Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Every-successful-replay route admission changes first-token latency rankings in clinical question answering

Rui Li
rui.li@hillresearch.ai
Hill Research
Jason Zhao
Hill Research
Shuang Cao
Hill Research
Alexandre Duprey
Hill Research
Ruihua Liu
Hill Research

Research Article

Keywords: clinical question answering, retrieval-augmented generation, evidence provenance, benchmark validity, medical informatics

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10990749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

To determine whether explicit evidence-admission requirements change latency rankings and the evidence selected for clinical review in clinical question answering.

Methods

MedRouteGuard checks provenance, temporal validity, conflict preservation, and dependency freshness for complete answer-and-evidence routes. We replayed 2,541 routes for 847 clinician-authored questions and compared rankings with admission labels enforced or ignored while holding executions fixed. Supporting evaluations included a blinded 320-route label audit, a matched deployed-system comparison, and an independent public medication-safety workload with 326 held-out questions and blinded review of 180 paired packets.

Results

Admission changed 61/847 (7.2%) route winners and shifted retrospective-oracle 95th-percentile request-to-first-generated-token latency from 2.89 to 3.24 s. Specialist validation yielded 92.3% sensitivity and 93.3% specificity. The independent workload showed 21/326 (6.4%) winner changes. Admission-selected packets had 7/180 versus 18/180 material evidence-validity errors (paired difference, -6.1 percentage points; 95% confidence interval, -10.4 to -1.9); among 14 output-discordant pairs, errors were 2/14 versus 13/14. Matched deployed first-token P95 was 3.24 versus 4.18 s, with 98.9% versus 98.5% answered coverage.

Conclusions

Evidence-task eligibility altered benchmark rankings and reviewed packets. First-generated-token timing is an internal processing milestone; the retrospective results do not establish patient benefit. Clinical latency benchmarks should specify route-admission rules. A public, offline companion artifact supports inspection of these evaluation rules.

BACKGROUND

Clinical question answering supports evidence review only when an answer can be inspected in its clinical and temporal context. A fast route that omits support, crosses a requested time boundary, suppresses a registered contradiction, or reuses stale evidence may complete a different task from a route that returns a valid answer-and-evidence packet. Treating these routes as interchangeable can turn an evidence-validity failure into an apparent performance gain.

Retrieval-augmented generation (RAG) has improved medical question answering through long-context, rationale-guided, specialty, and multi-step retrieval.[1–4] ClinicalMind demonstrates the use of large language models and knowledge graphs for real-time clinical analytics.[5] These advances motivate evaluating which completed evidence tasks should qualify for comparison when reporting the speed of a clinical information system.

Existing methods address important parts of this question. Grounding metrics and selective prediction assess executed outputs or answer/abstain decisions.[6, 7] Database provenance and temporal data management formalize lineage and time.[8–10] Question routing and ontology-guided agent memory improve the allocation of retrieval and context resources.[11, 12] EviDex makes continuously updated evidence paths auditable and freshness-aware, while FoggyCache studies approximate computation reuse.[13, 14] These mechanisms inform route design; eligibility of the complete clinical answer-and-evidence task remains a separate benchmark decision.

Study objective

We evaluated whether an explicit evidence-admission rule changes latency rankings and the evidence selected for clinical review. MedRouteGuard requires provenance, temporal validity, conflict preservation, and dependency freshness before a completed route receives benchmark credit. Its primary fixed-candidate experiment compares identical executions with admission labels enforced or ignored. Supporting studies assess label validity, recurrence in an independent public medication-safety workload, matched deployment performance, and output quality. Additional file 1 provides extended methods, results, and technical context.

METHODS

Study design

We conducted a retrospective evaluation with five prespecified evidence blocks (Table 1). Maintained-service question construction and benchmarking covered 1 January through 31 December 2024. A chronologically later public medication-safety workload tested transfer. The intended use was bounded evidence review by a qualified analyst or clinician; no model output was returned to care teams. The primary estimand was whether enforcing evidence-task membership changed the median-latency route winner and the P95 request-to-first-generated-token latency over a fixed candidate pool.

The five blocks separated distinct claims rather than treating all comparisons as evidence for one omnibus endpoint. The fixed-candidate block identified the membership effect within one execution universe. The label audit compared that deterministic membership object with a specialist standard. The independent workload tested recurrence under a later question distribution and a separately constructed public evidence graph. The matched deployment block evaluated operational performance inside a shared admission contract, while the output-quality block assessed a study-specific text-retrieval

comparator. This separation prevents performance differences between independently configured systems from being mistaken for the controlled membership effect.

Table 1
Prespecified evidence blocks, contrasts, and analysis units

Block	Population and contrast	Primary analysis unit and endpoint
Route membership	847 questions; 2,541 fixed question–route candidates; admission labels enforced versus ignored	Question; route-winner change and paired P95 difference
Label validity	320 clause-enriched routes; deterministic admission label versus blinded specialist standard	Route; sensitivity and specificity
Medication safety	326 held-out questions and 978 fixed candidates; 180 paired frozen packets	Question; winner change and paired material-error difference
Matched deployment	847 questions; MedRouteGuard versus OfflineGraph-Tuned under a shared evidence contract	Question; paired P95 difference and answered coverage
Output quality	500 questions; MedRouteGuard versus TextRAG	Question for acceptability; evidence unit for provenance

System and admission boundary

Each request was compiled into linked entities, relations, a requested time scope, and an evidence-output policy. Three route families were available: graph-first, hybrid graph-and-text, and text fallback. The maintained evidence graph stored source locations, temporal intervals, registered conflict groups, and dependency watermarks for cached or derived objects.

The locked 31 December 2024 graph 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. At ingestion, a resolvable source required a stable identifier and a document span or governed-record locator. Missing or contradictory dates remained unknown. Conflict groups linked normalized claims sharing a subject and relation but differing in polarity, value, recommendation, or population. Every cached subgraph or summary retained the transitive closure of source versions, graph-edge update sequences, and conflict identifiers on which it depended.

Admission protected the complete answer-and-evidence route with four conjunctive clauses. Provenance required resolvable support for every answer-bearing claim. Temporal validity required every used evidence object to intersect the requested time scope. Conflict preservation required every registered answer-relevant contradiction to be returned or explicitly flagged. Freshness required a complete dependency closure with no version beyond the stored watermark. Missing metadata produced an unknown-state rejection rather than a favorable default.

An answer-bearing claim was a minimal proposition asserting an entity–relation–value or polarity, recommendation, population qualifier, time statement, or quantitative result. Conjoined clauses were separated when their support, polarity, population, or time could differ. The generation record retained each claim and its evidence identifiers; a deterministic round-trip pass then matched visible assertions to retained claims. Graph evidence aligned only when normalized subject, relation, polarity or value, population, and temporal fields agreed. Any unmatched assertion, unresolved source, incomplete conflict status, or missing dependency failed the applicable clause and remained visible in the diagnostic record.

Pre-execution admission tested route capabilities and reusable state. A latency model ranked only pre-admissible candidates. Post-execution admission tested returned evidence, claim–evidence alignment, conflict coverage, and a final watermark. Formally, final admission was

$$\mathrm{Adm}_{qPr} = \mathrm{PreAdm}_{qt}\,(P) \wedge\ \mathrm{PostAdm}_{qPtr}.$$

The system buffered output until post-admission. If a route failed, it attempted the next candidate without restarting the original request clock. Request-to-first-generated-token was the historical timestamp of the first token generated by the route that ultimately passed; it was not user-visible release latency.

All answer generation used Azure OpenAI GPT-4-Turbo (gpt-4-1106-preview) through the Azure OpenAI API, with temperature 0.2, a 128K context window, and a maximum of 2,048 output tokens. The evidence-bounded prompt supplied the question, temporal scope, retrieved evidence identifiers and content, and registered conflict flags, and required claim-linked support in the returned packet. The model version and prompt bundle were locked on 31 December 2024 for the maintained-service benchmark; corresponding configurations and representative prompts are archived with the code. The admission decision was deterministic and did not use a learned probability score, so probability calibration was not applicable. The execution and fixed-candidate evaluation workflow is shown in Fig. 1.

Fixed-candidate intervention

For each of 847 clinician-authored questions, the compiler enumerated all three route families before replay, yielding 2,541 candidates. Each candidate received one excluded warm-up and five measured attempts, for 12,705 scheduled measurements. Graph snapshot, prompt bundle, generator version, Azure workers, cache state, gate computations, randomized request blocks, and concurrency were fixed. Eleven measurements (0.09%) remained missing after retry; every question retained analyzable candidates. Candidate latency was the median across successful measured replays.

The prespecified every-successful-replay rule admitted a candidate only if every successful measured replay passed all four clauses. The retrospective gate-on oracle selected the fastest admitted candidate median; the gate-off oracle selected the fastest completed candidate median while retaining gate-computation time. Thus, the intervention changed membership labels, not executions, timing instrumentation, or candidate availability. Four-of-five and majority-label rules were sensitivity analyses.

The clock began at original request acceptance. Complete-answer ended after the final generated token and claim alignment; complete-evidence ended after every evidence object and diagnostic had been validated and serialized. Evidence assembly and generation could proceed in parallel. A terminal scoped-insufficiency response was timed for deployed coverage analyses but did not receive admitted-route credit. In the fixed-candidate primary cohort, every question retained a gate-on route, so terminal no-answer latency did not enter either retrospective P95.

Clinical validation and independent medication-safety evaluation

The held-out label audit contained 320 candidate routes, 80 enriched for each admission clause. Specialists blinded to the system label judged whether the returned packet was clinically valid for the specified clause and available metadata. The reference class for sensitivity was a clinically invalid route that should be rejected.

Sampling was stratified by clause, system decision, and severity to estimate label validity rather than natural prevalence. Two specialists independently reviewed each route. Disagreement on clause validity or material severity triggered a third adjudication before unblinding. Sensitivity, specificity, predictive values, false-admission rate, and false-rejection rate were calculated from the adjudicated route labels.

The independent workload used time-versioned FDA labeling, ClinicalTrials.gov records, and open-access biomedical literature.[15–17] A separate team screened 451 questions, excluded 29 duplicate, non-time-indexed, or non-adjudicable questions, used 96 earlier questions for development, and held out 326 questions after 1 July 2025. Three route families yielded 978 fixed candidates, evaluated with the same replay and gate-on/gate-off design.

A prespecified stratified sample of 180 held-out questions entered a blinded two-period crossover review. Each condition used the successful frozen packet nearest the selected route's median latency; identical selected routes reused the same frozen output. Ten medication-safety clinicians each reviewed 36 pairs in counterbalanced order after a one-week washout. Every output received two primary ratings; disagreements on material error, unsafe omission, acceptability, or severity triggered adjudication by a third clinician. A material evidence-validity error was an out-of-window, conflict-suppressing, or unsupported evidence defect capable of changing the review conclusion, urgency, contraindication assessment, or dose decision.

The 180 questions comprised 41 drug-interaction, 38 contraindication, 34 dose-adjustment, 32 warning-temporal, and 35 evidence-conflict questions. Reviewers saw the question, frozen answer, and evidence packet but not the route, gate label, latency, or condition. Applicability of contradiction visibility and temporal-scope endpoints was fixed independently from the indexed reference standard. Missing rating fields could not be submitted; no primary rating was missing.

Supporting deployed and output-quality evaluations

The matched deployed comparison used the same 847 questions, evidence graph, admission contract, Azure worker classes, Redis cache budget, and endpoint protocol. MedRouteGuard selected among the three pre-admissible families; OfflineGraph-Tuned used a fixed graph-first serving policy with tuned Neo4j hints and ColBERT-based linking. A blinded 240-question audit checked whether latency differences were compatible with output quality.

The deployment used five Azure Standard E32ads v5 Neo4j workers, two Standard E16ads v5 query-service workers, one Standard NC24ads A100 v4 vector-index worker, and a 64-GiB Redis cache. Timed endpoints included request-to-first-generated-token, complete answer, and complete evidence. Coverage retained all 847 questions; failures remained in request-level failure reporting even when another replay for the question succeeded.

For the separate output-quality block, five board-certified specialists reviewed 500 unique questions and one output from MedRouteGuard and TextRAG, presented as 1,000 blinded single-answer packets. Three independent ratings per output assessed correctness, completeness, safety, and actionability. Overall acceptability was judged separately by at least two of the three reviewers, based on task fulfilment, supported key conclusions, and absence of consequential error or serious safety defects; safety disagreements underwent independent adjudication. Provenance precision and recall used returned evidence units. A separate 200-question audit classified citation-error mechanisms.

Statistical analysis and ethics

P50, P95, and P99 latency used question-level candidate medians. Confidence intervals used 1,000 paired bootstrap resamples of question identifiers with measured replays resampled within question. Acceptability used paired risk differences and exact McNemar tests. Label-validity intervals used Wilson score intervals. Medication-safety binary outcomes used paired risk differences and two-sided exact tests on discordant adjudicated pairs; a common-packet sensitivity forced identical frozen packets to share one outcome. All intervals were two-sided 95% intervals. The five prespecified evidence blocks addressed distinct questions; no joint rejection claim was made. Detailed estimands, missingness rules, power rationale, and reviewer-aware models are in Additional file 1.

The packet review additionally reported conditional effects among the 14 questions for which the two retrospective oracles selected different routes and outputs. Those intervals used paired resampling of the active questions and were interpreted as conditional, not population-prevalence estimates. Period and randomized sequence were evaluated jointly in a secondary mixed model; a separate carryover-proxy sensitivity omitted sequence because the two are not jointly identifiable in a two-period crossover. Unsafe omission and all other secondary endpoints were descriptive and were not included in a multiplicity-controlled rejection claim.

The institutional Clinical Informatics Research Ethics Committee approved the retrospective evaluation and waived individual consent because data were de-identified before research access, there was no

patient contact, and outputs were not returned to care teams. The study followed the Declaration of Helsinki, CHART guidance, and applicable retrospective STROBE items;[18, 19] the completed STROBE checklist accompanies the submission.

Public companion artifact

The versioned MedRouteGuard artifact is publicly archived in Zenodo under the MIT License.[20] It provides source code, configuration files, tests, pinned public-source samples, synthetic fixtures, and an offline demonstration that requires no paid model API. Deterministic simulation illustrates admission, replay aggregation, fallback accounting, and reporting; the packaged paper-reference aggregates identify the reported study results separately. This lightweight workflow supports inspection of the evaluation rules without requiring the governed clinical records or a full commercial-model rerun. Data and code access are detailed below.

RESULTS

Admission changed fixed-candidate latency rankings

Enforcing the every-successful-replay rule changed the median-latency route winner for 61/847 questions (7.2%; 95% CI, 5.7%–9.0%). The retrospective-oracle P95 increased from 2.89 s with admission labels ignored to 3.24 s with labels enforced, a paired shift of + 0.35 s (95% CI, + 0.18 to + 0.53). Excluded-candidate speed events occurred for 98/847 questions (11.6%), and candidate-latency bucket membership changed for 83/847 (9.8%). Every question retained at least one admitted candidate, so the two P95 estimates used the same question denominator and candidate-median endpoint.

The direction persisted when admission required four of five passing replay labels (+ 0.32 s; 56 winner changes) or a majority (+ 0.28 s; 49 winner changes). Relaxing provenance, temporal scope, conflict preservation, or freshness alone changed 34, 31, 16, and 24 winners, respectively. Detailed overlap, question-family, public-only, and cache/update analyses are reported in Additional file 1.

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 median winner gap was 0.27 s. Associated failed-clause tags among the 98 speed-event questions were provenance 44, temporal 31, conflict 19, and freshness 28; these counts were non-exclusive because one excluded candidate could fail more than one obligation.

Admission labels agreed with specialists

In the 320-route audit, pooled sensitivity for rejecting clinically invalid routes was 92.3% (95% CI, 86.7%–95.6%) and specificity for admitting valid routes was 93.3% (88.6%–96.1%). The pooled confusion counts were 131 true rejections, 12 false rejections, 11 false admissions, and 166 true admissions. Forty-one

routes required third-reviewer adjudication; agreement before adjudication was $\kappa = 0.82$ (0.77–0.87). Conflict preservation had the lowest clause-specific sensitivity (87.9%) and widest uncertainty.

Pooled positive and negative predictive values were 91.6% (85.9%–95.1%) and 93.8% (89.2%–96.5%). 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. Because the sample was enriched by clause and severity, predictive values are conditional on this audit and should not be interpreted as deployment prevalence.

The independent workload reproduced ranking and packet effects

In 326 held-out medication-safety questions, admission changed 21 winners (6.4%) and shifted P95 from 3.43 to 3.88 s, a paired difference of + 0.45 s (+ 0.23 to + 0.68). No question had a zero-admissible set. In the 180-question packet review, material evidence-validity errors were 7/180 with gate-on selection and 18/180 with gate-off selection, a paired difference of -6.1 percentage points (-10.4 to -1.9). Unsafe omissions were 4/180 versus 13/180.

The two selection rules produced different routes and frozen outputs for 14/180 questions. Within these active pairs, material errors were 2/14 versus 13/14 and unsafe omissions were 1/14 versus 10/14. The corresponding absolute reductions were 78.6 percentage points (42.9–100.0) and 64.3 points (28.6–92.9). The remaining 166 pairs reused identical frozen packets; condition-only rating disagreements were balanced, and forcing one common label for each identical packet left both full-sample risk differences unchanged.

Exact paired tests within the active pairs yielded 1 gate-on-only versus 12 gate-off-only material-error discordances ($P = 0.0034$) and 1 versus 10 unsafe-omission discordances ($P = 0.0117$). Contradiction visibility when required was 45/49 versus 35/49, and temporal-scope validity was 58/61 versus 48/61. Median complete-evidence review time was 4.7 versus 5.6 minutes. Reviewer agreement was $\kappa = 0.79$ (0.73–0.85), and period, order, and carryover diagnostics did not show material imbalance. Table 2 summarizes the primary and supporting outcomes across the five evidence blocks. The cross-workload evidence chain and the output-discordant clinical outcomes are shown in Fig. 2.

Table 2
Primary and supporting outcomes across the five evidence blocks

Evidence block	Primary estimate	Supporting result
Fixed-candidate membership	61/847 winner changes; P95 + 0.35 s [+ 0.18, + 0.53]	98/847 excluded-candidate speed events
Label validity	Sensitivity 92.3% [86.7, 95.6]	Specificity 93.3% [88.6, 96.1]
Medication-safety replication	21/326 winner changes; P95 + 0.45 s [+ 0.23, + 0.68]	Material errors 7/180 versus 18/180
Matched deployment	First-token P95 3.24 versus 4.18 s	Answered coverage 98.9% versus 98.5%
Output quality	Acceptable 94.2% versus 87.4%	Evidence F1 0.87 versus 0.74

Supporting deployment and quality results

Under the shared evidence contract, MedRouteGuard reduced matched deployed request-to-first-generated-token P95 from 4.18 to 3.24 s, a paired difference of -0.94 s (-1.19 to -0.69). Complete-answer P95 was 6.82 versus 8.06 s, and answered coverage was 838/847 (98.9%) versus 834/847 (98.5%). These timestamps describe recorded processing milestones; user-visible release occurred only after answer and evidence validation.

In the separate 500-question audit, acceptable outputs were 471/500 (94.2%) for MedRouteGuard and 437/500 (87.4%) for TextRAG, a paired difference of + 6.8 percentage points (+ 4.2 to + 9.5). Evidence precision, recall, and F1 were 0.91, 0.84, and 0.87 versus 0.78, 0.71, and 0.74. The matched 240-question deployed-system audit showed 225 versus 220 acceptable outputs, a paired difference of + 2.1 points (-1.0 to + 5.4). Full distributions, provenance counts, failure taxonomy, coverage, and throughput appear in Additional file 1.

DISCUSSION

This study identifies a benchmark-denominator problem: route execution and route eligibility are different properties. When the clinical evidence contract controlled membership, 7.2% of maintained-service questions and 6.4% of independent medication-safety questions changed latency winners. Because the primary intervention reused identical executions and changed only prespecified labels, the ranking effect cannot be attributed to different prompts, hardware, candidate generators, or runtime conditions.

The specialist audit connects the deterministic boundary to clinically interpretable evidence obligations. The independent packet review further shows why membership matters beyond a numerical latency shift. The overall paired error difference was modest in absolute terms because only 14 of 180 selections produced different frozen outputs, but the error contrast was concentrated within those active pairs. This localization, together with balanced disagreements on identical packets, argues against interpreting

repeated-rating variation as a system-output effect. The analysis remains retrospective and supports an association between selection rules and reviewed packets, not a patient-outcome or online-treatment effect.

The contribution complements clinical analytics platforms such as ClinicalMind by testing the eligibility of completed evidence tasks within a fixed execution universe.[5] It also differs from contract-checked graph editing for verifier-guided reasoning, where a structural gate establishes whether a reasoning graph can run before semantic verification.[21] Here, the checked object is a clinical answer-and-evidence route, and its admission label determines benchmark membership. MedRouteGuard differs from routing objectives that score candidate accuracy, cost, or latency and from post-hoc grounding scores that characterize an output after execution. Here, capability and state constrain routes before execution, returned evidence determines final admission, and the resulting label controls benchmark credit while the original request clock retains failed work. The deployed comparison indicates that this evidence boundary is compatible with high coverage and lower recorded tail-latency milestones, but it does not establish user-visible release latency or isolate every implementation difference between the deployed configurations.

Several limitations bound interpretation. The maintained-service and medication-safety cohorts were purposefully stratified rather than prevalence samples. The every-successful-replay rule estimates conservative consistency; vote-based rules answer different reliability questions. The four clauses evaluate available provenance, time, conflict, and dependency metadata, not the truth of every source. The packet effect was concentrated in 14 questions and therefore has wide conditional intervals. The deployed comparison includes policy, query-hint, and linking differences despite a shared graph and evidence contract. Finally, the evaluated generator, Azure configuration, and three route families may not represent other institutions, models, or workloads.

For clinical information-system evaluation, latency reports should state the completed task, the eligibility rule, the handling of failed attempts and fallback, and the coverage denominator. The present evidence suggests that these are not reporting details: changing membership can change both the tail estimate and the evidence packet selected for review. Prospective multi-site studies should test the boundary under other models, evidence stores, and user-visible release endpoints.

CONCLUSIONS

An explicit admission rule for complete clinical answer-and-evidence routes changed latency rankings on fixed executions, agreed with specialist validity judgments, and altered medication-safety evidence packets. Benchmarking speed without specifying route eligibility can therefore credit evidence omission, temporal mismatch, conflict suppression, or stale reuse as performance.

Abbreviations

API

application programming interface
CI
confidence interval
F1
harmonic mean of precision and recall
GPU
graphics processing unit
LLM
large language model
P95
95th percentile
QA
question answering
RAG
retrieval-augmented generation
SLO
service-level objective
TTFT
time to first token.

Declarations

Ethics approval and consent to participate

The institutional Clinical Informatics Research Ethics Committee approved this retrospective study and waived individual informed consent. The study used de-identified data and followed the Declaration of Helsinki.

Clinical trial number: not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The public data used for the independent medication-safety evaluation are available from openFDA, ClinicalTrials.gov, and Europe PMC. Pinned snapshots of these public sources, synthetic fixtures, and deterministic demonstration outputs accompany the lightweight artifact archived in Zenodo at https://doi.org/10.5281/zenodo.22681885. Study aggregates are reported in the article and Additional file 1. Maintained-service per-question route ledgers, replay-level annotations, evidence packets, and reviewer-pairing records are subject to institutional data-use and privacy controls and are not publicly

available. Inquiries about restricted materials should be directed to the corresponding author; access remains subject to institutional authorization and applicable data-use and privacy controls.

The MedRouteGuard companion source code, configuration files, offline demonstration workflows, tests, and documentation are available in Zenodo at https://doi.org/10.5281/zenodo.22681885 under the MIT License.

Project: MedRouteGuard. Archived code and project documentation: https://doi.org/10.5281/zenodo.22681885. Programming language: Python. Requirements: Python 3.14 and the dependencies listed in requirements.txt. The documented offline workflow was tested on macOS. License: MIT; commercial and non-academic use is permitted under the license.

Competing interests

The authors declare no financial or non-financial competing interests.

Funding

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

Authors' contributions

RL, JZ, SC, AD, and RHL contributed to conceptualization, methodology, software, validation, formal analysis, data curation, clinical review, visualization, writing–original draft, and writing–review and editing. RL supervised the work and led project administration. All authors accept accountability for their contributions and the integrity of the work. RL, Rui Li; JZ, Jason Zhao; SC, Shuang Cao; AD, Alexandre Duprey; RHL, Ruihua Liu.

Acknowledgements

OpenAI Codex was used to improve the writing and presentation of this manuscript. The authors remain responsible for its final content. Azure OpenAI GPT-4-Turbo was the evaluated answer generator; its version, settings, and human validation procedures are described in Methods and Additional file 1.

References

1. Zhang G, Xu Z, Jin Q, et al. Leveraging long context in retrieval augmented language models for medical question answering. NPJ Digit Med. 2025;8:239. https://doi.org/10.1038/s41746-025-01651-w.
2. Sohn J, Park Y, Yoon C 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. pp. 12739–53. https://doi.org/10.18653/v1/2025.naacl-long.635

3. Wada A, Tanaka Y, Nishizawa M, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Digit Med. 2025;8:395. https://doi.org/10.1038/s41746-025-01802-z.
4. Wind S, Sopa J, Truhn D, et al. Multi-step retrieval and reasoning improves radiology question answering with large language models. NPJ Digit Med. 2025;8:790. https://doi.org/10.1038/s41746-025-02250-5.
5. Cao S, Li R, Wu R, et al. Real-time clinical analytics at scale: A platform built on large language models-powered knowledge graphs. JAMIA Open. 2026;9(1):ooaf167. https://doi.org/10.1093/jamiaopen/ooaf167.
6. Es S, James J, Espinosa-Anke L et al. 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. pp. 150–8. https://doi.org/10.18653/v1/2024.eacl-demo.16
7. 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. pp. 2151–9. (Proceedings of machine learning research; vol. 97). Available from: https://proceedings.mlr.press/v97/geifman19a.html
8. Buneman P, Khanna S, Tan WC. Why and where: A characterization of data provenance. In: International conference on database theory [Internet]. 2001. pp. 316–30. https://doi.org/10.1007/3-540-44503-X_20
9. Cheney J, Chiticariu L, Tan WC. Provenance in databases: Why, how, and where. Found Trends Databases [Internet]. 2009;1(4):379–474. https://doi.org/10.1561/1900000006.
10. Jensen CS, Snodgrass RT. Temporal data management. IEEE Trans Knowl Data Eng [Internet]. 1999;11(1):36–44. https://doi.org/10.1109/69.755613
11. 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. pp. 3621–34. https://doi.org/10.18653/v1/2025.findings-naacl.200
12. Cao S, Li R. Ontology-guided long-term agent memory for conversational RAG. In: Proceedings of machine learning and systems 8 [Internet]. 2026. pp. 233–50. Available from: https://proceedings.mlsys.org/paper_files/paper/2026/hash/2fb4be70fc9668e9ec2c71b34fb127d4-Abstract-Conference.html
13. Li R, Cao S, Liu R, et al. EviDex: Provenance-weighted evidence-path indexing for fresh and auditable retrieval under continuous updates. Proc ACM Meas Anal Comput Syst. 2026;10(2):1–23. https://doi.org/10.1145/3805636.

14. Guo P, Hu B, Li R et al. FoggyCache: Cross-device approximate computation reuse. In: Proceedings of the 24th annual international conference on mobile computing and networking. ACM; 2018. pp. 19–34. https://doi.org/10.1145/3241539.3241557
15. U.S. Food and Drug Administration. FDALabel: Full-text search of drug product labeling [Internet]. 2024. Available from: https://nctr-crs.fda.gov/fdalabel/ui/search
16. U.S. National Library of Medicine. ClinicalTrials.gov [Internet]. 2026. Available from: https://clinicaltrials.gov/
17. National Library of Medicine. PubMed central open access subset [Internet]. 2024. Available from: https://pmc.ncbi.nlm.nih.gov/tools/openftlist/
18. Huo B, Collins G, Chartash D, et al. Reporting guideline for chatbot health advice studies: The CHART statement. BMC Med. 2025;23:447. https://doi.org/10.1186/s12916-025-04274-w.
19. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. https://doi.org/10.1016/S0140-6736(07)61602-X.
20. Li R, Cao S. MedRouteGuard: companion artifact for Every-successful-replay route admission changes first-token latency rankings in clinical question answering [Internet]. Zenodo; 2026. https://doi.org/10.5281/zenodo.22681885
21. Li R, Cao S. From trajectories to graphs: Contract-checked editing for verifier-guided LLM reasoning. In: Proceedings of the 64th annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics; 2026. pp. 43259–306. https://doi.org/10.18653/v1/2026.acl-long.2004