



Evaluating an On-Premises Retrieval-Augmented Generation System for Health Records Retrieval: Retrieval Performance, Response Latency and Perceived Usability in Kenyan Level 5 Hospitals

Kevin Comba Gatimu^{1*}, James Mwikya², Stanley Mogaka²

¹Department of Computer Science and Informatics, Kirinyaga University, Kerugoya, Kenya

²School of Pure and Applied Sciences, Kirinyaga University, Kerugoya, Kenya

*Corresponding author: kevin.comba.gatimu@outlook.com | Co-authors: jmwikya@kyu.ac.ke; smogaka@kyu.ac.ke

DOI : <https://doi.org/10.55248/gengpi.07.1026.3266>

ABSTRACT :

Retrieval-Augmented Generation (RAG) systems are increasingly proposed for clinical information retrieval, yet evaluations are overwhelmingly conducted on cloud infrastructure in high-income settings, reported gains from advanced retrieval techniques are seldom subjected to ablation, and automated judges are routinely drawn from the same model as the system under test. This article reports the third objective of a design science study: evaluation of the retrieval performance, response latency and perceived usability of a developed On-Premises RAG system for health records retrieval in Level 5 hospitals in Kenya. Twelve curated queries spanning six classes, each paired with a ground-truth reference computed by structured query over a synthetic corpus of 1,603 indexed chunks, were executed through the live pipeline under three configurations, being the full advanced pipeline, a naive vector-search baseline and a no-rerank ablation, with retrieval depth held at eight chunks throughout. Two deterministic, judge-free measures were computed for all three configurations, and three judged measures were computed for the naive configuration by Qwen2.5-7B-Instruct, a judge independent of the Phi-4 Mini Instruct (3.8B) generator in both instance and model family and served locally, so that no health-record content left the machine. Contrary to expectation, the naive baseline outperformed both advanced configurations on key-fact context precision, at 0.687 against 0.633 and 0.591 over the nine queries for which the measure is defined, and on key-fact accuracy, at 0.491 against 0.400 and 0.275 over all twelve. It was also the only configuration to meet the thirty-second design target, at a mean of 13.99 seconds against 34.33 seconds for the no-rerank ablation. Judged context recall of 0.533 over ten defined scores and faithfulness of 0.599 over eleven, the latter below the criterion of 0.80 adopted by the study, together indicate that incomplete retrieval rather than unsupported generation was the binding constraint. Performance varied sharply by class: catalogue lookup recorded key-fact accuracy of 0.875 while aggregate counting recorded 0.167 and identifier-keyed per-patient retrieval 0.244, the last with judged context recall of 0.000. Thirteen intended users at the two study hospitals recorded completion of between 84.6 and 100.0 per cent of six task scenarios, rated source verifiability highest at a median of 5, and returned a mean System Usability Scale total of 80.6; every one of them identified per-patient retrieval as routine, which is the class on which the system performed worst. The study concludes that added pipeline complexity conferred no measured benefit at this corpus scale, and that aggregate and identifier-keyed questions require structured and exact-match retrieval alongside semantic search.

Keywords: retrieval-augmented generation; RAGAS; ablation analysis; response latency; System Usability Scale; on-premises deployment; health records retrieval; Kenya

1. Introduction

1.1 Background of the Study

As Retrieval-Augmented Generation has moved from research construct to clinical proposal, the question of how such systems should be assessed has become as consequential as how they should be built. RAG, formalised by Lewis et al. (2020), grounds generated text in evidence retrieved at inference time, and clinical studies report potential improvements in grounding, although source support remains incomplete in some evaluations (Amugongo et al., 2025). Because a RAG system comprises a retrieval component and a generation component whose failures are distinct, evaluation requires instruments that assess each separately as well as end to end. The RAG Assessment (RAGAS) framework of Es et al. (2024) has become the de facto standard for this purpose, decomposing system behaviour into faithfulness, context precision and context recall and employing a language model as judge so that evaluation scales without exhaustive human annotation. A system intended for clinical use must additionally satisfy an operational constraint and be found usable by the people who would operate it, so retrieval quality, response latency and perceived usability are treated here as three facets of a single evaluation rather than as separate studies.

Evaluation practice in the field nonetheless carries four systematic weaknesses. It is geographically concentrated in high-income settings; it is architecturally optimistic, since advanced techniques such as reranking and contextual compression are frequently adopted on the strength of aggregate benchmark gains rather than setting-specific ablation; it is often infrastructurally incompatible with the systems most in need of assessment, since cloud-hosted judge models cannot lawfully or practically be used where the data under evaluation must not leave the premises; and it frequently draws the judge from the same model as the generator, a configuration that Zheng et al. (2023) identify as subject to self-enhancement bias, that Wataoka, Takahashi and Ri (2024) trace to familiarity rather than quality, and that Spiliopoulou et al. (2025) show extends to model family as well as instance. Broader benchmarks have exposed how far RAG systems remain from expert performance on temporal, set-based and multi-hop questions (Yang et al., 2024), and evaluation surveys now recommend multi-dimensional suites for clinical deployment (Zhang et al., 2025), yet neither addresses evaluation under data-residency constraint.

In Kenya, this last point is binding rather than theoretical. The Data Protection Act, 2019 classifies health information as sensitive personal data under Section 44, and Section 50 of the Act, read with Regulation 26 of the Data Protection (General) Regulations, 2021, designates the provision of primary or secondary health care a strategic interest of the state requiring that such processing be carried out through a server and data centre located in Kenya. A system built to keep patient data on the premises cannot be evaluated by transmitting that data to an external judge model, since doing so would reproduce at evaluation time precisely the exposure the architecture was designed to eliminate. Evaluation must therefore itself be conducted on-premises. Companion studies established the design requirements and architecture for such a system and developed it in full; the present article reports how well it performs, how quickly it responds, and how the staff who would use it judged it.

1.2 Statement of the Problem

The ideal situation is one in which a hospital can determine, before clinical deployment and using only local resources, how reliably a retrieval system grounds its answers, how accurately it answers, how long it takes to answer, and whether the staff who would use it find it usable, each disaggregated by the kinds of question clinicians actually ask. The current reality is that published clinical RAG evaluations report aggregate scores on general-domain or high-income clinical datasets, rarely ablate the pipeline stages whose contribution they implicitly assume, seldom report latency alongside quality, and are executed on infrastructure that resource-constrained and residency-bound facilities cannot use.

The knowledge gap has three parts. First, the performance of a fully on-premises RAG pipeline on health-records retrieval in an African county referral hospital context has not been measured, so the practical cost of the privacy-preserving architectural choice is unknown. Second, the marginal contribution of advanced retrieval techniques at the corpus scale and query mix of such a facility has not been established: the literature reports their benefit at benchmark scale and infers that it transfers, but the inference is untested where corpora are small and queries are heterogeneous. Third, automated retrieval scores and the judgements of intended users have not been obtained for the same system in the same setting, so it is not known whether the tasks a benchmark measures are the tasks staff actually perform.

If this gap persists, facilities risk three errors. They may adopt elaborate pipelines whose stages confer no benefit and may actively discard relevant evidence, incurring latency and complexity for nothing; observing high faithfulness scores, they may conclude that a system is accurate when it is merely well grounded in whatever it happened to retrieve; or, observing favourable usability ratings, they may treat perceived usability as evidence of retrieval correctness. This study addresses the gap by evaluating the developed system under three pipeline configurations, disaggregated by query class, entirely on local infrastructure, and by placing the technical results beside the judgements of thirteen intended users.

1.3 Objective of the Study

To evaluate the retrieval performance, response latency and perceived usability of the developed On-Premises Retrieval-Augmented Generation system through technical benchmarking and assessment by intended users.

The corresponding research question is: what retrieval performance, response latency and perceived usability does the developed On-Premises RAG system exhibit under technical benchmarking and assessment by intended users?

1.4 Justification and Scope

A quantitative evaluation using an established framework, complemented by structured assessment from intended users, is warranted on four grounds. The construct of retrieval quality is decomposable and measurable, and RAGAS supplies validated instruments for each component whose scores correlate strongly with human judgement (Es et al., 2024). Comparative configuration analysis requires a scoring procedure that can be applied identically across conditions, which automated and deterministic scoring permit and expert appraisal does not at feasible cost. The residency constraint requires that evaluation execute locally, which a locally hosted judge model makes possible. And because automated scores say nothing about whether

the benchmark measures the work that is actually done, an assessment by intended users is necessary to establish which classes of question arise in practice.

The evaluation is delimited to retrieval quality, end-to-end response latency and perceived usability and acceptance. Clinical outcomes, patient safety and task efficiency in live workflow were not assessed, and no observed task times were available from the expert session, so no estimate of time saved is offered. It was conducted on a synthetic corpus modelled on Kenyan hospital record structures, no actual patient records being used at any stage, and in a controlled environment configured to match the server specifications of the two study hospitals rather than in live clinical workflow. The technical benchmark comprises twelve scored queries drawn from a master inventory of sixteen, two in each of six classes, and this modest size is acknowledged throughout as constraining the strength of inference; no inferential significance testing is reported. The scope does not extend to a clinical impact assessment, a longitudinal deployment study or a randomised controlled trial.

2. Literature Review and Conceptual Framework

2.1 Theoretical Framework

The study is anchored in the evaluation activity of the Design Science Research Methodology of Peffers et al. (2007), the fifth of its six activities, in which the artefact is assessed against the objectives defined for it. Design science holds that evaluation is not an appendix to construction but the mechanism by which an artefact yields knowledge: what is learned is the relationship between design choices and observed performance under stated conditions. This orientation carries a methodological consequence adopted here, namely that evaluation must be able to disconfirm design decisions. An evaluation reporting only aggregate scores for a single configuration cannot do so, whereas an ablation design that varies pipeline composition while holding corpus, queries, generation model, and judge model constant can attribute score differences to the stages under test.

The measurement framework is supplied by RAGAS (Es et al., 2024), which decomposes RAG behaviour into components that fail independently, and is supplemented here by two deterministic, judge-free measures and by direct instrumentation of the running application. Faithfulness measures the proportion of statements in a generated response supported by the retrieved context, and so isolates hallucination resistance without reference to ground truth. Context precision measures the proportion of retrieved chunks relevant to the reference answer, and context recall the extent to which retrieved context covers the claims in that reference; together they characterise retrieval quality from the complementary directions of noise and coverage. The two deterministic measures, key-fact context precision and key-fact accuracy, apply the same average-precision and coverage logic by lexical matching of salient reference facts, and require no model call: they carry no judge variance and are exactly reproducible from the archived captures, which is what allows them to carry a cross-configuration comparison. They are lexical proxies and may under-credit a correct paraphrase or over-credit an answer that echoes the tokens of the reference. The decomposition matters because the metrics can move independently: a system may be highly faithful to poorly retrieved context, producing answers that are well grounded and wrong. Distinguishing the two is the analytical purpose the framework serves.

2.2 Review of Related Empirical Studies

Es et al. (2024) developed and validated RAGAS by comparing automated judgements against human evaluation across diverse question-answering tasks, finding strong correlation and establishing the framework as a scalable alternative to manual assessment. The validation was conducted on general-domain datasets, and applicability to health-records retrieval in African settings, where clinical terminology, documentation conventions and query patterns differ from the validation data, was not established. Yang et al. (2024) introduced the Comprehensive RAG Benchmark across five question types and 2,655 questions, revealing substantial gaps between RAG systems and human experts on temporal, set-based and multi-hop questions; set-based questions are the benchmark analogue of the aggregate counting queries examined here. Xiong et al. (2024) provided the MIRAGE benchmark for zero-shot medical question answering, and Zhang et al. (2025) synthesised the evolution of RAG evaluation, recommending a minimum suite of faithfulness, answer relevance, context precision and end-to-end accuracy for clinical deployments; the suite adopted here substitutes reference-based context recall and deterministic key-fact accuracy for answer relevance, once ground-truth references had been curated for every query. A separate literature governs the choice of judge. Zheng et al. (2023) identify self-enhancement bias as a principal limitation of the LLM-as-a-judge paradigm; Wataoka, Takahashi and Ri (2024) show that models assign systematically higher evaluations to text they find more familiar, so that self-preference expresses familiarity rather than quality; Spiliopoulou et al. (2025) show the bias extends to model family; and Koo et al. (2024) show that models of smaller parameter scale exhibit stronger cognitive biases as evaluators and correlate less well with human preference ordering. Perceived usability is measured with the System Usability Scale of Brooke (1996), whose totals are interpreted against the benchmark distribution reported by Bangor et al. (2008).

Evidence on the contribution of advanced techniques is more equivocal than its frequent citation suggests. Wolk (2025) compared twelve RAG

pipelines: hallucination rates were 11.2 per cent for hybrid fusion, 18.4 for dense retrieval and 20.1 for sparse retrieval; self-reflective RAG was lowest at 5.8 per cent. Oche et al. (2025) identified robustness limitations in text-based RAG under long-context and multi-hop reasoning and recommended hybrid retrieval, while Wampler et al. (2025) established chunk size, embedding model choice and retrieval depth as critical performance determinants, a finding whose corollary, that performance is highly sensitive to configuration, implies that transfer of benchmark-derived defaults to a different corpus and query mix cannot be assumed. Gao et al. (2024) reinforce this caution, observing that hallucination persists in retrieval-heavy tasks across all three architectural generations. Park et al. (2025) demonstrated improved retrieval precision through structured clinical query formulation, though under continuous cloud connectivity. Korom et al. (2025) evaluated AI Consult, a safety-net clinical decision support tool, in a quality-improvement study covering 39,849 patient visits at Penda Health clinics in Nairobi. Clinicians with access to the tool made 16 per cent fewer diagnostic errors and 13 per cent fewer treatment errors than clinicians without access. These clinical-error outcomes are distinct from retrieval-quality metrics.

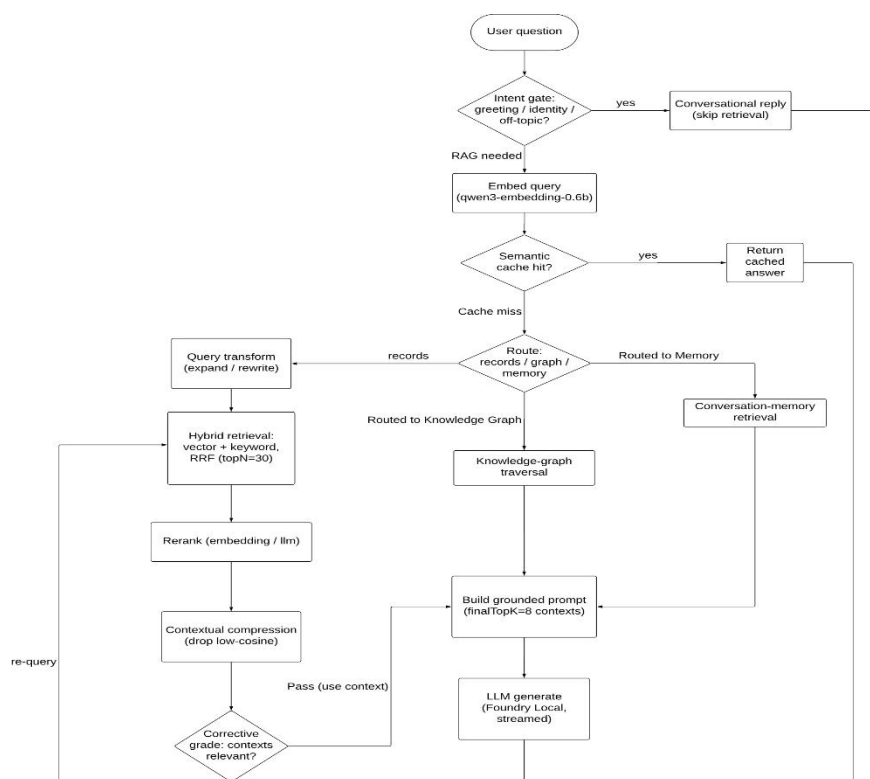
2.3 Identification of Research Gaps

Four gaps follow. First, RAGAS has not been applied to health-records retrieval in an African healthcare setting; Es et al. (2024) validated it on general-domain data, and the benchmarks of Yang et al. (2024) and Xiong et al. (2024) address general and medical question answering rather than institutional record retrieval, leaving the framework's behaviour on this task in this context unestablished. Second, the marginal contribution of advanced retrieval techniques at small corpus scale is untested: Wołk (2025), Oche et al. (2025) and Wampler et al. (2025) motivate configuration-specific evaluation, but their findings do not determine which stages benefit the corpus and hardware configuration used here, so facilities have no evidence on which to decide whether to enable them. Third, no published evaluation has been executed wholly on-premises with a locally hosted judge drawn from a different model family than the generator, a design that the residency obligations governing Kenyan health data make necessary and that the self-preference literature makes methodologically required. Fourth, automated retrieval scores and structured judgements by intended users have not been reported for the same system, so the relationship between what a benchmark measures and what staff actually do remains unexamined. This article addresses all four.

2.4 Conceptual Framework

The evaluation framework treats the pipeline configuration and the query class as the independent variables and the retrieval measures as the dependent variables, with the corpus, query set, generation model and judge model held constant so that observed differences are attributable to configuration. Figure 1 shows the query pipeline whose stages constitute the manipulated variable: the three configurations correspond to all stages enabled, all advanced stages disabled, and all stages enabled except reranking. Retrieval depth is held at eight context chunks in every configuration, so differences are not attributable to how much was retrieved. The framework distinguishes two dependent constructs that the literature often conflates, namely grounding, captured by faithfulness, and accuracy, captured by key-fact accuracy and underwritten by context precision and recall; their separation is the analytical core of the study, since a system may score highly on the first while failing on the second. Two further dependent constructs complete the objective: end-to-end response latency, assessed against a thirty-second design criterion, and perceived usability and acceptance, assessed by intended users.

Figure 1: Advanced RAG Query Pipeline Showing the Stages Manipulated Across Configurations



3. Methodology

3.1 Research Design

The study employed a quantitative, comparative evaluation design corresponding to the evaluation activity of the Design Science Research Methodology (Peppers et al., 2007), with a structured assessment by intended users as a second strand. Three pipeline configurations were compared on a common query set over a common corpus, with the generation model and the judge model held constant across conditions and retrieval depth fixed at eight chunks, so that score differences were attributable to retrieval configuration alone. The design is an ablation in the strict sense: the full configuration and the no-rerank configuration differ in exactly one stage, permitting the contribution of reranking to be isolated, while the naive baseline establishes the floor against which the advanced pipeline as a whole is compared. The second strand is a single-session assessment by thirteen intended users at the two study hospitals, working with the naive configuration over synthetic task scenarios. That strand did not compare pipeline configurations under controlled conditions and is not presented as such a comparison.

3.2 Evaluation Dataset and Corpus

The corpus comprised synthetic health records representing the electronic medical record database of a fictitious institution, generated programmatically to reproduce the structure, field coverage and referential integrity of Kenyan hospital record data and ingested into the system as 1,603 chunks of approximately 512 tokens with source lineage metadata. It comprises 60 synthetic patients and 120 clinical encounters across 25 referentially complete relational tables, is exactly reproducible from seed 20260816, and its integrity was verified before ingestion. No actual patient records were used at any stage, consistent with Section 44 of the Data Protection Act, 2019, and a Data Protection Impact Assessment under Section 31 of the Act and Regulations 49 to 53 of the 2021 Regulations was completed before the system was operated.

The master query inventory comprises sixteen health-records retrieval queries, of which twelve constitute the captured technical benchmark, two drawn from each of six classes. Each is paired with a ground-truth reference answer computed directly from the seeded corpus by structured query against the source database rather than authored by hand, which is what makes the references verifiable and the benchmark reproducible; because the references are data-dependent, they must be regenerated if the corpus is re-seeded, and they were verified against the saved seed before the results reported here were compiled. The six classes were chosen to span the retrieval scenarios encountered in health-records practice rather than sampled at random, so that performance could be attributed to query structure. Aggregate-count queries ask whole-database questions such as the total number of registered patients. Distribution queries ask for a grouped profile, such as the most common diagnoses. Catalogue-lookup queries ask for enumerable reference knowledge, such as the medications indicated for hypertension. Per-patient queries are keyed to an individual patient number and require exact identifier matching. Cross-condition queries ask about co-occurrence between diagnoses. Negative-case queries ask about facts the corpus does not contain, so that a correct answer is a correct report of absence. References for the aggregate, distribution and cross-condition classes state the computed value and name the records contributing to it, rather than stating the value alone; the reason is given in Section 5.1.

3.3 Variables and Measurement

The independent variables were pipeline configuration, at three levels, and query class, at six levels. The dependent variables comprised five retrieval measures, each bounded between 0 and 1 and defined in Table 1, together with end-to-end response latency in seconds and the perceived usability and acceptance ratings returned by intended users. Two families of retrieval measure were used and are never pooled. The deterministic measures, key-fact accuracy and key-fact context precision, are computed without a language model by matching salient reference facts against the generated answer and against the retrieved chunks respectively; they were computed for all three configurations. The judged measures, context precision, context recall and faithfulness, are computed by the locally served judge against the curated reference, and were computed for the naive configuration. Judged context precision and deterministic key-fact context precision apply different criteria of relevance, are reported under distinct labels throughout, and are not interchangeable. The denominators of the measures differ by definition rather than by accident: key-fact accuracy is computed over all twelve queries; key-fact context precision over the nine for which it is defined, a reference asserting an absence containing no fact that a retrieved chunk could rank above another; judged context recall over the ten for which it is defined, the two negative-case queries being excluded on the same ground; and faithfulness over the eleven responses from which a claim could be extracted. Every mean reported below is accompanied by the number of defined scores from which it was computed. The answer relevancy metric anticipated at the proposal stage was superseded by reference-based context recall and deterministic key-fact accuracy once ground-truth references had been curated for every query.

Table 1: Evaluation Measures, Definitions and Acceptance Criteria

Measure	Definition	Reference required	Construct captured	Acceptance criterion
Faithfulness (judged)	Response claims inferable from the retrieved context, divided by total response claims	No	Grounding / hallucination resistance	≥ 0.80
Context precision (judged, with reference)	Mean precision at k weighted by the relevance indicator over the retrieved chunks, judged against the whole reference	Yes	Retrieval noise	Reported; no a priori threshold
Context recall (judged)	Reference claims supported by the retrieved context, divided by total reference claims	Yes	Retrieval coverage	Reported; no a priori threshold
Key-fact context precision (judge-free)	Average precision over the retrieved chunks, a chunk counting as relevant where it carries at least one salient reference fact	Yes	Retrieval noise, lexical criterion	Reported; no a priori threshold
Key-fact accuracy (judge-free)	Salient reference facts present in the response, divided by total salient reference facts	Yes	End-to-end accuracy	Reported; no a priori threshold
End-to-end response latency	Wall-clock time from query submission to response completion	Not applicable	Operational performance	Mean ≤ 30 s
System Usability Scale	Ten-item standard instrument scored 0 to 100 (Brooke, 1996)	Not applicable	Perceived usability	Reported; interpreted against Bangor et al. (2008)

3.4 Experimental Procedure

Evaluation was executed in two stages. In the capture stage, each of the twelve queries was submitted to the live application through the same interface used by clinical staff, and the retrieved contexts, the generated answer, the retrieval route and the sequence of pipeline stages executed were recorded verbatim to a dataset file, one record per query, so that scoring operated on the system's actual behaviour rather than on a reconstruction. Capture was repeated under each of the three configurations. The full configuration enabled query transformation, hybrid dense-lexical retrieval, reranking, contextual compression, corrective retrieval grading and query routing. The naive baseline disabled all of these, corresponding to plain vector search followed by generation. The no-rerank ablation enabled all stages except reranking. In the scoring stage, the captured datasets were graded, the deterministic measures being recomputed directly from the captured answers and contexts for all three configurations and the judged measures being computed for the naive configuration.

The generation model under evaluation was Phi-4 Mini Instruct (3.8B) throughout, and the judge model was Qwen2.5-7B-Instruct. Three criteria governed the choice of judge. Independence: the two models originate from different developers and different training lineages, so the judge is independent of the generator at the level of family as well as instance, which the self-preference literature makes necessary. Capacity: the judge carries roughly twice the parameter count of the generator, since decomposing a response into atomic claims and verifying each against retrieved context is a

more demanding task than the generation it assesses. Locality: both models were served by Microsoft Foundry Local on the evaluation machine, so that neither the health-record content nor the generated responses left the local environment at any point during evaluation, preserving at evaluation time the privacy property the architecture claims at run time. Because the available accelerator memory did not accommodate both models simultaneously, the harness unloaded the generation model and loaded the judge between the capture and scoring stages; this is a constraint of the evaluation environment rather than of the deployed system, which serves only the generation model. An earlier configuration in which the 3.8-billion-parameter generator graded its own output was subject to self-preference bias and insufficient judging capacity simultaneously, and its scores are not reported here. Table 2 summarises the three configurations.

Table 2: Pipeline Configurations Compared

Configuration	Stages enabled	Queries scored	Measures computed	Purpose
Full pipeline	Query transformation, hybrid retrieval, reranking, compression, corrective grading, routing	12	Deterministic; judged from a separately archived run	Performance of the designed system as specified
Naive baseline	Plain vector search and generation only	12	Deterministic and judged	Floor against which advanced techniques are compared
No-rerank ablation	All stages except reranking	12	Deterministic	Isolates the marginal contribution of reranking

3.5 Data Analysis

Retrieval results were reported as mean scores for each measure under each configuration, together with the number of defined scores, since the denominators differ by definition. Comparison across configurations rests on the two deterministic measures, which were computed for all three; the judged measures characterise the naive configuration and are not used to rank configurations. Disaggregation by query class was used to characterise performance across retrieval scenarios and to locate the sources of end-to-end error. Latency was reported as a mean end-to-end response time per configuration against the thirty-second design criterion. Expert data were reported as recorded task completion counts, medians with interquartile ranges for agreement items, and the summary statistics of the published System Usability Scale totals. Given twelve scored queries, two per class, and thirteen expert participants, results are interpreted as indicative rather than confirmatory, and no inferential significance testing was conducted, the per-cell counts not supporting it.

4. Results

4.1 Overall Performance by Pipeline Configuration

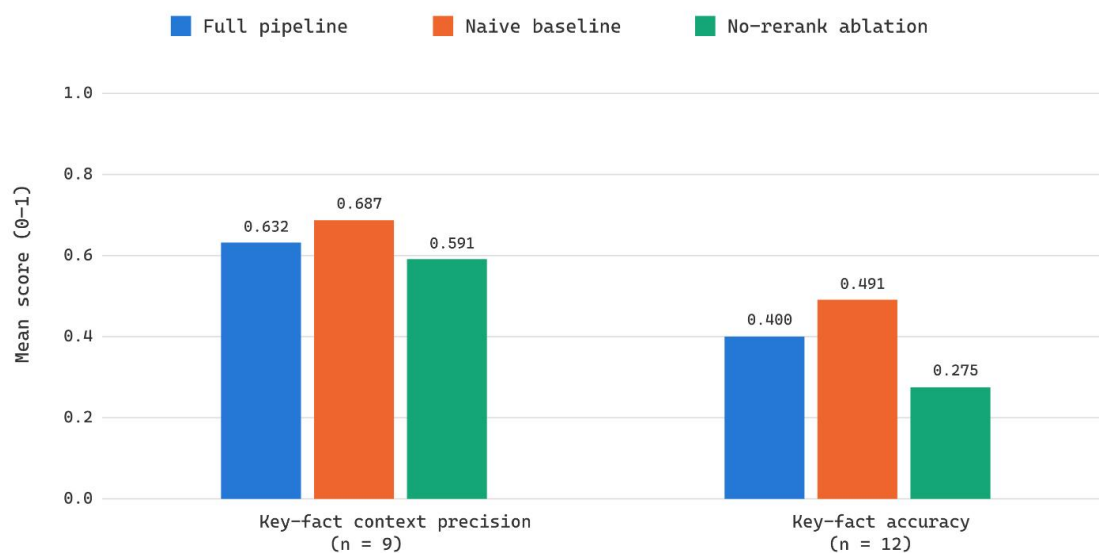
Table 3 presents the mean scores for each configuration and Figure 2 presents the comparison graphically. Two findings emerge. First, and most consequentially, the naive baseline outperformed the full advanced pipeline on both measures available across all three configurations. Key-fact context precision ordered the configurations naive (0.687), full (0.633) and no-rerank (0.591), and key-fact accuracy ordered them identically at 0.491, 0.400 and 0.275. Since retrieval depth was constant at eight chunks, the advanced stages cost approximately 0.054 of context precision and 0.091 of key-fact accuracy relative to plain vector search, and the no-rerank ablation was weakest on both, indicating that the loss arises from the interaction of stages rather than from reranking alone. Second, retrieval is functioning but incompletely. Under the naive configuration, judged context recall of 0.533 over the ten queries for which the measure is defined indicates that rather more than half of the claims required by the reference answers were present in the retrieved context, and faithfulness of 0.599 over eleven indicates that the majority of generated statements were supported by whatever content had been retrieved. Faithfulness falls below the criterion of 0.80 adopted by the study, and the criterion is therefore recorded as unmet. Low recall and low faithfulness identify distinct concerns: the first that information required by the reference was not retrieved, the second that generated content was not supported by what was retrieved.

Table 3: Results by Configuration, with Source and Coverage

Measure	Full pipeline	Naive baseline	No-rerank ablation
Key-fact context precision	0.633 (n = 9)	0.687 (n = 9)	0.591 (n = 9)

Measure	Full pipeline	Naive baseline	No-rerank ablation
(judge-free)			
Key-fact accuracy (judge-free)	0.400 (n = 12)	0.491 (n = 12)	0.275 (n = 12)
Judged context precision	0.0696 (n = 12), archived run	0.063 (n = 12)	Not computed
Judged context recall, negative cases excluded	0.362 (n = 10), archived run	0.533 (n = 10)	Not computed
Judged faithfulness	0.5395 (n = 12), archived run	0.599 (n = 11)	Not computed
Mean end-to-end response time	Not available	13.99 s	34.33 s

Note. Parentheses give the number of defined scores from which each mean was computed; the denominators differ by definition, not by omission. Key-fact measures are deterministic and require no model call. Judged values for the full configuration derive from a separately archived run and are disclosed as such rather than presented as a simultaneous repeat; no judged run was found for the no-rerank configuration. Generation model: Phi-4 Mini Instruct (3.8B). Judge model: Qwen2.5-7B-Instruct, served locally. Retrieval depth: eight context chunks throughout. No inferential ranking of judged scores is offered.



Note. Judge-free measures. Values are shown to three decimal places.

Figure 2: Key-Fact Context Precision and Key-Fact Accuracy by Pipeline Configuration

Contrary to expectation, the full advanced pipeline held no advantage on this evaluation set. The naive baseline led on both deterministic measures, which are the only two computed across all three configurations and therefore the only basis on which the configurations can be compared. The comparison is descriptive: it establishes that these three saved captures ordered in this way, and it does not establish that any individual pipeline stage caused the loss, nor that naive retrieval is superior in general.

4.2 Ablation Analysis

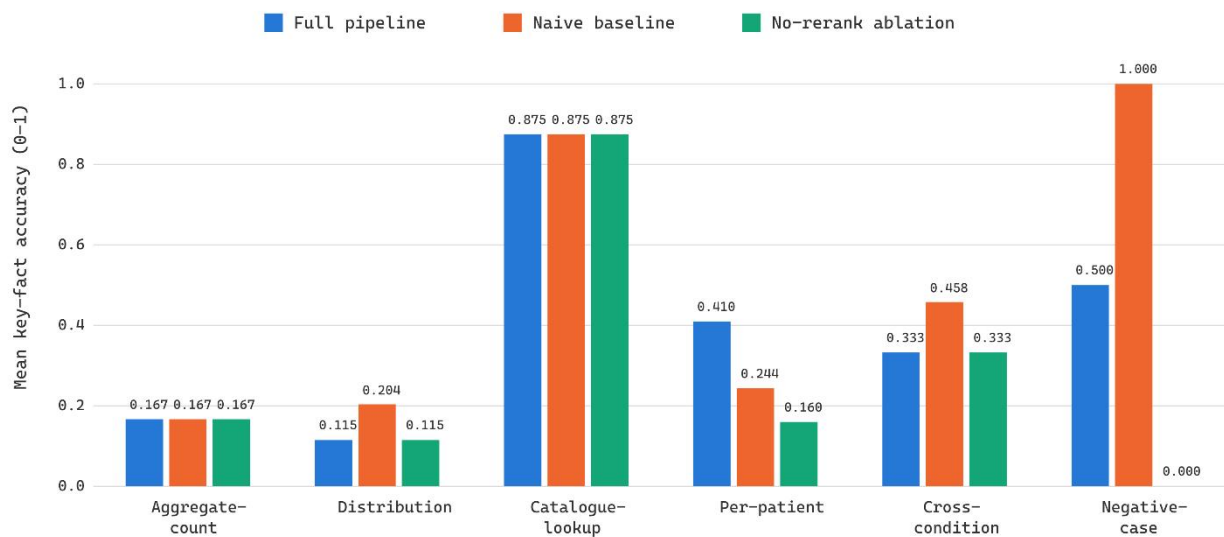
Because the full configuration and the no-rerank configuration differ in exactly one stage, the comparison isolates the contribution of reranking. Disabling reranking lowered key-fact context precision from 0.633 to 0.591 and key-fact accuracy from 0.400 to 0.275, so reranking was not the source of the deficit relative to the naive baseline; removing it made matters worse. Read with the naive result, the pattern indicates that the loss arises from the interaction of the advanced stages rather than from any one of them, and that the combination of query rewriting, compression and corrective grading is what discards chunks carrying reference facts. The direction is coherent and diagnostic. These stages are designed to manage abundance, promoting the most relevant candidates and displacing marginal ones; at a corpus of 1,603 chunks with a heterogeneous query mix, the candidate pool is small enough that filtering of this kind removes evidence the answer required. The result is therefore scale-dependent rather than a refutation of the techniques, and it should not be generalised to production-scale corpora. It does, however, establish for this facility and this corpus that the additional stages were not merely inert but counterproductive.

4.3 Performance by Query Class

Table 4 and Figure 3 disaggregate performance by query class, locating the sources of the modest aggregate scores. With two queries in each class, these results characterise the shape of the system's behaviour rather than quantifying it precisely.

Table 4: Key-Fact Accuracy and Judged Scores by Query Class

Query class	Queries	Key-fact accuracy, full	Key-fact accuracy, naive	Key-fact accuracy, no- rerank	Judged context recall, naive	Judged faithfulness, naive
Aggregate-count	2	0.167	0.167	0.167	0.667 (2)	1.000 (2)
Distribution	2	0.115	0.204	0.115	0.167 (2)	0.500 (2)
Catalogue-lookup	2	0.875	0.875	0.875	1.000 (2)	0.459 (2)
Per-patient	2	0.410	0.244	0.160	0.000 (2)	0.000 (1)
Cross-condition	2	0.333	0.458	0.333	0.834 (2)	0.700 (2)
Negative-case	2	0.500	1.000	0.000	Not defined (0)	0.637 (2)
All classes	12	0.400	0.491	0.275	0.533 (10)	0.599 (11)



Note. Two queries per class ($n = 2$); twelve queries overall. Zero indicates a score of 0.000.

Figure 3: Key-Fact Accuracy by Query Class and Pipeline Configuration

Performance varied sharply by class rather than uniformly. Under the naive configuration, catalogue-lookup queries recorded key-fact accuracy of 0.875, the strongest non-negative class in the benchmark; identifier-keyed per-patient queries recorded 0.244 with judged context recall of 0.000 across both queries in the class; and aggregate-count queries recorded 0.167 under every configuration, with distribution queries between 0.115 and 0.204. The pattern is internally consistent and identifies two failure mechanisms precisely. Aggregate and distribution answers were faithful to what had been retrieved, judged faithfulness for the aggregate class being 1.000, but reported on the retrieved sample rather than on the corpus, because a fixed window of eight chunks cannot enumerate every record relevant to a whole-database count. Per-patient answers failed because an exact identifier such as SYN-2024-0035 carries little semantic signal for an embedding model to match, so identifier-keyed lookup is poorly served by similarity search. The two negative-case queries recorded key-fact accuracy of 1.000 under the naive configuration; this is a result obtained on two benchmark items and is not evidence that absence is reliably detected across clinical records. Given two queries in each class, these disaggregated results are indicative rather than conclusive.

4.4 Operational Performance

Because the system is intended for deployment on hospital hardware, operational characteristics were recorded during evaluation. Table 5 reports the mean end-to-end response time for each configuration against the thirty-second design criterion. The criterion was met by the naive configuration at a mean of 13.99 seconds and was not met by the no-rerank configuration at 34.33 seconds; no comparable figure was available for the full configuration. This is a substantive result rather than merely an operational one. The naive baseline is not only the more accurate configuration on both measures reported across all three pipelines; it is also roughly two and a half times faster, and it is the only configuration evaluated that satisfies the response-time criterion. The advanced retrieval stages therefore imposed a cost on accuracy and on latency simultaneously while conferring no measured benefit on either.

Table 5: Reported Operational Performance

Measure	Naive baseline	No-rerank ablation	Full pipeline
Mean end-to-end response time	13.99 s	34.33 s	Not available
Mean meets the thirty-second design criterion	Yes	No	Cannot be determined
Resident memory, proportion of 32 GB	~70%	~70%	~70%
Processor utilisation	~10%	~10%	~10%
Graphics processor utilisation	80–100%	80–100%	80–100%

Note. Evaluation hardware: Lenovo Yoga Slim 7; Intel Core Ultra 7 155H; 32.0 GB LPDDR5-6400; Intel Arc integrated graphics with 17.9 GB shared memory; Windows 11 25H2. Inference was served through OpenVINO on the integrated graphics processor. The capture harness does not instrument per-query elapsed time, so the figures are coarse end-to-end means derived from the modification timestamps of the captured dataset files, written at the end of their respective configuration runs during a single sequential session; the full configuration is not reported because no preceding timestamp anchors the start of its run.

Two further observations follow. Resident memory occupancy of approximately 70 per cent of 32 gigabytes confirms that the system operates within the memory envelope specified as a design constraint, but with a working margin under 10 gigabytes. That margin is sufficient for the evaluated corpus of 1,603 chunks and should not be assumed sufficient for a production corpus at a Level 5 hospital, which would be substantially larger, since the vector index grows with the chunk count; memory scaling with corpus size was not measured. Inference executed throughout on an integrated graphics processor at 80 to 100 per cent utilisation, so the experiment demonstrates neither operation on a processor alone nor performance on the hospitals' own servers. A distinction must also be preserved between two costs. The response latency reported above is the deployment cost, experienced by a clinician awaiting an answer. Scoring the benchmark with the language-model judge consumed 2,828 seconds, or 47.1 minutes, for one configuration, with the judge running serially on the accelerator. That is an evaluation cost, incurred offline and never experienced in clinical use, and it is reported because the recommendation for routine local evaluation depends on what such a run costs on hospital-grade hardware.

4.5 Assessment by Intended Users

The system was assessed by thirteen purposively selected intended users working with the naive configuration over synthetic task scenarios. The panel comprised five administrators, four information technology or health-records officers, two medical or clinical officers and two nurses, seven drawn from the county referral hospital and six from the faith-based hospital, so that it included health professionals as well as administrative and technical users. The panel did not compare pipeline configurations under controlled conditions, and the results below are not presented as such a comparison. Table 6 summarises the recorded outcomes. Completion of the six observed task scenarios was recorded at between 84.6 and 100.0 per cent, with assistance or reformulation noted against each task whose count fell below thirteen; because the observer records disclose no consistent rule governing whether an assisted or reformulated attempt was counted as complete, the counts are reported as recorded completion and are not relabelled as unaided success, and no pooled success rate is computed from them. Of the five ratings of response quality, source verifiability attracted the highest median at 5, the other four returning medians of 4. The thirteen published System Usability Scale totals ranged from 75.0 to 85.0 with a mean of 80.6 and a sample standard deviation of 3.6.

Table 6: Summary of the Assessment by Intended Users (n = 13)

Measure	Value	Note
Recorded task completion, six scenarios	84.6–100.0%	Lowest on the aggregate-count scenario;

Measure	Value	Note
		two participants reformulated before completing
Relevance of the retrieved records	Md = 4 (IQR 1)	Perceived relevance during synthetic tasks
Correctness of the answer, given the records retrieved	Md = 4 (IQR 1)	Perceived, not independently established
Sources permitted the answer to be verified	Md = 5 (IQR 1)	Highest-rated quality item
System Usability Scale total	M = 80.6, SD = 3.6, range 75.0–85.0	Scored 0–100; not a percentage of clinical accuracy
That processing remains on the premises matters to me	Md = 5 (IQR 0)	Acceptance item
That the system operates offline matters to me	Md = 5 (IQR 0)	Acceptance item
Most frequently endorsed obstacles to adoption	Trust in the answers (3); fit with workflow (3)	Multiple selections permitted

Note. Agreement items were rated on a five-point scale. A high median accompanied by an interquartile range of 0 does not establish unanimity. These ratings describe perceived relevance, correctness and verifiability during synthetic tasks; they do not independently establish clinical correctness, and they do not constitute a warrant for acting upon a response in the care of a patient. No elapsed task times were available, so no estimate of time saved is offered.

The panel also rated how far each of the six query classes reflects work they actually perform, and Table 7 places those ratings beside the technical result for the same class. The juxtaposition is not an analysis of agreement, the two assessments not having been obtained over paired responses. Its force lies in a single alignment: per-patient retrieval was rated routine by all thirteen participants, the highest routine rating of any class, and is the class on which the system recorded the weakest result, at key-fact accuracy of 0.244 and judged context recall of 0.000 under the configuration they used. Aggregate-count and distribution queries, rated routine by nine and eight participants respectively, recorded 0.167 and 0.204. Catalogue lookup, the strongest class at 0.875, was rated routine by ten. The classes the panel identified as most routine are therefore, with one exception, the classes on which the system performed worst, and the juxtaposition identifies a priority for development rather than a finding of readiness.

Table 7: Reported Realism of Each Query Class alongside Technical Accuracy

Query class	Rated routine, n of 13	Routine (%)	Key-fact accuracy, naive	Reading
Aggregate-count	9	69.2	0.167	Indicates the value of corpus-wide structured counting
Distribution	8	61.5	0.204	Indicates the need for structured grouping and completeness
Catalogue-lookup	10	76.9	0.875	The strongest non-negative class in this benchmark
Per-patient	13	100.0	0.244	The highest reported routine relevance with weak accuracy
Cross-condition	9	69.2	0.458	Indicates the value of explicit patient joins and set intersection
Negative-case	6	46.2	1.000	Two benchmark successes; limited evidence

Acceptance of the display of sources does not validate the automated judge, nor does it qualify the retrieval weaknesses the benchmark records. The

panel assessed the system in a single session over synthetic tasks; no outcome was measured over live patient records or within routine workflow, and the coincidence between the size of this panel and the size of the requirements sample reported in the companion design study is not treated as evidence that the same individuals took part in both.

5. Discussion

5.1 Interpretation of Key Results

Four findings carry the interpretation. The first is that grounding and accuracy came apart. Faithfulness of 0.599 alongside key-fact accuracy of 0.491 under the naive configuration, and judged faithfulness of 1.000 on the aggregate class alongside key-fact accuracy of 0.167 on that same class, show a system that largely says only what its retrieved context supports while frequently failing to answer correctly, because the context it retrieved was incomplete. The constraint on end-to-end performance lies in retrieval, not in generation, and the distinction has immediate practical force: a facility observing only faithfulness would misread the system's quality, and reports must therefore present retrieval and accuracy measures alongside faithfulness.

The second is that the advanced pipeline did not earn its complexity on this evaluation set, and did not earn it twice over. It was the less accurate configuration on both deterministic measures, and the no-rerank ablation was also the slower one, at 34.33 seconds against 13.99 seconds for the naive baseline, failing the thirty-second design criterion that the naive configuration met. The advanced stages therefore imposed a cost on accuracy and on latency simultaneously while conferring no measured benefit on either. This is a scale-dependent result rather than a refutation of the techniques: reranking, compression and corrective grading are designed to manage abundance, and where retrieval is scarce rather than abundant they address a problem that does not obtain.

The third is that the judged context-precision figure of 0.063 requires interpretation rather than literal reading. The judged measure assesses each retrieved chunk against the reference answer in its entirety, asking whether that chunk alone was useful in arriving at the whole answer. Where a reference is a synthesis, such as a count, a list or a distribution, no single chunk can satisfy it, and the judge returns zero for every chunk even when precisely the correct chunks were retrieved; this was verified directly, a chunk naming one of the reference medications being rejected on the ground that it did not list the others. The interpretation is confirmed by the one benchmark query whose reference is a single proposition rather than a synthesis, concerning the co-occurrence of hypertension and diabetes, which scored 0.759 under the same metric and the same judge. The deterministic variant applies the same average-precision formula but counts a chunk as relevant if it carries at least one salient reference fact, and returns 0.687 over the same retrievals. Read together, the pair establishes that the retriever was surfacing relevant chunks and that the low judged figure reflects the interaction between a whole-reference criterion and a synthesis reference. That observation does not establish that every low judged score is an artefact of metric framing: reference framing, judge error and genuine retrieval deficiency are three distinct explanations, and separating them would require relevance labels assigned by clinicians to the same chunks. This is why both measures are reported.

The fourth is that the two weakest classes fail for the same structural reason, in opposite directions, and that the intended users identified one of them as their most routine task. Aggregate queries need evidence that is exhaustive, and a fixed top-k window cannot be exhaustive; per-patient queries need evidence that is exact, and semantic similarity is approximate by construction. Neither is a defect of the generation model, and neither is remediable by tuning the pipeline: both mark the boundary of what semantic retrieval can do, and both are addressed by complementing it with deterministic database aggregation and exact-match identifier indexes rather than by further optimising the vector path. That every one of the thirteen intended users reported per-patient retrieval as routine, while the system recorded key-fact accuracy of 0.244 and judged context recall of 0.000 on that class, converts a technical weakness into a development priority. Neither strand of evidence yields that observation alone: the benchmark establishes the weakness without indicating how often the operation is performed, and the panel establishes the frequency without detecting the weakness.

The observed faithfulness scores indicate that source grounding remains incomplete. This is consistent with the caution in Amugongo et al. (2025) that RAG does not guarantee full support from retrieved sources. Because the present evaluation did not include a non-RAG control, it cannot establish that retrieval reduced unsupported generation. The advanced configurations also failed to outperform the naive baseline on the reported deterministic measures. Wolk (2025) evaluated different pipelines and outcomes, so those findings are not a direct test of the present configuration. Corpus size and pipeline design are possible explanations for differences between studies, but their effects were not isolated here. The results support evaluating each configuration for its intended task rather than assuming that additional stages improve performance.

The class-level findings align closely with two independent strands. Yang et al. (2024) reported that RAG systems perform poorly on set-based questions in the Comprehensive RAG Benchmark, and Oche et al. (2025) documented robustness limitations under long-context reasoning; the systematic undercounting observed here is the same phenomenon in a health-records setting, with the added diagnostic detail that faithfulness remains at 1.000 while the count is wrong. The evaluation also differs from Korom et al. (2025) in an instructive way: that study measured clinical outcomes for

a cloud-dependent system and found benefit, whereas this one measures retrieval quality for an on-premises system and finds a specific, addressable limitation. The two are complementary rather than competing. Finally, the usability result should not be read as corroborating the technical one. A mean System Usability Scale total of 80.6, which falls in the upper band of the distribution reported by Bangor et al. (2008), was obtained alongside mean key-fact accuracy of 0.491; the two measure different constructs, and their divergence illustrates rather than resolves the relationship between perceived usability and retrieval correctness.

5.3 Theoretical Implications

Three implications follow. First, the study supplies empirical grounds for treating grounding and accuracy as separate constructs in RAG evaluation theory rather than as correlated indicators of a single quality dimension. The near-inverse pattern observed in the aggregate class, faithfulness at 1.000 with key-fact accuracy at 0.167, shows that the two can dissociate almost completely. Evaluation frameworks that report a composite score, or practice that treats faithfulness as a proxy for reliability, will therefore systematically misrepresent system quality. The claim is stronger than a recommendation to report both: faithfulness is conditional on the retrieved context and therefore cannot bound answer quality at all.

Second, the results identify a retrieval-aggregation mismatch as a structural property of semantic RAG rather than an implementation defect. Fixed top-k retrieval optimises for the most similar evidence, whereas aggregate questions require all qualifying evidence and identifier lookups require one exact record. These are different information-access problems, and no amount of embedding or reranking improvement reconciles them, because the mismatch is between the retrieval objective and the question type. RAG architecture theory should therefore incorporate query-class routing to structurally different retrieval mechanisms as a first-class design element, rather than treating a single semantic path as universal. Third, the study indicates that a metric's denominator and its reference framing are themselves theoretical commitments: the divergence between judged context precision of 0.063 and deterministic key-fact context precision of 0.687 over identical retrievals arises entirely from what each measure counts as relevant, which argues for reporting the criterion of relevance alongside the score.

5.4 Practical and Policy Implications

For hospital management, the results support piloting the system as an augmentation layer over existing electronic medical record systems for narrative record retrieval and catalogue lookup, where the demonstrated strength lies, while routing counting, distribution and patient-record lookup queries to conventional structured reports until aggregation-aware and identifier-aware retrieval is integrated and evaluated. The evidence does not support adopting the system for analytical reporting or patient-record lookup in its present form, and a decision on adoption should be deferred until evidence has been obtained over real records and within workflow. Usability and acceptance ratings, however favourable, are not a sufficient basis for that decision.

For implementers, three practices follow directly. The naive configuration should be adopted as the deployment default on both accuracy and latency grounds, advanced stages being enabled only where a facility-specific ablation demonstrates benefit; pipelines should be built to permit stage-level ablation so that such an evaluation is possible at all, a requirement that is architectural rather than procedural. Patient-identifier lookups should be routed to exact-match indexes and counting and distribution queries to deterministic database aggregation, a change of routing rather than of model. And the deterministic measures should be adopted as a zero-cost continuous quality-assurance tier, run on every configuration change and corpus update, with judged evaluation reserved for periodic audit: the deterministic measures require no model call and were computable for all three configurations, whereas judged scoring of a single configuration consumed 2,828 seconds, or 47.1 minutes, on the evaluation hardware, which makes a full three-configuration judged evaluation an unattended overnight job rather than a project in itself.

For facilities and vendors, locally executed evaluation of this kind should be adopted as standing quality assurance before and during clinical deployment, reporting faithfulness alongside retrieval-quality measures so that grounding strength is not mistaken for answer accuracy, and stating the denominator and the criterion of relevance for every figure reported. For regulators, including the Office of the Data Protection Commissioner and the Ministry of Health, this study demonstrates that meaningful performance evaluation can be conducted entirely on-premises without transmitting health data to external services, which means that evidence of pre-deployment evaluation can reasonably be required of clinical artificial intelligence systems without creating a conflict with the localisation requirement of Section 50 of the Act and Regulation 26 of the 2021 Regulations.

6. Conclusion, Limitations and Future Work

6.1 Summary of Main Findings

This article evaluated the retrieval performance, response latency and perceived usability of a developed On-Premises RAG system for health records retrieval. Twelve curated queries across six classes, each with a ground-truth reference computed by structured query over a synthetic corpus of 1,603

chunks, were executed through the live pipeline under three configurations at a constant retrieval depth of eight chunks. The naive baseline recorded the highest key-fact context precision at 0.687 over nine defined scores, against 0.633 for the full pipeline and 0.591 for the no-rerank ablation, and the highest key-fact accuracy at 0.491 over twelve, against 0.400 and 0.275; it was also the only configuration to meet the thirty-second design criterion, at a mean of 13.99 seconds against 34.33 seconds for the no-rerank ablation. Under the naive configuration, judged context precision was 0.063 over twelve scores, context recall 0.533 over ten and faithfulness 0.599 over eleven, the last falling below the criterion of 0.80. Disaggregation attributed the weakest results to aggregate counting, which was faithful but reported on the retrieved sample rather than the corpus, and to identifier-keyed per-patient lookup, which semantic similarity served poorly; catalogue lookup was strongest at 0.875. Thirteen intended users recorded completion of between 84.6 and 100.0 per cent of six task scenarios, rated source verifiability highest at a median of 5, returned a mean System Usability Scale total of 80.6 with a sample standard deviation of 3.6, and returned medians of 5 with an interquartile range of 0 on the statements that on-premises processing and offline operation matter; all thirteen identified per-patient retrieval as routine. The study concludes that hallucination in the strict sense was not the dominant failure mode and that incomplete retrieval was; that advanced retrieval stages conferred no measured benefit at this corpus size and query mix while costing both accuracy and latency; and that fixed top-k semantic retrieval is structurally mismatched to whole-database aggregation and exact identifier lookup. The evaluation objective was therefore achieved, and yielded a basis for further development and controlled testing rather than evidence of clinical effectiveness or readiness for deployment.

6.2 Limitations

Six limitations qualify these conclusions. First, and most consequentially, the technical benchmark comprised twelve queries with two in each class, which supports indicative rather than confirmatory inference, precludes significance testing, and means that individual query behaviour exerts substantial influence on class means. Second, the judged measures were computed for the naive configuration only, with a separately archived run supplying judged figures for the full configuration and none found for the no-rerank ablation, so the cross-configuration comparison rests on the deterministic measures alone; a single archived run does not establish reproducibility across repeats. Third, the judge was a seven-billion-parameter model selected because the residency constraint required local hosting, and no automated judge has been validated against expert human adjudication for Kenyan health-record retrieval, so some proportion of the low judged scores may reflect judge behaviour rather than system behaviour. Fourth, the corpus was synthetic and small at 1,603 chunks, and both the absolute scores and the ablation result are plausibly scale-dependent, so the finding that advanced techniques do not help should not be generalised to production-scale corpora. Fifth, the latency figures are coarse end-to-end means derived from the modification timestamps of the captured dataset files rather than from per-query instrumentation, no figure was available for the full configuration, and a mean does not establish that every query met the target; inference also ran throughout on an integrated graphics processor, so operation on a processor alone was not demonstrated. Sixth, the assessment by intended users was a single session over synthetic tasks with no controlled comparison of configurations, the observer records disclose no consistent rule governing whether an assisted or reformulated attempt counted as complete, and no elapsed task times were available, so no pooled success rate and no estimate of time saved can be computed; clinical outcomes and performance in routine workflow remain outside the evidence.

6.3 Future Research Directions

Four directions follow. First, aggregation-aware and identifier-aware retrieval should be completed and evaluated, classifying each incoming question by class and dispatching counting, distribution and co-morbidity questions to deterministic aggregation over the relational store, identifier-keyed questions to exact-match lookup, and narrative and catalogue questions to semantic retrieval, then composing the results into a single grounded answer with its sources displayed. The benchmark used here, scored by class, would test the improvement directly: the hypothesis is that the three weak classes rise substantially while catalogue-lookup performance is retained. Such a study should also report what the routing decision costs in latency, since the thirty-second criterion must still be met, and how the router behaves on questions that fall between classes, which is where a scheme of this kind is most likely to fail. Second, the evaluation should be repeated on a substantially larger and statistically powered query set with balanced class representation, over actual de-identified institutional records at production scale, so that the scale-dependence of the ablation finding can be tested directly; elapsed time should be instrumented within the capture harness itself and judged metrics computed for all three configurations. Third, the judge requires validation against adjudication by clinicians on a shared sample, since whether its scores track clinical judgement is at present assumed rather than demonstrated, and no automated judge should be used to certify a clinical system until that question is settled. Fourth, evaluation should be extended to clinician-facing outcomes through a counterbalanced study within workflow, recording unaided first-attempt success, assisted completion, reformulation, final factual correctness and elapsed task time as separate quantities, so that the relationship between the retrieval measures reported here and practical clinical value can be established.

REFERENCES :

- Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6), e0000877. <https://doi.org/10.1371/journal.pdig.0000877>
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.
- Data Protection (General) Regulations, 2021, Legal Notice No. 263. (2021). Republic of Kenya. Kenya Gazette Supplement No. 186. <https://www.odpc.go.ke/wp-content/uploads/2023/05/Kenya-Gazette-Supplement-No.-186-Data-Protection-General-Regulations-2021.pdf>
- Data Protection Act, No. 24 of 2019. (2019). Republic of Kenya. Kenya Gazette Supplement. <https://www.odpc.go.ke/wp-content/uploads/2023/01/Kenya-Data-Protection-Act-2019.pdf>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3900–3906). <https://arxiv.org/abs/2309.15217>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. <https://arxiv.org/abs/2312.10997>
- Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., & Kang, D. (2024). Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 517–545). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.29>
- Korom, R., Kiptinness, S., Adan, N., Said, K., Ithuli, C., Rotich, O., Kimani, B., King'ori, I., Kamau, S., Atemba, E., Aden, M., Bowman, P., Sharman, M., Soskin Hicks, R., Distler, R., Heidecke, J., Arora, R. K., & Singhal, K. (2025). AI-based clinical decision support for primary care: A real-world study. *arXiv preprint arXiv:2507.16947*. <https://arxiv.org/abs/2507.16947>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* 33 (pp. 9459–9474). Curran Associates.
- Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*. <https://arxiv.org/abs/2507.18910>
- Park, S., Kim, J., Lee, H., & Choi, Y. (2025). Enhancing clinical decision support with adaptive iterative self-query retrieval for retrieval-augmented large language models. *PMC*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12383471/>
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Spiliopoulou, E., Fogliato, R., Burnsky, H., Soliman, T., Ma, J., Horwood, G., & Ballesteros, M. (2025). Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. *arXiv preprint arXiv:2508.06709*. <https://arxiv.org/abs/2508.06709>
- Wampler, D., Nielson, D., & Seddighi, A. (2025). Engineering the RAG stack: A comprehensive review of the architecture and trust frameworks for retrieval-augmented generation systems. *arXiv preprint arXiv:2601.05264*. <https://arxiv.org/abs/2601.05264>
- Wataoka, K., Takahashi, T., & Ri, R. (2024). Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*. <https://arxiv.org/abs/2410.21819>
- Wolk, K. (2025). Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment. *Electronics*, 14(21), Article 4227. <https://doi.org/10.3390/electronics14214227>
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*. <https://arxiv.org/abs/2402.13178>
- Yang, X., Sun, K., Xin, H., Sun, Y., Bhatt, N., Jiang, X., Sun, S., Dong, L., & Choi, Y. (2024). CRAG: Comprehensive RAG benchmark. *arXiv preprint arXiv:2406.04744*. <https://arxiv.org/abs/2406.04744>
- Zhang, Y., Li, X., & Chen, H. (2025). Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*. <https://arxiv.org/abs/2504.14891>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in neural information processing systems* 36 (pp. 46595–46623). Curran Associates.