the same evidence graph, admission contract, hardware envelope, cache budget, and endpoint protocol. Brackets show 95% bootstrap confidence intervals for P95. Lower is better.

System	P50 (s)	P95 (s)	P99 (s)
MedRouteGuard	1.42	3.24 [3.08, 3.41]	4.87
OfflineGraph-Tuned	1.53	4.18 [3.94, 4.43]	6.92

Relative to OfflineGraph-Tuned, MedRouteGuard reduced P95 latency from 4.18 s to 3.24 s and P99 from 6.92 s to 4.87 s on the same graph state (Table S26). The advantage was concentrated in tail latency. In the aggregate blinded matched audit (Supplementary Table S14), the paired acceptability difference was +2.1 percentage points [− 1.0, +5.4], and surfaced-provenance estimates favored MedRouteGuard.

The paired first-generated-token P95 difference for MedRouteGuard minus OfflineGraph-Tuned was − 0.94 s [− 1.19, − 0.69]. This deployed-system contrast uses the same 847-question denominator and admission contract; admitted-answer counts differed and are reported separately.

The request-to-first-generated-token endpoint ends at the retained generation event and is not a release-time measure. Box 2 reports complete-answer and complete-evidence endpoints under the same matched question, cache, request-order, and worker protocol.

The companion endpoints test whether the first-token latency advantage persists through complete answer and evidence delivery while preserving coverage and matched-audit acceptability.

Box 2 | Matched complete-answer, complete-evidence, coverage, and Q2 audit outcomes. Timing entries are P95 with paired-bootstrap 95% confidence intervals; coverage outcomes retain the 847-question denominator.

Outcome	MedRouteGuard	OfflineGraph-Tuned
Time to complete answer	6.82 s [6.51, 7.15]	8.06 s [7.69, 8.45]
Time to complete evidence bundle	5.74 s [5.43, 6.08]	7.12 s [6.76, 7.50]
Paired complete-answer P95 difference	− 1.24 s [− 1.58, − 0.91]	
Throughput at 10 s complete-answer target	8.6 QPS	7.0 QPS
Zero admissible set	6/847	9/847
Scoped evidence insufficiency	6/847	9/847
Abstention	9/847	13/847
Timeout or terminal failure	2/4,235 requests	4/4,235 requests
Answered coverage	838/847 (98.9%)	834/847 (98.5%)
Family coverage	186/239/166/137/110	186/238/165/136/109
Acceptable outputs in 240-query audit	225/240	220/240
Discordant acceptable pairs	11 MedRouteGuard-only versus 6 OfflineGraph-Tuned-only	

The complete-answer and complete-evidence results retained the direction of the first-generated-token comparison. The terminal flow was 838 admitted answers and nine no-answer abstentions for MedRouteGuard, versus 834 and 13 for OfflineGraph-Tuned. The six and nine scoped-insufficiency outcomes were subtypes of those abstentions and also carried the zero-admissible diagnostic. The 2/4,235 and 4/4,235 request-level terminal failures occurred among measured replays and overlapped questions that retained other successful replays. In the 240-question matched audit, 225 MedRouteGuard outputs and 220 OfflineGraph-Tuned outputs were acceptable; the paired +2.1-percentage-point difference [− 1.0, +5.4] supports compatibility of the observed tail-latency advantage with maintained output quality.

Figure S6 adds complete-answer throughput to these matched latency and coverage endpoints.

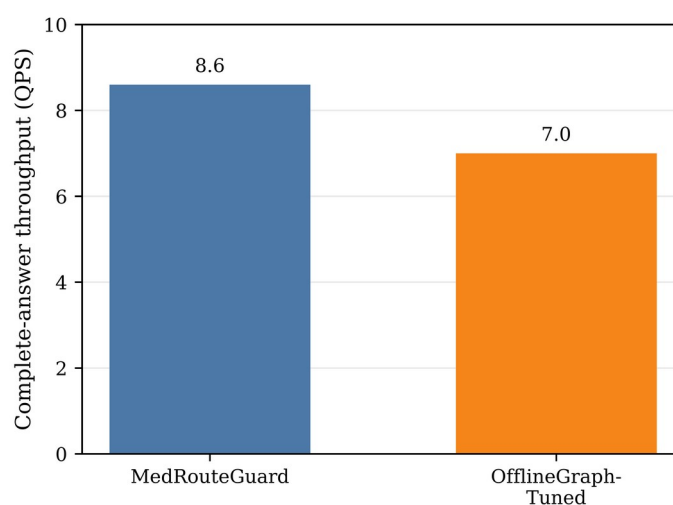


Fig. S6 Complete-answer throughput for the matched deployed-system comparison. At the prespecified 10 s target, MedRouteGuard sustained 8.6 QPS and OfflineGraph-Tuned sustained 7.0 QPS under the same graph, admission contract, cache budget, and worker envelope

Because OfflineGraph-Tuned also differs in graph-first policy, Neo4j hints, and ColBERT-based linking, the comparison is between complete deployed configurations on the same questions and evidence contract. Q3 separately quantifies the paired membership contrast obtained by enforcing versus ignoring prespecified route labels on a fixed, exhaustively replayed candidate universe. The 240-question blinded audit in Box 2 and Supplementary Table S14 provides the matched quality bridge.

S12.3 Controlled fixed-candidate intervention: admission changes latency winners

Q3 directly tested whether the latency ranking was invariant to the clinical task definition. We held the graph snapshot, prompts, hardware, candidate routes, gate computations, and latency estimator fixed. Every candidate was replayed, then the retrospective route oracle was recomputed with candidate-level admission labels enforced or ignored. Stable rankings would preserve both winner identity and the candidate-pool latency estimand. Table S27 tests that condition as a controlled membership comparison.

Table S27 Same-stack route-membership diagnostic on the preserved candidate pool. Graph snapshot, hardware, prompts, candidate plan families, gate computations, and latency estimator are held fixed; the retrospective oracle either enforces or ignores candidate-level admission labels when assigning benchmark membership. Clause counts are question-level, non-exclusive counts among speed-event questions and reflect associated candidate-level tags; they do not identify the clause state of the replay that produced a qualifying fast latency.

Metric	Gate-relaxation summary
Candidate-level observed-route-oracle P95	2.89 s with admission labels ignored versus 3.24 s with the every-successful-replay rule enforced.
Questions with an excluded-candidate speed event	98/847 (11.6%).
Candidate-level route-winner changes	61/847 (7.2%).
Candidate-level bucket-set changes (≤ 3 s, 3–5 s, 5–10 s, > 10 s)	83/847 (9.8%).
Candidate-level median winner gap	0.27 s.
Question-level, non-exclusive failed-clause tags associated with speed events	Provenance 44, temporal 31, conflict 19, freshness 28.
Q3 paired-latency companion. The analysis unit is the question; five measured replays are nested within each question–route candidate. Gate checks remain timed in both arms, so the contrast changes retrospective benchmark membership only.	
Paired quantity	Estimate
Gate-on marginal P95	3.24 s [3.08, 3.41]
Gate-off marginal P95	2.89 s [2.74, 3.05]
Gate-on minus gate-off P95	+0.35 s [+0.18, +0.53]
Questions in paired latency quantile	847/847
Missing candidate replays	11/12,705 replay measurements (0.09%); no question lost
Excluded-candidate speed event	98/847 (11.6%), 95% CI 9.6%–13.9%
Route-winner flip	61/847 (7.2%), 95% CI 5.7%–9.0%
Candidate-pool bucket-set change	83/847 (9.8%), 95% CI 7.9%–11.9%

The Q2 deployed-system P95 and Q3 gate-on observed-route-oracle P95 are distinct estimands. Both equal 3.24 s in the archived summaries, but the deployed planner and retrospective oracle need not select the same route for a question. The shared value is therefore reported with its analysis label wherever it recurs.

The ranking-invariance condition failed under the prespecified rule. The paired retrospective-oracle P95 increased by 0.35 s when the every-successful-replay membership rule was enforced. An excluded-candidate speed event occurred for 98 of 847 questions (11.6%), the median-latency route winner changed for 61 (7.2%), and the set of occupied candidate-latency buckets changed for 83 (9.8%). These events are non-nested: a replay-level speed event need not change a candidate median, and candidate-pool membership can change a latency bucket without changing the fastest median route. Provenance and temporal failures were most frequent among associated excluded candidates; freshness failures concentrated in update-sensitive questions, and conflict failures identified candidates with incomplete contradiction coverage. Applying the conservative consistency rule therefore changed the retrospective candidate-pool latency estimand and oracle winner. The clause counts are question-level and non-exclusive: one question can contribute to multiple clause rows when an associated excluded candidate violates more than one requirement, so $44 + 31 + 19 + 28$ can exceed 98 by construction. Supplementary Tables S7 and S8 report overlap counts and question-family concentrations, while Supplementary Table S9 repeats the route-membership diagnostic on a public-only slice.

The effect persisted under single-clause relaxations of the same fixed candidate pool. Ignoring provenance alone changed 34/847 (4.0%) winners and yielded a 3.02 s apparent P95; the corresponding values were 31/847 (3.7%) and 3.07 s for temporal scope, 16/847 (1.9%) and 3.14 s for conflict preservation, and 24/847 (2.8%) and 3.10 s for freshness. Winner changes were most frequent in temporal (12/112, 10.7%), safety (14/139, 10.1%), and aggregation questions (15/168, 8.9%), compared with entity lookup (6/187, 3.2%). The membership result therefore appears across every clause family and concentrates where clinical evidence composition is most demanding.

The fixed-replay aggregation sensitivity preserved the direction and localization of the effect. Requiring at least four of five successful replay labels changed 56/847 (6.6%) winners and yielded gate-on P95 of 3.21 s, a +0.32 s paired shift [0.15, 0.49] from gate off; the majority rule changed 49/847 (5.8%) winners and yielded 3.17 s, a +0.28 s shift [0.11, 0.44]. Thus, the conservative five-of-five rule produced the largest estimate, but neither the winner-change signal nor the positive P95 shift depended on a single failed replay.

Supplementary Tables S2–S5 report the online-planner, warm/cold cache, update-stress, and runtime-ablation diagnostics that explain why the denominator-corrected route set is operationally nontrivial. Table S27 and its paired companion define the direct candidate-pool boundary audit.

S12.4 Clinical validity of the admissibility labels

The held-out expert review tested whether deterministic route labels corresponded to clinically defensible evidence validity. Among 320 candidate routes, 41/320 (12.8%) required third-reviewer adjudication and agreement before adjudication was $\kappa = 0.82$ [0.77, 0.87]. Box 3 reports the raw confusion counts and uncertainty using clinically invalid as the positive reference-standard class.

Box 3 | Blinded clinical validity of route-admission labels. TP/FP/FN/TN refer to system rejection against the adjudicated expert standard; sensitivity is the share of clinically invalid routes rejected and specificity is the share of clinically valid routes admitted.

Clause	Counts	Sensitivity (95% CI)	Specificity (95% CI)
Provenance	35/3/2/40	94.6% [82.3, 98.5]	93.0% [81.4, 97.6]
Temporal	33/2/3/42	91.7% [78.2, 97.1]	95.5% [84.9, 98.7]
Conflict	29/4/4/43	87.9% [72.7, 95.2]	91.5% [80.1, 96.6]
Freshness	34/3/2/41	94.4% [81.9, 98.5]	93.2% [81.8, 97.7]
Pooled	131/12/11/166	92.3% [86.7, 95.6]	93.3% [88.6, 96.1]

Pooled PPV and NPV were 91.6% [85.9, 95.1] and 93.8% [89.2, 96.5]. The false-admission rate was 11/142, 7.7% [4.4, 13.3], and the false-rejection rate was 12/178, 6.7% [3.9, 11.4]. These predictive values and pooled error rates are conditional on the clause-, admission-, and severity-enriched 320-route sample. Three of the 11 false admissions were judged clinically material, corresponding to 3/320 (0.9%) of the enriched audit; none produced patient-facing output. Conflict preservation had the widest uncertainty and lowest sensitivity, identifying conflict-group coverage as the highest-yield target for further gate refinement.

Box 4 | Diagnostic evidence matrix for candidate-level membership changes. Each clause links the enforced property to its validation and observed failure mode.

Clause and enforced property	Validation and observed failure
Provenance: resolvable support for every material claim	Held-out experts and same-stack relaxation exposed false admissions with missing citation paths; support assembly contributes timed work.
Temporal validity: in-window evidence or scoped insufficiency	Held-out temporal routes and the independent time-versioned workload exposed routes that crossed the requested interval; filtering and fallback contribute timed work.
Conflict preservation: registered contradictions remain visible	Blinded safety and aggregation routes showed the widest clause uncertainty; incomplete conflict coverage triggered candidate exclusion, while preserved conflicts increased synthesis work.
Freshness: every reused object has a current dependency watermark	Injected-update stress quantified the latency and false-invalidation cost of conservative cache reuse; archived replays contained 0/1,968 observed stale-hit events.

Box 4 shows that each clause excludes a distinct out-of-contract route behavior and incurs observable work. The held-out audit validates the resulting labels, while Table S27, Box 5, and Supplementary Tables S4 and S6 connect those labels to ranking, evidence validity, and update behavior.

S12.5 Independent replication links route membership to medication-safety errors

The independent 326-question workload tested whether candidate-level ranking non-invariance recurred under a distinct question distribution and time-versioned public evidence graph. The analysis repeated the retrospective candidate-pool audit with the same controlled membership estimand. Gate-on P95 was 3.88 s [3.62, 4.17], compared with 3.43 s [3.19, 3.70] when admission labels were ignored and 4.73 s [4.39, 5.09] for the matched fixed graph-first comparator. Excluded-candidate speed events occurred for 32/326 (9.8%), the median-latency winner changed for 21/326 (6.4%), and the candidate-pool bucket set changed for 27/326 (8.3%). All 21 winner changes were route-to-route comparisons; 0/326 gate-on questions had a zero-admissible set, and no route-to-no-answer transition contributed to the latency contrast. The 8/326 versus 4/326 abstentions below are output dispositions of selected routes, not missing candidate-median endpoints.

The paired gate-on minus gate-off P95 difference was +0.45 s [+0.23, +0.68]. As in Q3, this is a candidate-pool boundary estimate conditional on the independent workload and route generator.

The second analysis was a blinded review of frozen output packets selected by the two retrospective oracles, estimating packet-level differences associated with those selections under this workload.

Box 5 | Independent medication-safety evidence-review outcomes. Clinical outcomes use the 180-question blinded review of packets selected by the gate-on and gate-off retrospective oracles; candidate-pool outcomes use all 326 held-out questions.

Outcome	Gate on	Gate off
Clinically material evidence-validity error	7/180 versus 18/180 (gate on versus gate off)	
Paired error difference	− 6.1 percentage points [− 10.4, − 1.9]	
Unsafe omission	4/180 versus 13/180 (gate on versus gate off)	
Answer acceptable	169/180	158/180
Contradiction visible when required	45/49	35/49
Temporal scope valid when required	58/61	48/61
Selected-route output abstention	8/326 versus 4/326 (gate on versus gate off)	
Median complete-evidence review time	4.7 minutes	5.6 minutes
Paired review-time difference	− 0.9 minutes [− 1.3, − 0.5]	
Reviewer agreement	$\kappa = 0.79$ [0.73, 0.85]	

Independent-crossover inference companion. Discordant counts are gate-on-only versus gate-off-only packet labels. Active-pair effect intervals are 95% percentile intervals from paired resampling of the 14 route- and output-discordant questions. Condition odds ratios compare packets selected by the gate-on and gate-off retrospective oracles; period and sequence are estimated jointly, whereas the carryover proxy comes from a separate sensitivity model that omits sequence. Binary paired tests are two-sided.

Paired quantity	Estimate
Material-error discordant pairs	5 versus 16

Paired quantity	Estimate
Material-error conditional odds ratio	0.31 [0.12, 0.85]; exact $P = 0.027$
Unsafe-omission discordant pairs	3 versus 12
Unsafe-omission conditional odds ratio	0.25 [0.07, 0.89]; exact $P = 0.035$
Route- and output-discordant review pairs	14/180 (7.8%)
Material errors in route-discordant pairs	2/14 versus 13/14 (gate on versus gate off)
Unsafe omissions in route-discordant pairs	1/14 versus 10/14 (gate on versus gate off)
Same-output material-error label discordance	8/166 (4 gate-on-only, 4 gate-off-only)
Same-output unsafe-omission label discordance	4/166 (2 gate-on-only, 2 gate-off-only)
	Material-error and unsafe-omission differences remain
Common-outcome sensitivity	– 11 / 180 and – 9 / 180; identical packets contribute zero paired difference.
Output-discordant material-error exact test	1 gate-on-only versus 12 gate-off-only; exact $P = 0.0034$.
Output-discordant unsafe-omission exact test	1 gate-on-only versus 10 gate-off-only; exact $P = 0.0117$.
Missing primary ratings	0/720
Question–condition labels adjudicated	26/360 (7.2%)
Period effect on material error	OR 1.08 [0.57, 2.04]; $P = 0.81$
Randomized order effect on material error	OR 0.93 [0.49, 1.78]; $P = 0.83$
Carryover proxy after washout	OR 1.11 [0.58, 2.14]; $P = 0.75$

The packet review localized the observed clinical association to the 14 questions for which the retrospective rules selected different frozen outputs. The common-outcome sensitivity removed repeated-rating variation from the 166 identical packet pairs without changing either paired risk difference. Within the 14 route- and output-discordant pairs, material errors were 2 versus 13 and unsafe omissions were 1 versus 10. The corresponding absolute reductions were 78.6 percentage points [42.9, 100.0] and 64.3 points [28.6, 92.9], with uncertainty propagated by paired bootstrap over the 14 active questions. Condition-only discordances favored admission-

based selection by 12 to 1 for material error and 10 to 1 for the prespecified secondary unsafe-omission endpoint. Median evidence-review time was 4.7 versus 5.6 minutes as a secondary condition-level difference. Period, sequence, and carryover diagnostics showed no material crossover imbalance. Replication across the maintained service and independent public workload therefore supports candidate-membership sensitivity as a clinically relevant benchmark diagnostic.

S13 Methods

S13.1 Method overview

MedRouteGuard joins two technical objects that are usually evaluated separately: an operational admission boundary for complete answer-and-evidence routes and a fixed-candidate intervention that measures how that boundary changes benchmark membership. The online system first screens route capabilities and reusable state, ranks only pre-admissible candidates, and then validates the returned claim–evidence packet before release. The audit reuses the same executed candidate universe and changes only whether the resulting admission labels determine latency credit. This design identifies the benchmark-denominator effect independently of differences between separately configured systems, while the deployed comparison tests routing performance inside the admitted set.

S13.2 Study design and intended use

We conducted a retrospective evaluation comprising a maintained-service benchmark, a held-out clinical validation of route-admission labels, and an independent time-versioned medication-safety workload. Maintained-service interactions and primary-benchmark question construction covered 1 January–31 December 2024; the independent workload used the later, chronologically held-out period specified below. All primary benchmark runs used the locked graph state described below. The intended use is bounded evidence review by a qualified analyst or clinician, with returned support and any abstention or insufficiency reason available for inspection. The primary methodological question was whether a latency ranking remained invariant when completed routes that failed the clinical evidence-task definition were excluded before benchmark credit.

S13.3 Analysis map and estimands

Table S28 fixes the population, contrast, endpoint, and analysis unit for each evidence block. It also separates the deployed-system estimand from the retrospective membership intervention.

Table S28 Analysis map. Each row is a separately defined evidence block; numerical comparisons remain attached to the listed population and unit.

Analysis	Population and contrast	Primary reported endpoint	Analysis unit
Q1	500 benchmark questions; MedRouteGuard versus TextRAG	Paired acceptability and micro-averaged provenance	Unique question for paired acceptability; support unit for provenance counts
Q2	847 benchmark questions; deployed MedRouteGuard versus OfflineGraph-Tuned	Paired request-to-first- generated-token P95 difference	Question-level median replay latency, with P95 over 847 questions
Q3	847 questions and 2,541 fixed candidates; admission labels enforced versus ignored on the same executions	Retrospective-oracle P95 difference and winner flips	Question; candidate latency is the median of successful measured replays
Label audit	320 clause-, decision-, and severity-enriched routes; deterministic label versus blinded specialist standard	Sensitivity for invalid routes rejected and specificity for valid routes admitted	Candidate route
Medication boundary	326 held-out questions and 978 fixed candidates; admission labels enforced versus ignored	Retrospective-oracle P95 difference and winner flips	Question
Medication packets	180 paired frozen packets selected by the two retrospective oracles	Paired material evidence- validity error difference	Question-condition pair; 14 pairs had different routes and outputs

Q2 and Q3 each yield a value of 3.24 s at one reported operating point, but they are not interchangeable: Q2 is the deployed MedRouteGuard first-generated-token P95, whereas Q3 is the gate-on retrospective-oracle P95 over the fixed candidate replay universe.

S13.4 Clinical evidence graph

The maintained labeled property graph represents canonical clinical entities and evidence-bearing relationships. Entity nodes include drugs, conditions, interventions, trials, outcomes, and safety findings. Evidence edges carry a source identifier, source span or structured-record locator, temporal validity interval, extraction confidence, update timestamp, and optional conflict-group identifier. Derived edges summarize recurring evidence patterns but retain links to their supporting records. The evaluated 31 December 2024 snapshot contained 1,521,847 nodes, 3,814,293 evidence edges, and 412,058 derived edges assembled from public biomedical documents and a governed de-identified clinical-data component.

S13.5 Construction and quality control of gate inputs

Source resolvability was established at ingestion by requiring a stable source identifier and either a document span or governed-record locator; identifier resolution and locator syntax were checked independently of route execution. Validity intervals were derived from publication, effective, supersession, and encounter dates under source-specific rules, with open intervals permitted only when the source asserted continuing validity. Missing or contradictory dates remained unknown rather than being imputed. Answer-relevant conflict groups linked normalized claims that shared a subject and relation but differed in polarity, value, recommendation, or applicable population; a clinical curator reviewed group membership and relevance when the reference standard was locked.

Every cached subgraph or derived summary stored the transitive closure of source versions, graph-edge update sequences, and conflict-group identifiers on which it depended. The object watermark was fixed at materialization, checked transactionally at request time, and invalidated when any dependency advanced. Schema checks, interval checks, conflict-group review, and dependency-closure checks were completed before the benchmark lock by staff who did not set the gate thresholds. Updates after lock advanced the relevant watermarks but did not rewrite historical route labels. An absent locator, interval, conflict status, or dependency could therefore produce an explicit unknown-state rejection; it could not pass through a favorable default.

S13.6 Logical request compilation

A question q is compiled into

$$L(q) = (E, R, T, O),$$

where E contains linked entities, R contains relation and aggregation predicates, T specifies the requested validity interval, and O specifies the evidence-output policy. The output policy determines whether every claim requires edge-level support or whether passage-level support is permitted when graph coverage is incomplete. Compilation failures, unresolved temporal scope, and low-confidence entity links are retained as explicit routing signals.

S13.7 Candidate routes

MedRouteGuard enumerates three route families. A graph-first route is used for well-linked, low-fan-out questions. A hybrid route combines graph traversal with dense text retrieval when linking is ambiguous or when the graph contains incomplete support. A text-fallback route retrieves source passages when graph coverage is insufficient but document-level provenance remains available. Candidate generation is separated from route admission so that each family is evaluated under the same evidence constraints. Together, the families create a heterogeneous route universe in which evidence eligibility and latency-aware selection can be tested independently.

S13.8 Pre- and post-execution route admission

A physical route P must satisfy four evidence conditions:

1. Provenance: every answer-bearing claim is linked to returned evidence with a resolvable source identifier and location;
2. Temporal validity: every evidence object used by an admitted answer falls within the requested time scope;
3. Conflict preservation: known contradiction-bearing evidence is retained or flagged rather than silently removed; and
4. Freshness: cached subgraphs and summaries remain valid under dependency-aware update watermarks.

Admission consumes categorical execution records rather than thresholding a composite quality score. For this decision, an answer-bearing claim is a minimal proposition that asserts a clinical entity–relation–value or polarity, recommendation, population qualification, temporal statement, or quantitative result. Conjoined clauses are separate claims when they can differ in support, polarity, population, or time; headings and non-assertive formatting are not claims. The generation record retains each claim and the evidence identifiers used to

construct it before the answer is rendered. Rendering is restricted to those retained records: each visible declarative span carries its originating claim identifier, and a deterministic round-trip pass resegments the rendered answer under the same proposition rules. Admission requires a one-to-one match between rendered assertions and retained claim identifiers; an unmatched, duplicated, or newly introduced assertion fails provenance with PROV_UNRESOLVED. Thus, “every claim” refers to every assertion visible in the released answer rather than only the generator’s pre-render list.

Claim–evidence alignment is deterministic at the admission boundary. A graph evidence object aligns when its registered normalized subject, relation, polarity or value, population qualifier, and temporal qualifier agree with the applicable claim fields. A passage aligns when its cited source span is registered to the same normalized proposition. Every aligned object must also have a resolvable source identifier and location. Mixed, missing, or unresolved fields yield *unknown*, not a favorable alignment. The entity-linker’s confidence score does not enter *Adm*: the locked compiler emits a resolved canonical identifier or an unresolved routing signal, and unresolved cases expand to hybrid or text-fallback candidates. Final admission is determined from the returned claim–evidence packet, so there is no unreported learned confidence cutoff in the four-clause gate.

An answer-relevant conflict group is one whose normalized subject and relation match a compiled request predicate or an answer-bearing claim and whose members differ in polarity, value, recommendation, or applicable population. This rule is applied to the conflict groups registered in the locked graph snapshot; a missing group status is unknown. Table S29 specifies the complete clause logic and unknown-state propagation.

Table S29 Operational route-admission contract. All listed conditions are conjunctive. Unknown inputs fail the corresponding clause and remain visible through the emitted reason-code set.

Clause	Pre-execution pass rule	Post-execution pass rule	Unknown/failure output
Provenance	The route can return the source identifier and location required by $O(q)$.	Every answer-bearing claim has at least one aligned returned evidence object, and every such object has a resolvable source identifier and location.	Any missing claim link, unresolved identifier, location, or alignment field fails the clause and emits PROV_UNRESOLVED.

Clause	Pre-execution pass rule	Post-execution pass rule	Unknown/failure output
Temporal validity	The route can enforce	Every evidence object used by an admitted claim	Missing, contradictory, or out-of-scope temporal metadata fails the replay and emits TEMP_SCOPE.
	$T(q)$, propagate unknown dates, and emit a typed failure when in-scope support is unavailable.	intersects the requested validity interval. A scoped insufficiency response is a timed no-answer outcome, not a temporal-clause pass.	
Conflict preservation	The route can enumerate every registered answer-relevant conflict group under the subject-relation rule above.	Every active relevant group is represented in the returned evidence or explicitly flagged in the answer packet.	Missing group status or incomplete relevant-group coverage fails the clause and emits CONFLICT_UNKNOWN.
Freshness	Every candidate reuse object has a complete dependency closure, and no dependency version exceeds its stored watermark at request-time validation.	The same dependency closure is checked immediately before release; no dependency may have advanced since the request-time check.	A missing dependency or watermark emits FRESHNESS_UNKNOWN; an advanced dependency emits FRESHNESS_STALE. Both fail the clause.

The online procedure separates what can be established before execution from what can be observed only after retrieval and generation. Pre-admission requires every pre-execution rule in Table S29:

$$PreAdm_{qt}(P) = C_{prov}^{cap} \wedge C_{temp}^{cap} \wedge C_{conflict}^{cap} \wedge C_{fresh}^{pre}.$$

After execution r , post-admission applies every returned-packet rule:

$$PostAdm_{qptr} = C_{prov}^{ret} \wedge C_{temp}^{ret} \wedge C_{conflict}^{ret} \wedge C_{fresh}^{final}$$

At the fixed graph and watermark state, the final replay label is $Adm_{qpr} = PreAdm_{qt}(P) \wedge PostAdm_{qptr}$. Every failed or unknown term contributes its code to a set; codes are not overwritten by a later failure. If no pre-admissible route exists, or if every executed route fails post-admission, the system returns a scoped evidence-insufficiency response or abstention with the accumulated reason-code set.

The replay label, online fallback policy, and Q3 candidate label are distinct. Adm_{qPr} records whether one completed replay passes all four clauses. Online serving tries the next ranked candidate after an observable replay failure and retains all elapsed time; only a passing replay releases an admitted answer. When the candidate list is exhausted, the terminal scoped-insufficiency response or abstention is timed but receives neither admitted-answer status nor candidate-route credit. Q3 subsequently forms Adm_{qP} by intersecting the replay labels of every successful measured replay for candidate P ; that retrospective aggregation does not alter online fallback behavior.

S13.9 Latency-aware planner

For each pre-admissible route, MedRouteGuard estimates

$$\hat{T}(P) = T_{compile} + T_{link} + T_{cache} + T_{graph} + T_{vector} + T_{post} + T_{llm}.$$

The estimator uses graph depth, sampled fan-out, predicate selectivity, retrieval depth, cache state, and recent component latency. It produces the ordered list

$$\Pi(q) = \text{argsort}_{P: PreAdm_{qt}(P)=1} \hat{T}(P).$$

The online policy executes routes in $\Pi(q)$ and releases the first route whose returned packet satisfies post-admission:

$$P^{online}(q) = \text{first} \{ P \in \Pi(q) : PostAdm_{qPtr} = 1 \}.$$

The planner therefore ranks only capabilities and state known before execution; final admission remains an observed property of the completed answer–evidence packet.

This ordering defines the study’s enforcement boundary. The protected object is the complete answer-and-evidence route as a benchmark participant, rather than the generator, graph, or cache in isolation. The four clauses prevent a candidate that omits support assembly, crosses the requested interval, suppresses a registered contradiction, or reuses invalidated state from receiving final route credit. The planner and route families provide the executable instantiation used to measure that boundary.

S13.10 Operational admission sequence and measured object

Admission used a two-stage procedure whose rules were fixed before benchmark execution. First, the compiler resolved intent, entities, requested relations, temporal scope, and evidence-output requirements; candidate enumeration then produced graph-first, hybrid, and text-fallback plans. Pre-admission removed plans that could not expose resolvable source locations, enforce the requested interval, enumerate registered conflict groups, or validate every reused dependency watermark. The cost model ranked only the remaining plans. The highest-ranked plan was executed, and the returned evidence bundle was checked before generation for temporal scope, conflict coverage, source resolution, and watermark currency. Generation was constrained to that bundle, after which claim–evidence alignment and a final watermark check determined post-admission. A route entered the benchmark only after both stages succeeded.

The clock started at original request acceptance and was never restarted after route rejection. The execution record retained separate timestamps for request-to-first-generated-token, complete answer, complete evidence, and terminal response. First-generated-token is the historical timestamp of the first token produced by the route that eventually passed post-admission; it is determined from the completed trace and is not the user-visible release time. Output remained buffered until generation, claim alignment, evidence validation, and final freshness checks completed. Complete-answer ended after the final token and claim alignment, while complete-evidence ended after every evidence object and diagnostic had been validated and serialized. Evidence assembly and generation could proceed in parallel, so either completion timestamp could occur first. The instrumented trace includes compilation, candidate enumeration, pre-admission, cost estimation, retrieval or graph execution, post-admission, generation, serialization, failed attempts, and fallbacks; each reported endpoint includes only work completed before its own timestamp.

Missing metadata never passed by default. An unresolved source location emitted `PROV_UNRESOLVED`; absent or out-of-scope temporal metadata emitted `TEMP_SCOPE`; incomplete conflict-group coverage emitted `CONFLICT_UNKNOWN`; and an absent or advanced dependency watermark emitted `FRESHNESS_UNKNOWN` or `FRESHNESS_STALE`. The planner broke equal predicted costs by lower recent latency variance, then by the fixed graph-first, hybrid, and text-fallback order. Online requests executed the best predicted pre-admissible route and evaluated the next candidate only after an observable post-admission failure. If no candidate completed final admission, the service returned a scoped

evidence-insufficiency response when the unmet requirement was known and otherwise abstained with accumulated reason codes.

Q3 retrospective candidate-pool boundary audit

Q3 tested ranking invariance separately from the deployed online policy. The diagnostic would be null if enforcing the clinical task definition left the route winner and candidate-pool latency distribution unchanged. For each of the 847 locked questions, the compiler enumerated the graph-first, hybrid, and text-fallback routes before any diagnostic replay, yielding a fixed universe of 2,541 question–route candidates. We then attempted every candidate once for route-local warm-up and five times for measurement, defining 12,705 scheduled replay measurements. The graph snapshot, prompt bundle, GPT-4-Turbo version, Azure workers, Redis state, dependency watermarks, randomized request blocks, and concurrency were synchronized across routes. All admission checks were executed and timestamped during every replay, and their computation was retained in both retrospective arms. The headline candidate latency was the median request-to-first-generated-token timestamp across successful measured replays. Candidate-level membership used the conservative prespecified intersection rule

$$Adm_{qP} = \bigwedge_{r \in R_{qP}^{success}} Adm_{qPr},$$

so a question–route candidate entered the gate-on oracle only when every successful measured replay satisfied all four clauses. “Gate off” removed this candidate-level membership restriction in the retrospective analysis while retaining gate-computation time and the complete executed replay set.

The intersection rule defines the prespecified conservative candidate-consistency estimand rather than average replay eligibility: one observed evidence-contract failure excludes a route candidate from benchmark credit, whereas median latency summarizes its execution time across successful replays. Its target is consistency over the observed replay set; replay-level label invariance is a separate estimand. A fixed-execution aggregation sensitivity replaced the intersection with thresholds requiring at least 80% or a simple majority of successful replay labels (nominally four-of-five and three-of-five); threshold ceilings were applied to candidates with a missing replay. To test whether the membership result depended on the joint four-clause label, we also recomputed the same candidate-level oracle while relaxing provenance, temporal, conflict, or freshness labels one at a time. Both sensitivities reused the fixed executions and timing summaries; no route was re-run or re-timed for a sensitivity condition.

Let $\tilde{T}_{qP}$ be the median observed request-to-first-generated-token latency across successful measured replays of candidate P for question q , and let $b(\cdot)$ map a latency to the prespecified Table S27 bucket. If at least one candidate was admitted, the gate-on diagnostic winner was $P_q^{on} = \arg \min_{P: Adm_{qP}=1} \tilde{T}_{qP}$; otherwise $P_q^{on} = \emptyset$ denoted the timed scoped-insufficiency or abstention outcome after all candidates failed admission. Its observed terminal-response latency was included in the gate-on question summary. The gate-off winner was $P_q^{off} = \arg \min_{P: Exec(P)=1} \tilde{T}_{qP}$. Ties at recorded precision were resolved by lower latency variance and then the fixed graph-first, hybrid, text-fallback order.

The three diagnostic counts were deliberately non-nested. An *excluded-candidate speed event* meant that at least one successful replay from a candidate excluded by the intersection rule completed faster than the gate-on question summary; this metric links replay latency to candidate membership and does not assign the candidate's clause failure to that particular replay. A winner flip meant $P_q^{on} \neq P_q^{off}$. A candidate-pool bucket-set change meant that $\{b(\tilde{T}_{qP}): Adm_{qP}=1\}$, augmented by the no-answer bucket when $P_q^{on} = \emptyset$, differed from $\{b(\tilde{T}_{qP}): Exec(P)=1\}$. Thus, bucket-set changes measure altered latency-distribution membership rather than only winner movement. One automatic retry was allowed after a 60-second timeout. A route required at least one successful measured replay to remain in the candidate-pool analysis; missing replays were not imputed, and a route with no successful replay was unavailable to both arms. The paired latency analysis retained 847/847 questions, with 11/12,705 replay measurements (0.09%) missing after terminal retry. The resulting paired retrospective route-oracle estimand quantifies ranking sensitivity to the prespecified candidate-consistency rule over the observed route universe.

S13.11 Dependency-aware cache validation

MedRouteGuard maintains subgraph and derived-summary caches. Each cache object records its contributing entities, edges, source versions, and a dependency watermark. Reuse is allowed only when no dependency has advanced beyond the recorded watermark and no newly registered conflict group intersects the cached evidence. Invalid objects are bypassed and scheduled for refresh. This mechanism prevents a stale cache hit from being treated as a legitimate latency improvement.

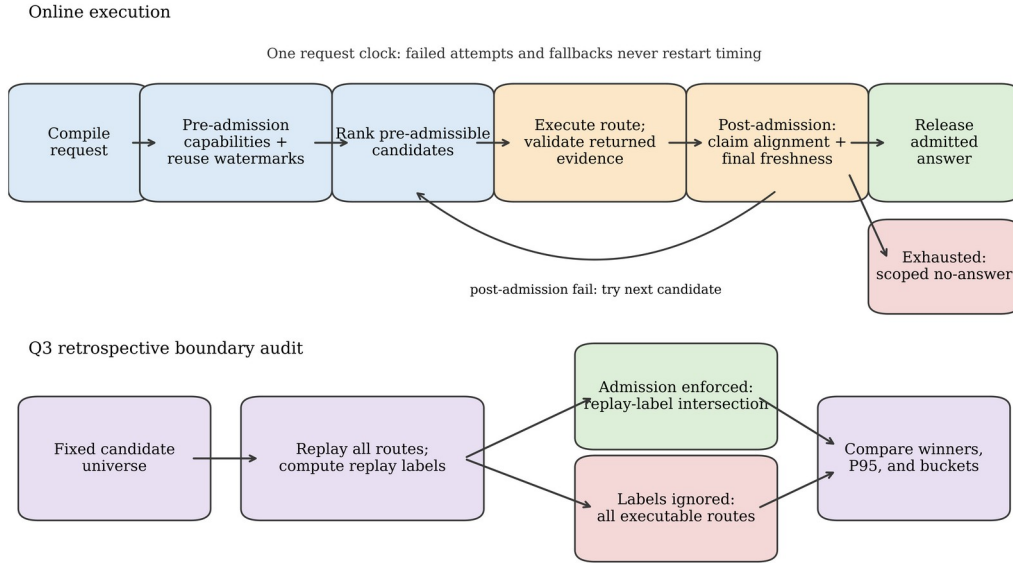


Fig. S7 MedRouteGuard online execution and Q3 replay design. The deployed policy compiles each request, applies pre-admission capability checks before ranking, validates returned evidence and buffered claims after execution, and continues to the next ranked route after an observable failure without resetting the clock. Q3 starts from one common replay-and-label node, branches in parallel to admission-enforced and labels-ignored membership analyses, and compares winners, P95, and candidate-pool buckets

Figure S7 separates online serving from the retrospective audit. Online serving ranks pre-admissible routes and observes post-admission after execution, whereas Q3 replays the fixed candidate universe and changes only which completed candidates receive retrospective benchmark membership.

S13.12 Retrospective evaluation cohort

The benchmark comprised 847 unique clinician-authored questions across the five prespecified families summarized in Table S30. Questions were authored during 1 February–30 November 2024 by 12 clinicians spanning internal medicine, clinical pharmacology, oncology, neurology, and radiology. The source frame contained 920 candidate questions. Eligible questions specified an evidence need that could be adjudicated against the locked snapshot and did not contain direct identifiers. We excluded administrative requests, questions requiring live patient management, unresolved reference answers, and paraphrases with the same intent–entity–time tuple; 73 candidate questions were removed before the test set was locked. Reference answers and required support units were constructed by two clinicians with third-reviewer adjudication. All 847 retained questions were unique after normalized semantic and entity-level deduplication.

This was a purposefully stratified benchmark rather than a consecutive or prevalence sample. The five family targets were fixed before evaluation to retain direct lookup while enriching multi-hop, aggregation, safety, and temporal tasks that exercise mutable-evidence constraints; all eligible, deduplicated questions within the locked targets were retained. Consequently, the reported 11.6%, 7.2%, and 9.8% rates are conditional on this question construction and candidate-route generator. The 500-question Q1 audit and 240-question matched Q2 audit were sampled without replacement within the five locked strata using a fixed randomization list.

Questions were evaluated against the fixed graph snapshot and system configuration. All systems shared the Azure worker classes and evaluation window detailed in Supplementary Note S1. Request-to-first-generated-token latency was the historical primary systems outcome; complete-answer and complete-evidence timing, answer correctness, completeness, safety, actionability, evidence precision, evidence recall, coverage, abstention, and constraint violations were additional outcomes.

Table S30 Deployment scope and retrospective evaluation cohort.

Item	Value
Graph nodes	1,521,847
Evidence edges	3,814,293
Derived edges	412,058
Public clinical documents	127,000
De-identified EMR encounters	58,000
Evaluation questions	847 clinician-authored questions
Question mix	187 entity lookup, 241 multi-hop, 168 aggregation, 139 safety, and 112 temporal questions

S13.13 Benchmark independence and configuration lock

Intent classification, entity linking, the 47 Cypher templates, planner features, cost estimation, and prompt wording were developed on 2,480 questions from service periods preceding the locked benchmark. Development records sharing a normalized intent–entity–time tuple with any of the 847 test questions were removed. Historical latency telemetry used for the cost model excluded final benchmark query identifiers, and no answer-quality, route-choice, or latency outcome from the 847-question set was used for feature selection,

threshold tuning, prompt revision, or baseline configuration. The compiler, graph snapshot, prompt bundle, generator version, candidate families, cache budget, and analysis plan were locked on 31 December 2024. The independent medication-safety workload used a separate chronological development set and remained sealed until its mapping and analysis rules were fixed.

S13.14 Comparators

Comparator roles were fixed by estimand. The study-specific text-only RAG comparator, TextRAG, combined BM25, ColBERT reranking, top-10 passage retrieval, passage-level citations, and the shared generator for the blinded answer and provenance audit. OfflineGraph-Tuned used the same evidence graph, admission requirements, hardware envelope, and evaluation questions but combined a fixed graph-first serving policy with hand-tuned Neo4j hints and ColBERT-based linking. It is the matched fixed graph-serving comparator for the Q2 deployed-system contrast. GraphRAG-OD represented a descriptive query-time graph-construction implementation, and an internal fixed-policy version of MedRouteGuard isolated route heterogeneity within the same service stack. Q3 supplied the paired candidate-pool boundary audit: every preserved candidate was replayed, and the observed route oracle was recomputed with and without label-based membership while all timed work was retained. Recent long-context, rationale-guided, recurrent, and specialty clinical RAG systems define the method-level context in Section S11; the numerical claims remain attached to the specified comparators and workloads evaluated here.

S13.15 Blinded clinical review

The 500-pair audit comprised 500 unique benchmark questions and one output from each of MedRouteGuard and TextRAG, for 1,000 independently presented system outputs. The family allocation was 110 entity, 142 multi-hop, 99 aggregation, 82 safety, and 67 temporal questions. Five board-certified specialists who had not authored the questions reviewed the outputs in system-blinded, single-answer packets containing the question, answer, and returned evidence. Each output received three independent ratings; 3,000 ratings total and 600 per reviewer. Reviewers completed 30 calibration items, discussed rubric disagreements without seeing study-system labels, and then evaluated disjoint randomized blocks. No reviewer rated both outputs from a pair consecutively.

Correctness anchors ranged from clinically wrong to fully correct; completeness from missing essential content to covering every required element; safety from likely to cause material harm to no material safety

defect; and actionability from unusable to directly usable for the stated evidence-review task. Scores of one and two denoted major defects, three denoted mixed or incomplete performance, four denoted clinically usable performance with only minor omissions, and five denoted no material defect. Dimension-level consensus was the median of the three ratings. Overall acceptability was a separate binary judgment: the answer had to fulfil the requested evidence-review task, support its key conclusions, and contain no consequential error or serious safety defect. At least two of three reviewers had to judge it acceptable, with safety disagreements adjudicated by a specialist who had not rated that output. Minor shortcomings could lower an individual dimension score without making the overall answer unacceptable. Missing ratings were not imputed; 0 primary ratings were missing.

For provenance, the reference answer was decomposed into atomic required support units. A returned unit was a true positive when it supported an answer claim and matched a required unit, a false positive when it was returned without claim-relevant support, and a false negative when a required unit was absent. Micro-averaged precision, recall, and F1 were the prespecified provenance summaries. A temporal error meant that a cited unit did not intersect the requested validity interval. A conflict error meant that an active, reference-standard conflict group relevant to the answer was omitted or presented without an explicit conflict flag.

The separate 200-question citation audit comprised 200 unique questions and two blinded outputs per question. Each of three citation experts inspected the returned support for every output, yielding $200 \text{ questions} \times 2 \text{ outputs} \times 3 \text{ experts} = 1,200 \text{ output-expert assessments}$. Its prespecified manuscript output was a descriptive mechanism taxonomy of MedRouteGuard citation-error instances—missing citation, citation mismatch, or unflagged conflict—rather than a second comparative provenance effect. Multiple error instances could arise from one output; the 500-pair audit above supplied the comparative provenance estimand. The matched Q2 audit comprised 240 unique benchmark questions, both MedRouteGuard and OfflineGraph-Tuned outputs, and two initial ratings per output, with a third specialist adjudicating discordant acceptability or safety judgments. All 960/960 scheduled primary ratings were complete, and 37/480 outputs (7.7%) required adjudication. The five-family counts reported in the Supplementary Information sum to 240; the unit for paired inference was the unique question, not the individual rating.

S13.16 Latency, coverage, and endpoint protocol

Each matched system processed the same randomized query blocks at a concurrency of 16. One route-local warm-up replay per question was excluded, followed by five measured replays per query; cache priming, graph snapshot, worker allocation, and request order were identical across systems. Runs used a 60-second request timeout and at most one automatic retry. A timed-out initial attempt remained in the complete-request duration when its retry succeeded; requests that still failed were counted as failures and were not silently removed from coverage denominators.

For each Q2 endpoint and system, the question-level value was the median across successful measured replays; P50, P95, and P99 were then calculated over the 847 paired question-level values. A terminally failed replay was not imputed, but its request-level failure remained reported; every question retained at least one successful terminal response in each system after the retry rule, so the paired latency population was 847/847.

The matched deployment used five Azure Standard E32ads v5 Neo4j 5.15.0 workers (32 vCPUs, 256 GiB RAM, AMD EPYC 7763, 128 GiB page cache), two Standard E16ads v5 Python 3.11/PyTorch 2.1.2 query workers (16 vCPUs, 128 GiB RAM), and one Standard NC24ads A100 v4 vector worker (NVIDIA A100 80 GB, CUDA 12.2, Faiss 1.7.4 HNSW). Redis 7.2.3 provided a 64 GiB cache. All answer generation used Azure OpenAI GPT-4-Turbo (gpt-4-1106-preview, temperature 0.2, 128K context, maximum 2,048 output tokens).

Request-to-first-generated-token spanned request acceptance through compilation, pre-admission, selected-route retrieval, evidence validation, generator prefill, and the first generated token from the route that later completed post-admission or from a structured no-answer response. It is a retained trace timestamp rather than the user-visible release endpoint. Time-to-complete-answer ended after the final token and claim alignment of the admitted answer or structured no-answer terminal response. Time-to-complete-evidence ended after every cited edge or passage, source locator, temporal annotation, conflict flag, or typed insufficiency or abstention reason with its accumulated gate diagnostics had been validated and serialized. The terminal-response timestamp marked release only after answer completion, evidence completion, serialization, and the final freshness decision; because complete-answer and complete-evidence could occur in either order, neither component timestamp is treated as terminal response. The Q2 latency quantiles used all 847 paired questions; each had at least one successful terminal response in both systems after the retry rule, although only 838 and 834, respectively, produced admitted answers. Coverage was the share of the 847 questions producing an

admitted answer; abstentions, scoped insufficiency responses, empty admissible sets, timeouts, and failures were reported separately. Throughput was measured at a fixed complete-answer service target, defined over those terminal responses, rather than inferred from first-generated-token latency.

Question-level terminal dispositions were mutually exclusive: admitted answer or no-answer abstention. “Scoped evidence insufficiency” was a reasoned subtype of abstention, and “zero admissible set” was an overlapping diagnostic flag on that subtype rather than another terminal category. Retry and timeout counts were request-level diagnostics across 4,235 measured requests and could overlap a question that still had successful replays; they did not replace the 847-question coverage denominator.

S13.17 Held-out clinical validity of admission labels

We sampled 320 candidate routes from a pool held out from clause design and threshold selection, with 80 routes per clause and 48 difficult multi-clause cases. The 48 multi-clause cases were a tagged subset of the 320 and were assigned to one prespecified primary clause for the mutually exclusive confusion-count rows. Sampling was stratified by query family, route family, system admission decision, and failure severity. Six clinical or evidence-synthesis specialists, none of whom authored benchmark questions or implemented the gate, reviewed each route’s request, evidence bundle, temporal metadata, conflict groups, and watermark trace while blinded to the system decision and latency. Two experts independently judged clinical admissibility and the violated clause; a third adjudicated disagreements.

The clinical reference standard required resolvable support for every material answer claim, evidence inside the requested time window, visibility of all known answer-relevant conflicts, and no reuse after a relevant dependency update. Reviewers could label the appropriate system outcome as admitted answer, scoped insufficiency response, or abstention. The primary estimands treated clinically invalid as positive and system rejection as a positive prediction: sensitivity measured invalid routes rejected, specificity measured valid routes admitted, false-admission rate was the fraction of clinically invalid routes admitted, and false-rejection rate was the fraction of clinically valid routes rejected. We also reported PPV, NPV, raw TP/FP/FN/TN counts, adjudication and agreement, and the severity of false admissions. The sample was deliberately enriched by clause, system decision, and severity; its predictive values and pooled error rates are therefore conditional on that design. This audit quantifies agreement between deterministic decisions and the clinical reference standard for the four specified clauses and available metadata inputs.

S13.18 Independent medication-safety workload

The independent workload used time-versioned FDA label revisions, ClinicalTrials.gov records, and open-access biomedical literature [40–42]. A separate clinical team screened 451 candidate questions; 29 duplicate, non-time-indexed, or non-adjudicable questions were excluded before chronological allocation. The remaining 422 questions comprised 96 earlier questions used to fix the mappings and 326 held-out evaluation questions: 74 drug-interaction, 68 contraindication, 61 dose-adjustment, 59 safety-warning temporal, and 64 evidence-conflict questions. The held-out period began after 1 July 2025, and evidence versions published after each question’s index time were unavailable to the route. Every held-out question retained at least one candidate under each retrospective oracle (0/326 zero-admissible gate-on questions), so both latency arms used the candidate-median first-token endpoint; no terminal no-answer latency entered either quantile. The cohort was purposefully stratified and enriched for medication-safety validity conditions, not sampled to estimate their natural prevalence; the 6.4% winner-change rate is conditional on this workload and candidate generator.

The same admissibility contract was applied to a separately built public evidence graph; gate-on and gate-off runs shared the graph state, GPT-4-Turbo generator, prompt bundle, candidate families, cache policy, and Azure worker envelope. The external boundary audit reused the full Q3 replay protocol: three candidate plan families produced 978 question–route candidates, each received one excluded warm-up and five measured replays, and 4,890 measurements were scheduled in randomized synchronized blocks. The 60-second timeout, one-retry rule, tie-break order, cache-state controls, candidate-availability rule, and missing-replay handling were identical to Q3. Four of 4,890 measured replays (0.08%) were missing after terminal retry; every candidate retained a successful measurement and all 326/326 questions entered the paired P95 analysis. Gate-on and gate-off P95 values were computed over question-level winners selected from candidate-median latencies.

A blinded two-period crossover of oracle-selected packets evaluated a prespecified stratified random sample of 180 held-out questions: 41 drug-interaction, 38 contraindication, 34 dose-adjustment, 32 safety-warning temporal, and 35 evidence-conflict questions. For each condition, the review packet used the successful replay of P_q^{on} or P_q^{off} whose first-generated-token latency was closest to that route’s median; a zero-admissible gate-on question used its recorded structured no-answer response. Candidate-level admission labels determined route eligibility only and were not imputed to the selected packet; packet validity was determined solely by blinded adjudication of the frozen output and evidence. When the two winners were the same route, the identical frozen answer-and-evidence output was copied into both blinded condition packets rather than regenerated. The

selected route and resulting output packet differed for 14/180 (7.8%) pairs; the remaining 166/180 pairs reused the same frozen output. Because the two periods were reviewed independently, labels were not forced to agree for identical packets: same-output pairs contributed 8/166 (4 gate-on-only, 4 gate-off-only) material-error label discordances and 4/166 (2 gate-on-only, 2 gate-off-only) unsafe-omission label discordances. These balanced same-output disagreements quantify rating variation rather than an output difference. The clinical estimand was therefore the paired difference in adjudicated packet labels under the two retrospective selection rules, not an online-policy treatment effect. This rule and the discordance analysis were fixed before outcome review. Two reference curators, blinded to system output, fixed endpoint applicability from the question and time-indexed reference standard: 49 questions required contradiction visibility and 61 required temporal-scope validity.

We added a common-packet robustness analysis to separate selection effects from repeated-rating variation. Each of the 166 identical frozen packets was constrained to have one common outcome label in both conditions, so only the 14 route- and output-discordant pairs could contribute a paired difference. The common label's value cancels from the paired contrast; this transformation leaves a paired risk difference unchanged whenever condition-only disagreements among identical packets are balanced, as they were for both material error and unsafe omission. We then repeated two-sided exact paired tests on the condition-only discordances within the 14 output-discordant pairs. The analysis changes no packet, route label, or primary adjudication; it tests whether the reported condition contrast survives removal of label variation on objects that were identical by construction.

Ten medication-safety clinicians each reviewed 36 question pairs. Every output received two independent primary ratings, and each clinician reviewed both conditions for an assigned question in counterbalanced order: 90 questions were gate on first and 90 gate off first, with balance within reviewer and question family. Each period comprised two blocks of 18 outputs per reviewer; paired conditions were separated by a one-week washout and never appeared in the same session. System labels, route labels, and latency were concealed, although recognition of repeated question content after washout could not be excluded.

A clinically material evidence-validity error was an adjudicated grade 2 or 3 defect on a four-level severity rubric: use of superseded or out-of-window evidence, unsupported resolution of a known conflict, or absent claim-level support that could change the evidence-review conclusion, urgency, contraindication assessment, or dose-adjustment decision. Grade 1 defects were non-decision-changing citation or completeness defects; grade 0 denoted no material defect. An unsafe omission was the prespecified subset in which an active contraindication, major interaction, required dose adjustment, or currently applicable warning in the reference standard was

absent and the remaining answer could plausibly clear an option that required avoidance, modification, or escalation. A clearly labeled insufficiency response was not an error unless it asserted an unsupported conclusion; omission of non-actionable background detail was not unsafe.

The two primary reviewers scored each output independently. Any disagreement on material-error status, unsafe omission, acceptability, or severity by two grades triggered adjudication by a third medication-safety clinician who had not rated that pair. Adjudicated question–condition labels formed the primary binary analysis data. Missing fields could not be submitted; an unusable session was rescheduled before unblinding, and no rating was imputed. Period and randomized sequence were evaluated jointly on the adjudicated labels. Because sequence and residual carryover are not jointly identifiable in a two-period, two-condition crossover, the period-by-prior-condition carryover proxy was evaluated in a separate sensitivity model that omitted sequence. Reviewer effects were retained in rating-level sensitivity and review-time models. Outcomes were clinically material evidence-validity error, unsafe omission, contradiction visibility, temporal validity, acceptability, abstention, complete-evidence review time, and reviewer agreement.

S13.19 Statistical analysis

We estimated P50, P95, and P99 latency over question-level replay summaries. Confidence intervals used 1,000 paired bootstrap resamples of query identifiers, with measured replays resampled within query, thereby preserving system pairing and replay clustering. Paired P95 differences were computed within each resample. The Q3 bootstrap used seed 20250117; all reported confidence intervals are two-sided 95% intervals, and confirmatory tests used two-sided $\alpha = 0.05$. Coverage, abstention, insufficiency, timeout, and failure rates retained the full question or request denominator. Quantile inference required independent questions, stable endpoint definitions, and exchangeable replays within question; replay-order and cache-state diagnostics were inspected before resampling.

Acceptability differences used the paired risk difference with a percentile bootstrap interval and a two-sided exact McNemar test on discordant questions. The common ordinal odds ratio pooled the four rating dimensions in a cumulative-link mixed model,

$$\text{logit}\{Pr(Y_{qsr d} \leq k)\} = \theta_k - \{\beta_{sys} S_s + \beta_{fam} F_q + \beta_{dim} D_d + u_q + v_r\},$$

where q , s , r , and d index question, system, reviewer, and dimension, and u_q and v_r are random intercepts. The no-system-by-dimension-interaction specification defines the reported pooled common odds ratio and is not interpreted as four separately estimated dimension effects. Fleiss' κ and its bootstrap interval quantified agreement. Provenance TP, FP, and FN counts were pooled for the reported micro-averaged precision, recall, and F1.

Gate-validity sensitivity, specificity, PPV, NPV, false-admission, and false-rejection intervals used Wilson score intervals. For clinically material errors and unsafe omissions, the primary crossover contrast used the conditional likelihood of the discordant adjudicated question pairs, with two-sided exact binomial tests. The common-packet robustness analysis subtracted condition-only discordances observed among identical packets from the corresponding full-sample discordance counts and repeated the same exact test on the remaining route- and output-discordant pairs. Because identical-packet discordances were balanced by condition, this operation preserved the full-sample paired risk differences while removing repeated-rating variation from exact inference. To quantify effect magnitude in the clinically operative subset, we calculated the gate-off minus gate-on risk difference among the 14 route- and output-discordant questions and enumerated its empirical paired-bootstrap distribution by resampling those question pairs with replacement. The reported 95% percentile interval is therefore conditional on selection changing the frozen output and propagates uncertainty over the active question pairs. Period and randomized sequence were evaluated in the secondary logistic mixed model

$$\text{logit} \{ \Pr(E_{qc} = 1) \} = \alpha + \beta C_c + \gamma R_{qc} + \delta S_q + b_q.$$

Here C_c denotes the gate-on or gate-off retrospective packet-selection rule, R_{qc} period, S_q randomized sequence, and b_q a question random intercept. The carryover sensitivity replaced S_q with the period-by-prior-condition proxy K_{qc} and was not interpreted as an independently identified carryover effect. A rating-level sensitivity model added crossed reviewer and question random intercepts before adjudication. Log review time used reviewer and question random intercepts. Marginal paired risk differences and review-time differences were obtained by standardizing fitted outcomes over the observed questions; intervals used a paired question bootstrap with reviewer assignments retained.

The prespecified primary estimands were Q1 paired acceptability, Q2 paired first-generated-token P95 difference, Q3 gate-on minus gate-off retrospective-oracle P95 difference, held-out gate sensitivity, and the paired material-error difference between oracle-selected packets in the independent crossover. Each addressed a

distinct research question in a separately defined analysis block; no joint rejection claim was made across the five estimands, and each retained its unadjusted 95% confidence interval. The 847-question systems blocks used the complete locked benchmark rather than a power-selected sample. The 500-pair audit targeted approximately three-percentage-point precision for an acceptability rate near 90%; the 240-pair audit was sized as a guardrail against a large paired quality loss; 80 routes per clause in the 320-route audit bounded clause-level sensitivity uncertainty; and 180 crossover questions targeted the prespecified material-error contrast under paired analysis. Unsafe omission was prespecified as a secondary crossover endpoint; its conditional odds ratio, 95% interval, and unadjusted exact paired P value are reported descriptively without a multiplicity-controlled significance claim. All other secondary endpoints are likewise descriptive estimates with 95% intervals and no threshold-based significance claim. Period, sequence, and carryover estimates were design diagnostics rather than confirmatory endpoints; their unadjusted P values were interpreted descriptively. Missing human ratings were not imputed, and failed system requests remained in coverage analyses. Analyses used Python 3.11 with SciPy 1.11.4 and statsmodels 0.14.1.

S13.20 Ethics, governance, and reporting guidance

The institutional Clinical Informatics Research Ethics Committee approved this retrospective study and waived individual informed consent. Use of the 58,000 encounters was limited to the approved governed environment and complied with the Declaration of Helsinki and institutional data-custodian controls. The question authors and expert reviewers evaluated system behavior as professional collaborators rather than research participants; participation was voluntary, uncompensated beyond normal salary, and reviewers reported no role in system implementation.

Sex, gender, race, ethnicity, and age were not model inputs or outcomes because the evaluation unit was a route over evidence metadata. The evaluated regime is route-level evidence validity for bounded evidence review. The manuscript follows CHART for model, prompt, query, performance, and safety reporting and STROBE for the applicable retrospective observational elements [43,44].

S14 Discussion

Clinical evidence requirements determine the population of routes represented by a latency benchmark. Ignoring the prespecified membership rule lowered retrospective-oracle P95 from 3.24 s to 2.89 s and changed 61/847 (7.2%) median-latency winners; the independent medication-safety workload reproduced 21/326 (6.4%) winner

changes. Executability and evidence-task membership are therefore empirically distinct properties, and conflating them can change both the tail estimate and the route credited as fastest.

The fixed-candidate intervention identifies that distinction more directly than a comparison between two independently configured systems. Q3 held the route universe, executions, graph state, prompts, workers, cache conditions, and gate computation constant, then changed only whether admission labels controlled membership. Relaxing provenance, temporal scope, conflict preservation, or freshness changed 34, 31, 16, and 24 winners, respectively. Four-of-five and majority replay aggregation retained 56 and 49 winner changes and positive P95 shifts, locating the largest estimate at the conservative boundary without making it depend on one failed replay. The concentration in safety, temporal, and aggregation questions ties the effect to evidence composition, while specialist validation at 92.3% sensitivity and 93.3% specificity connects the deterministic labels to clinically interpretable obligations. In contrast to post-hoc grounding or confidence-based abstention, the protected object here is the complete route, and the decision changes benchmark credit while preserving the timing of every failed attempt.

The medication-safety review shows why that denominator matters beyond a numerical ranking. The selection rules produced different frozen outputs for 14/180 questions. Within those active pairs, admission-based selection reduced material evidence-validity errors by 78.6 percentage points [42.9, 100.0] and unsafe omissions by 64.3 points [28.6, 92.9]; condition-only discordances favored admission-based selection by 12 to 1 for material errors ($P=0.0034$) and 10 to 1 for unsafe omissions ($P=0.0117$). The intervals are necessarily wider than the point estimates because they propagate uncertainty across only the questions whose output changed. The 166 identical packet pairs contributed balanced rating disagreements, and assigning each identical packet one common outcome left both full-sample paired differences unchanged. The clinical contrast is therefore localized to changed evidence packets rather than repeated ratings of identical outputs.

The deployed-system results establish that this evidence boundary is compatible with earlier recorded non-release processing milestones, high coverage, and strong output quality. Under the same graph, questions, hardware, cache budget, and four-clause contract, MedRouteGuard reduced first-generated-token P95 by 0.94 s and complete-answer P95 by 1.24 s relative to the matched fixed graph-serving comparator, with 98.9% versus 98.5% answered coverage. The matched 240-question audit retained the quality direction, and the separate text-retrieval comparison yielded 94.2% acceptable outputs and 0.87 evidence F1. Planning within the eligible set

therefore moved these component-level timestamps earlier without making a user-visible release-latency claim or weakening the shared evidence contract.

The estimates describe two purposefully stratified retrospective workloads, three route families, and the stated Azure configuration. The every-successful-replay rule targets conservative candidate consistency; replay-vote or average-eligibility rules define different reliability estimands. The packet analysis estimates retrospective selection rules, with its output effect concentrated in 14 pairs, and the four-clause gate evaluates available provenance, temporal, conflict, and watermark metadata rather than source truth or patient outcomes. The deployment comparison establishes ordering at first-generated-token, complete-answer, and complete-evidence milestones, whereas user-visible release occurs only after both completion paths and the final freshness decision; terminal-response quantiles were not a prespecified reported endpoint and no release-latency claim is inferred from the component milestones. Within these boundaries, the ranking effect replicated across both evidence services, specialists agreed with the admission object, and changed membership selected materially different evidence packets.

For mutable clinical evidence services, route eligibility should be specified before latency credit is assigned. MedRouteGuard operationalizes that boundary through pre-execution capability and reuse checks, post-execution provenance, temporal, conflict-preservation, and freshness validation, same-clock timing across failed attempts, and latency-aware ordering within the eligible set. Across the maintained-service and public medication-safety workloads, the resulting route-admission estimand changed first-token rankings, matched specialist validity judgments, and selected packets with 6.1 percentage points fewer material evidence-validity errors across all 180 reviewed questions; the 14 output-discordant pairs localized that association. Reporting membership, failed attempts, sensitivity, latency, quality, and coverage together therefore prevents a benchmark from crediting evidence omission, temporal mismatch, conflict suppression, or stale reuse as speed.

References

1. Aronson AR, Lang FM. An overview of MetaMap: Historical perspective and recent advances. *BMC Medical Informatics and Decision Making* [Internet]. 2010;17(3):229–36.
<https://doi.org/10.1136/jamia.2009.002733>
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* [Internet]. 2020;36(4):1234–40.
<https://doi.org/10.1093/bioinformatics/btz682>

3. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8.
<https://doi.org/10.18653/v1/W19-1909>
4. Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, et al. Cypher: An evolving query language for property graphs. In: Das G, Jermaine CM, Bernstein PA, editors. Proceedings of the 2018 international conference on management of data, SIGMOD conference 2018, houston, TX, USA, june 10-15, 2018 [Internet]. ACM; 2018. p. 1433–45. <https://doi.org/10.1145/3183713.3190657>
5. Microsoft. GraphRAG: Retrieval-augmented generation with graphs.
<https://github.com/microsoft/graphrag>; 2025.
6. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Albers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics [Internet]. 2015;16:138. <https://doi.org/10.1186/s12859-015-0564-6>
7. 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 (EMNLP-IJCNLP) [Internet]. 2019. p. 2567–77. <https://doi.org/10.18653/v1/D19-1259>
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature [Internet]. 2023;620:172–80. <https://doi.org/10.1038/s41586-023-06291-2>
9. Huang X, Zhang J, Li D, Li P. Knowledge graph embedding based question answering. In: Proceedings of the 12th ACM international conference on web search and data mining [Internet]. 2019. p. 105–13.
<https://doi.org/10.1145/3289600.3290956>
10. Saxena A, Tripathi A, Talukdar P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In: Proceedings of the 58th annual meeting of the association for computational linguistics [Internet]. 2020. p. 4498–507. <https://doi.org/10.18653/v1/2020.acl-main.412>
11. Yasunaga M, Ren H, Bosselut A, Liang P, Leskovec J. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics [Internet]. 2021. p. 535–46.
<https://doi.org/10.18653/v1/2021.naacl-main.45>

12. Hu Y, Lei Z, Zhang Z, et al. GRAG: Graph retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: NAACL 2025. Association for Computational Linguistics; 2025. p. 4145-4157. <https://doi.org/10.18653/v1/2025.findings-naacl.232>
13. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering [Internet]. 2024; <https://doi.org/10.1109/TKDE.2024.3352100>
14. Chen S, Zheng H, Cui L. When and how to augment your input: Question routing helps balance the accuracy and efficiency of large language models. In: Findings of the association for computational linguistics: NAACL 2025. Association for Computational Linguistics; 2025. p. 3621–34. <https://doi.org/10.18653/v1/2025.findings-naacl.200>
15. Cao S, Li R. Ontology-guided long-term agent memory for conversational RAG. Proceedings of Machine Learning and Systems. 2026;8:233-250. https://proceedings.mlsys.org/paper_files/paper/2026/hash/2fb4be70fc9668e9ec2c71b34fb127d4-Abstract-Conference.html
16. Geifman Y, El-Yaniv R. SelectiveNet: A deep neural network with an integrated reject option. In: Proceedings of the 36th international conference on machine learning [Internet]. PMLR; 2019. p. 2151–9. (Proceedings of machine learning research; vol. 97). Available from: <https://proceedings.mlr.press/v97/geifman19a.html>
17. 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. Association for Computational Linguistics; 2024. p. 150–8. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
18. Buneman P, Khanna S, Tan WC. Why and where: A characterization of data provenance. In: International conference on database theory [Internet]. 2001. p. 316–30. https://doi.org/10.1007/3-540-44503-X_20
19. Cheney J, Chiticariu L, Tan WC. Provenance in databases: Why, how, and where. Foundations and Trends in Databases [Internet]. 2009;1(4):379–474. <https://doi.org/10.1561/19000000006>
20. Snodgrass RT. Developing time-oriented database applications in SQL [Internet]. Morgan Kaufmann; 1999. Available from: <https://www.sigmod.org/publications/anthology/vol6/systems/timeOriented/timeOriented.htm>

21. Jensen CS, Snodgrass RT. Temporal data management. *IEEE Transactions on Knowledge and Data Engineering* [Internet]. 1999;11(1):36–44. <https://doi.org/10.1109/69.755613>
22. Gupta A, Mumick IS. Maintenance of materialized views: Problems, techniques, and applications. *IEEE Data Engineering Bulletin* [Internet]. 1995;18(2):3–18. Available from: <https://sigmod.org/publications/dblp/db/journals/debu/GuptaM95.html>
23. Fan W, Wang X, Wu Y. Incremental graph pattern matching. *ACM Transactions on Database Systems* [Internet]. 2013;38(3):18:1–47. <https://doi.org/10.1145/2489791>
24. McSherry F, Murray DG, Isaacs R, Isard M. Differential dataflow. *CIDR* [Internet]. 2013; Available from: https://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper111.pdf
25. Avnur R, Hellerstein JM. Eddies: Continuously adaptive query processing. In: *Proceedings of the 2000 ACM SIGMOD international conference on management of data* [Internet]. 2000. p. 261–72. <https://doi.org/10.1145/335191.335420>
26. Deshpande A, Ives Z, Raman V. Adaptive query processing. *Foundations and Trends in Databases* [Internet]. 2007;1(1):1–140. <https://doi.org/10.1561/1900000001>
27. Marcus R, Negi P, Mao H, Zhang C, Alizadeh M, Kraska T, et al. Neo: A learned query optimizer. In: *Proceedings of the VLDB endowment* [Internet]. 2019. p. 1705–18. <https://doi.org/10.14778/3342263.3342644>
28. Zhang G, Xu Z, Jin Q, Chen F, Fang Y, Liu Y, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*. 2025;8:239. <https://doi.org/10.1038/s41746-025-01651-w>
29. 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 (volume 1: Long papers)* [Internet]. Albuquerque, New Mexico: Association for Computational Linguistics; 2025. p. 12739–53. <https://doi.org/10.18653/v1/2025.naacl-long.635>
30. 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* [Internet]. Suzhou, China: Association for Computational Linguistics; 2025. p. 4006–33. <https://doi.org/10.18653/v1/2025.findings-emnlp.214>

31. Wind S, Sopa J, Truhn D, Lotfinia M, Nguyen TT, Bressem KK, et al. Multi-step retrieval and reasoning improves radiology question answering with large language models. *npj Digital Medicine*. 2025;8:790. <https://doi.org/10.1038/s41746-025-02250-5>
32. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*. 2025;8:187. <https://doi.org/10.1038/s41746-025-01519-z>
33. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. *npj Digital Medicine*. 2025;8:137. <https://doi.org/10.1038/s41746-025-01536-y>
34. Wada A, Tanaka Y, Nishizawa M, Yamamoto A, Akashi T, Hagiwara A, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. *npj Digital Medicine*. 2025;8:395. <https://doi.org/10.1038/s41746-025-01802-z>
35. Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic acids research* [Internet]. 2004;32(suppl_1):D267–70. <https://doi.org/10.1093/nar/gkh061>
36. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* [Internet]. 2006;121:279–90. Available from: <https://pubmed.ncbi.nlm.nih.gov/17095826/>
37. Liu S, Ma W, Moore R, Ganesan V, Nelson SJ. RxNorm: Prescription for electronic drug information exchange. *IT Professional* [Internet]. 2005;7(5):17–23. <https://doi.org/10.1109/MITP.2005.122>
38. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics* [Internet]. 2008;41(5):706–16. <https://doi.org/10.1016/j.jbi.2008.03.004>
39. Breit A, Ott S, Agibetov A, Samwald M. OpenBioLink: A benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* [Internet]. 2020;36(13):4097–8. <https://doi.org/10.1093/bioinformatics/btaa274>
40. U.S. Food and Drug Administration. FDALabel: Full-text search of drug product labeling. U.S. Food and Drug Administration, National Center for Toxicological Research; <https://nctr-crs.fda.gov/fdalabel/ui/search>; 2024.
41. U.S. National Library of Medicine. ClinicalTrials.gov. <https://clinicaltrials.gov/>; 2026.

42. National Library of Medicine. PubMed central open access subset [Internet]. 2024. Available from: <https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>
43. Huo B, Collins G, Chartash D, Thirunavukarasu A, Flanagan A, Iorio A, et al. Reporting guideline for chatbot health advice studies: The CHART statement. BMC Medicine. 2025;23:447. <https://doi.org/10.1186/s12916-025-04274-w>
44. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. The Lancet. 2007;370(9596):1453–7. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X)