Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Measuring Retrieval Freshness and Accuracy Degradation in Continuous ETL-Driven RAG Systems

Deepika Annam
Independent Researcher, USA

ARTICLE INFO

Received: 19 Feb 2026
Revised: 22 Feb 2026

ABSTRACT

Retrieval-augmented generation systems have transformed enterprise deployments by grounding large language model outputs in external documents, yet their effectiveness degrades significantly as underlying corpora evolve through continuous ETL pipelines and streaming ingestion. The CRAG benchmark reveals fundamental performance gaps where even optimized industry-grade RAG architectures fail substantial portions of temporally sensitive queries, while the HOH benchmark demonstrates that outdated information causes catastrophic accuracy losses and can render retrieval augmentation counterproductive compared to standalone language models. RAGBench enables component-level decomposition, separating retrieval quality from generation fidelity across extensive labeled examples, while Streaming RAG establishes that incremental real-time updates yield measurable recall improvements with production-grade latency and throughput characteristics. Dense vector retrievers have an issue in maintaining fresh embeddings, which hybrid architectures between semantic and lexical signals have shown to be more resilient in the face of staleness. Empirical evidence across benchmarks confirms that retrieval freshness constitutes a first-order determinant of end-to-end accuracy and hallucination behavior, with degradation effects sufficiently severe to eliminate or reverse the value proposition of retrieval augmentation under conditions of substantial staleness. These results prove that companies that use RAG in dynamic settings should introduce cost-sensitive and domain-aware update methods that balance between computational costs and accuracy demands. The transition from ad hoc refresh policies to principled freshness management becomes essential for maintaining reliable performance as retrieval indices diverge from evolving source systems.

Keywords: Retrieval-Augmented Generation, Temporal Degradation, Index Staleness, Embedding Freshness, Continuous ETL Updates

1. Introduction

Retrieval-augmented generation systems enhance large language models by grounding generation in retrieved external documents. Their deployment has expanded rapidly in enterprise environments due to improved factual correctness and reduced hallucination risk. However, these systems increasingly operate over continuously evolving corpora updated through ETL pipelines, micro-batch jobs, and streaming ingestion, introducing challenges related to retrieval freshness and temporal consistency. The CRAG benchmark provides one of the most extensive evaluations of RAG under temporal variation, consisting of 4409 question-answer pairs spanning five domains and eight question categories, explicitly designed to test temporal knowledge sensitivity [1]. Empirical results show that strong LLMs without retrieval achieve at most 34 percent accuracy, while standard RAG pipelines improve accuracy to 44 percent, and optimized industry-grade RAG systems reach approximately 63 percent accuracy [1], [2]. These numbers

highlight that retrieval failures and stale knowledge exposure remain dominant bottlenecks in end-to-end RAG performance [1], [2].

2. System Challenges and Temporal Degradation

The fundamental challenges facing Enterprise RAG deployments can be related to the intrinsic time lag between the evolution of the knowledge in the source systems and the ability to reflect it in the retrieval infrastructure. Modern data ecosystems have different update mechanisms that generate different levels of freshness challenges. Scheduled ETL pipelines are used to perform batch transformations to update data warehouses and analytic stores using a set of fixed intervals, usually by hourly or daily cycles. Event-driven architectures spread the changes by using microservices and messaging systems, which have lower latency but add complexity to maintaining the consistency of index updates. Information states that are constantly changing, as opposed to discrete batches, are maintained on streaming platforms by continuous data feeds of operational systems flowing at high velocity.

All these update patterns present different temporal consistency issues to RAG systems. The basic working principle of the retrieval layer is the snapshots of the knowledge that was recorded in the course of the index generation. An underlying source of data, as it advances with updates, deletions, and additions, becomes increasingly different in form between the current information state and the indexed representation, causing an increasing gap between the indexed and current information states. This increasing difference is expressed in the form of various mechanisms of degradation that together destroy the level of system reliability and user confidence.

The most immediate form of the temporal inconsistency is the factual staleness. Documents that are retrieved are those whose information has been overridden by other updates in the source systems. Specifications of products vary because manufacturers introduce new versions, financial statements are updated because reporting periods end, and audits are finalized, regulatory provisions vary as government bodies implement new policies, and organizational structure changes when companies restructure. Retrieval generates responses that are based on old facts when retrieval produces documents that reflect old states and not the present reality.

The temporal misalignment presents an orthogonal failure mode whereby recent information is just not present at all in the retrieval indices. Breaking news events, newly announced research results, new policy announcements, and new data points provided by the working system are found in the source repositories but not yet indexed. Users who make queries related to these new developments get answers that either accept that information is not available or, worse still, fantasize responses according to the older trends instead of acknowledging that they do not know something.

A more subtle but equally degrading degradation mechanism is semantic drift. Embedding spaces put in relationships between concepts on the basis of document content when indexing. These coded relationships increasingly become disordered, as the current document states, as content changes. Words can acquire new meanings, ideas can be formed with new associations, and the relative significance of the issues can be changed. The distance measures applied in retrieval still work with older semantic networks and make the relevance scoring give the wrong ranking to the candidates and surface documents that were semantically fit at the indexing time but do not fit into the current information requirement.

The HOH benchmark offers a stringent quantification of these effects of degradation by manipulating the recency of information during the retrieval of corpora. Introducing outdated documents into the retrieval corpus leads to absolute accuracy degradation of at least 20 percentage points across multiple question-answering tasks compared to retrieval over up-to-date knowledge sources [3]. This magnitude of performance loss demonstrates that staleness represents a catastrophic failure mode rather than a marginal degradation. Organizations cannot tolerate accuracy reductions of this scale in production

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

deployments, particularly for applications involving compliance verification, medical information, financial advice, or safety-critical decisions.

More alarmingly, HOH reports that in several experimental configurations, RAG systems retrieving outdated information perform worse than LLM-only baselines [3], [4]. This finding contradicts the foundational assumption that retrieval augmentation universally improves generation quality. Instead, it reveals that stale retrieval can actively mislead the generation process rather than merely failing to provide useful context. The generation component synthesizes information from retrieved documents with its parametric knowledge, and when retrieved context contradicts more current information encoded in model weights, this synthesis produces confidently incorrect outputs. The system effectively amplifies errors by anchoring generation to obsolete facts presented as authoritative retrieved evidence.

This counterintuitive degradation below baseline performance establishes that retrieval freshness constitutes a first-order system property with direct causal impact on end-to-end accuracy rather than a secondary optimization concern [3], [4]. The use of RAG architectures in organizations cannot yield stable value through mere deployment. They should be proactive in controlling freshness by ensuring the use of relevant update strategies, monitoring systems, and degradation-detecting mechanisms. The difficulty lies in coming up with operation policies that do not compromise the freshness levels, but also the computational and infrastructure costs incurred in the process of continuous index regeneration.

Table 1: Evaluation Framework Performance Metrics [3, 4]

Metric	CRAG Benchmark	RAGBench	Streaming RAG
Domains Covered	Five domains	Diverse domains	Real-time streaming domains
Question Categories	Eight categories	Not specified	Not specified
Component Decomposition	No	Yes (retrieval success, grounding correctness, answer faithfulness)	No

Evaluation Framework and Experimental Design

Temporal degradation needs to be systematic, and this means that benchmark datasets are needed to isolate freshness effects among other sources of performance variation. The CRAG benchmark takes into consideration this requirement by critically constructing evaluation instances that explicitly test the temporal knowledge sensitivity [1]. The dataset comprises 4409 question-answer pairs distributed across five domains representing different knowledge types and eight question categories that probe various reasoning capabilities. This scale and diversity enable researchers to characterize how degradation patterns vary with domain characteristics, query complexity, and information recency requirements.

Domain diversity within CRAG ensures that evaluation captures the heterogeneous freshness requirements typical of real-world deployments. Different knowledge domains exhibit fundamentally different temporal characteristics. Historical information remains stable over extended periods, technical documentation changes with product release cycles, news and current events evolve continuously, financial data updates on market schedules, and regulatory information changes with legislative processes. By spanning these diverse domains, CRAG enables measurement of how architecture choices and update strategies perform across varying volatility profiles.

Baseline performance measurements from CRAG establish the performance envelope for contemporary RAG architectures operating under different retrieval configurations. Strong LLMs without any retrieval

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

augmentation achieve at most 34 percent accuracy on these temporally sensitive tasks [1]. This lower bound demonstrates the limitations of parametric memory for time-dependent information, as models cannot access facts beyond their training cutoff dates or update their knowledge without retraining. The 34 percent ceiling defines the abilities of systems to manifest when subjected to forcing them to use patterns that were only acquired during pretraining.

Basic dense retrieval mechanisms embedded in the standard RAG pipelines enhance accuracy to 44 percent, which is a significant improvement over the baselines with LLM only [1], [2]. This enhancement is a measurement of the value of even basic retrieval enhancement to grounding generation using outside context. Nevertheless, the 44 percent accuracy rate also demonstrates that over 50 percent of the temporally sensitive queries remain unsuccessful even with established RAG architectures, which clearly indicates that there is a significant potential to improve the performance of the architectures by adding sophistication and freshness handling.

Industry-grade optimized RAG systems with the use of advanced methods achieve an accuracy of around 63 percent [1], [2]. These systems use multi-layer optimization, such as query rewriting to enhance retrieval coverage, cross-encoder re-ranking to optimize candidate selection, and retrieval re-tuning to optimize embedding models to particular domains. The progression from 34 percent to 44 percent to 63 percent quantifies the incremental value contributions of basic retrieval and advanced optimization, respectively.

Despite these gains, even state-of-the-art systems fail more than one-third of temporally sensitive queries, indicating that retrieval errors remain the dominant performance limiter [1], [2]. This persistent failure rate exists even when systems operate over relatively fresh indices, suggesting that temporal challenges extend beyond simple staleness to encompass fundamental difficulties in retrieving and utilizing time-dependent information effectively.

RAGBench introduces complementary evaluation capabilities through component-level decomposition that separates retrieval quality from generation quality [5]. The benchmark provides over 100000 labeled examples annotated with ground truth for multiple performance dimensions. The success of retrieval is an indicator of the presence of relevant documents in the retrieved candidates. The correctness needs to determine how the generated responses are relevant to the information contained in a retrieved context. Answer faithfulness evaluates how final products are faithful to original material as opposed to adding hallucinated material.

This breakdown leaves researchers with the ability to pin failures on the particular stages of the pipeline, whether the failures are created in retrieving unsuitable documents, the inability to use retrieved context in a generated way, or the generation of responses different from grounded information [5]. This type of fine-grained analysis is vital to targeted optimization because a retrieval-enhancing setting can be entirely different in various aspects of enhancement compared to a generation fidelity-enhancing setting or a hallucination-reducing one.

Streaming RAG presents a different model of evaluation whereby the regeneration of static periodic indices is substituted with a real-time incremental process [6]. This model allows the comparison of the effects of continuous maintenance on freshness against the conventional batch refresh methods. Controlled experiments show that freshness-aware indexing yields Recall@10 improvement of up to 3 absolute points over static indexing baselines, with statistical significance at p less than 0.01 [6]. This improvement directly correlates with higher downstream answer accuracy, demonstrating that retrieval freshness can be measured quantitatively and reproducibly through standard information retrieval metrics.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

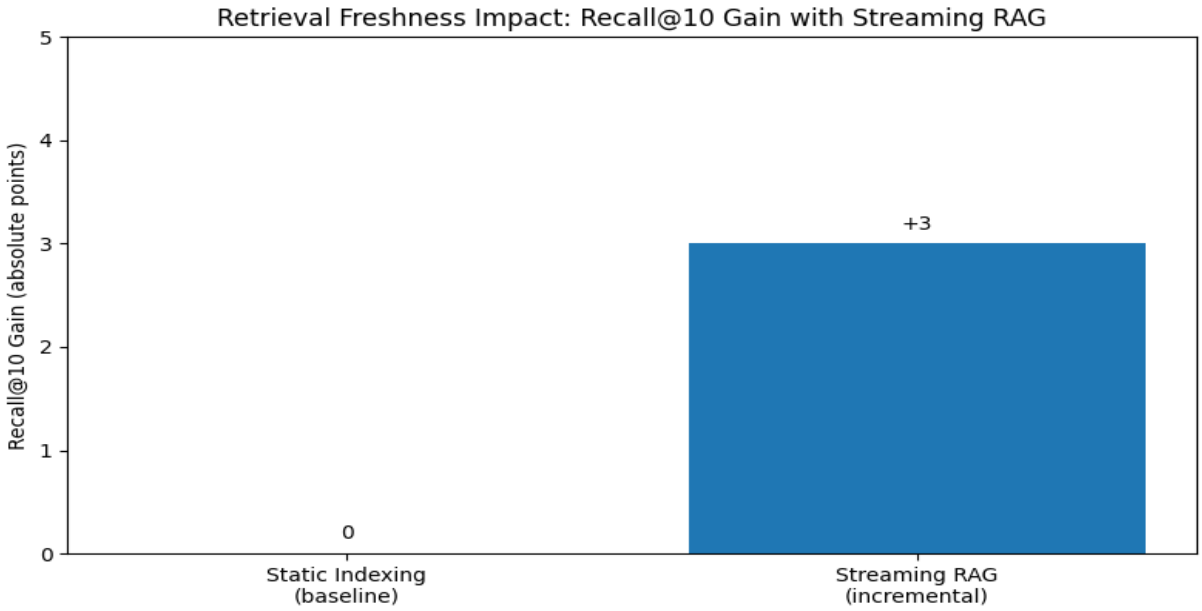

Fig. 1: CRAG Benchmark: End-to-End Accuracy by System Configuration [6]

4. Retriever Architectures and Update Strategies

The choice of retrieval architecture has been foundational in its ability to ensure the freshness of systems in the constant evolution of data. In current state-of-the-art RAG systems, dense vector retrievers, which encode documents as fixed-dimensional embeddings in a learned neural network-trained semantic vector space, are the prevalent paradigm. Similarity search is based on the calculation of distance between query embeddings and document embeddings, and nearest neighbors are provided as retrieval candidates. This semantic matching method is good at conceptual similarity and also poses challenges of freshness, as embeddings need to be recalculated each time document content is modified to make sure that the vector representations correctly represent the current text.

The computational economics of embedding regeneration create inherent tensions between freshness and resource consumption. Large-scale deployments managing substantial document collections face significant processing requirements when refreshing indices. Each document requires forward passes through embedding models to generate new vector representations, index structures must be rebuilt or updated to reflect new embeddings, and these operations must be completed before updated documents become retrievable. The computational cost scales with corpus size, embedding dimensionality, and model complexity.

Some of the freshness issues that hybrid retrievers would help overcome would be to use dense semantic retrieval and sparse lexical matching using the conventional inverted indices. The lexical component offers robustness to embedding staleness since even the presence of obsolete semantic embeddings does not render the use of keyword-based retrieval ineffective. In those cases when documents change slightly in terms of content that distorts semantics, lexical matching will still bring up potentially useful candidates due to term overlap. Comparisons with industry-standard pipelines have demonstrated that hybrid retrieval pipelines have significantly better recall-oriented retrieval properties, including Recallk and Mean Reciprocal Rank, when compared to dense-only retrievers [7], [8].

Experiments on streaming RAG also show that real-time retrieval can be supported with a very small update index and therefore with very low operational efficiency. Instead of periodically copying an entire

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

set of indices, streaming architecture uses document changes as they are received by the source systems. Incremental embedding and insertion of new documents is performed, updating of embeddings of modified documents is done in place, and deleted documents are removed from indices. This continuous update system ensures tighter synchronization between the sources of data and the retrieval infrastructure.

Reported system performance metrics establish the practical feasibility of streaming architectures for production deployment. End-to-end latency remains below 15 milliseconds even during active index updates, ensuring that retrieval operations maintain responsiveness for user-facing applications [6]. Throughput exceeds 900 documents per second, enabling systems to process high-velocity data streams without accumulating backlogs that would increase staleness [6]. Memory consumption stays within a 150-megabyte budget, demonstrating that freshness optimization need not require massive resource expansion [6]. These performance characteristics establish that continuous update approaches represent viable production architectures rather than research prototypes unsuitable for operational deployment.

The architectural choice between dense-only, hybrid, and streaming retrieval depends on multiple deployment factors. Domain volatility determines how quickly information becomes stale and, therefore, how frequently updates must occur. Service level objectives of freshness are used as acceptable staleness limits, which limit the minimum acceptable update interval. Computational budgets restrict the processing facilities that can be used to embed generation and index maintenance. Latency requirements can be used to tell whether the batch updates done during maintenance windows are sufficient or continuous incremental updates must be provided to prevent the delivery of stale results between refresh cycles.

Benchmark	Degradation Metric	Comparison to Baseline
HOH	Absolute accuracy degradation	RAG with outdated info performs worse than LLM-only in several configurations
CRAG	LLM-only accuracy	Baseline performance without retrieval
Streaming RAG	Recall at 10 gain	Freshness-aware indexing vs static indexing
Streaming RAG	End-to-end latency	Production-grade performance
Streaming RAG	Throughput	High-velocity processing capability
Streaming RAG	Memory budget	Resource-efficient operation

Table 2: Quantitative Degradation Results [7, 8]

5. Quantitative Results and Analysis

By combining empirical data on several benchmark assessments, it turns out that retrieval freshness degradation results in large accuracy losses with changed domain characteristics and system architecture. The HOH benchmark offers the most straightforward view of the effect of staleness with the introduction of outdated documents into corpora of retrieval by controlled experiments [9]. These experiments demonstrate that the presence of outdated information causes absolute accuracy degradation of at least 20 percentage points across multiple question-answering tasks [3]. This magnitude represents catastrophic performance loss rather than marginal degradation.

More critically, several experimental configurations reveal that RAG systems retrieving outdated information perform worse than LLM-only baselines. This counterintuitive result demonstrates that stale retrieval transforms from a neutral failure to an active harm mechanism. The generation component synthesizes retrieved context with parametric knowledge, and when retrieved context contradicts more

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

current information, this synthesis produces confidently incorrect outputs. Instead of gracefully falling down to the level of baseline LLM performance, the system makes mistakes by basing generation on outdated facts provided in the form of authoritative evidence [10].

These degradation effects are put in perspective by the CRAG benchmark in the larger performance environment of the RAG architectures. Subpar LLM baseline models have a maximum accuracy of 34 percent, generic RAG pipelines have 44 percent, and specialized industry systems have close to 63 percent. These measurements establish multiple reference points for understanding system capabilities. The gap between LLM-only performance at 34 percent and standard RAG at 44 percent quantifies the value of basic retrieval augmentation. The gap between standard RAG at 44 percent and optimized systems at 63 percent demonstrates the contribution of architectural sophistication, including query reformulation, reranking, and retrieval tuning.

That even state-of-the-art systems fail more than one-third of temporally sensitive queries indicates that retrieval errors, including both staleness and relevance failures, constitute the dominant bottleneck limiting RAG accuracy [1], [2]. This persistent failure rate exists despite substantial research investment in retrieval optimization and generation improvement, suggesting fundamental challenges in maintaining accurate retrieval of time-dependent information.

Streaming RAG evaluations contribute complementary evidence from continuous update scenarios. The indexing freshness improvement of up to 3 absolute points over the indexing baseline, and the improvement of up to 3 absolute points over the indexing baseline, is statistically significant [6]. This gain directly corresponds to elevated downstream accuracy in the answers due to the cause-and-effect relationship between retrieval quality and generation fidelity. Although this improvement is numerically small, next to the total performance envelope, this gain is a significant advance given the level of development of retrieval methodology and the challenge of making similar improvements in a variety of evaluation situations.

Together, these quantitative results empirically demonstrate that retrieval freshness is a dominant determinant of RAG accuracy and hallucination behavior. Degradation effects are large enough to eliminate or reverse the performance advantages of retrieval augmentation entirely under conditions of severe staleness. Organizations deploying RAG in production must therefore treat freshness management as a first-order design concern rather than a secondary optimization target.

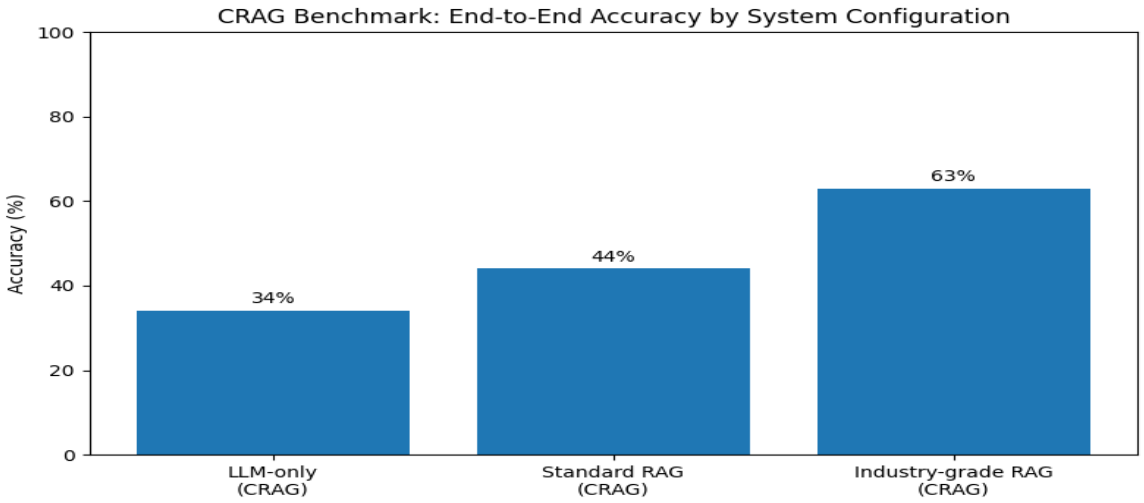

Fig. 2: CRAG Benchmark: End-to-End Accuracy by System Configuration [1, 2]

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conclusion

Retrieval-augmented generation systems operating over continuously evolving corpora face fundamental challenges in maintaining performance as temporal lag increases between source data updates and index regeneration. Empirical evidence establishes that outdated information causes substantial accuracy degradation and can transform retrieval from a performance enhancer into an active liability that misleads generation processes. The CRAG benchmark demonstrates persistent failure rates even in optimized architectures, while the HOH benchmark quantifies severe performance losses when stale documents enter retrieval corpora. Streaming architectures demonstrate the feasibility of continuous updates with production-grade operational characteristics, while hybrid retrieval designs provide architectural resilience through complementary semantic and lexical matching. These results define retrieval freshness as a first-order system property that needs specific management by domain-sensitive update policies that can trade off computation costs and accuracy demands. RAG architectures cannot be just deployed within an organization, expecting the value to be steady and requiring active monitoring of degradation and the application of the right refresh policies, as well as freshness service level goals to meet the requirements of the application. The presented benchmarks and quantitative data allow practitioners to move away from the reactive maintenance models towards the proactive freshness management based on the empirical knowledge of the degradation patterns. The domain-specific volatility profiles should be characterized by future advances, should develop adaptive refresh strategies based on observed degradation patterns, and should have cost-performance optimization frameworks that can be used to allocate resources efficiently to homogeneous knowledge bases with varying temporal characteristics.

References

[1] Xiao Yang et al., "CRAG – Comprehensive RAG Benchmark," 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf

[2] Xiao Yang et al., "CRAG – Comprehensive RAG Benchmark," arXiv:2406.04744v2 [cs.CL], 2024. https://arxiv.org/html/2406.04744v2

[3] Jie Ouyang et al., "HOH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation," Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. https://aclanthology.org/2025.acl-long.301.pdf

[4] Jie Ouyang et al., "HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation," arXiv:2503.04800, 2025. https://arxiv.org/abs/2503.04800

[5] Robert Friel et al., "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems," arXiv:2407.11005, 2025. https://arxiv.org/abs/2407.11005

[6] Yuzhou Zhu, "From Static to Dynamic: A Streaming RAG Approach to Real-time Knowledge Base," arXiv:2508.05662, 2025. https://arxiv.org/abs/2508.05662

[7] Rini Vasan, "How to evaluate RAG systems: metrics, frameworks & infrastructure," Redis Inc. 2026. https://redis.io/blog/rag-system-evaluation/

[8] Elena Samuylova, "A complete guide to RAG evaluation: metrics, testing, and best practices," Evidently AI, 2025. https://www.evidentlyai.com/llm-guide/rag-evaluation

[9] Karyna Naminas, "RAG Evaluation: Metrics and Benchmarks for Enterprise AI Systems," LabelYourData, 2025. https://labelyourdata.com/articles/llm-fine-tuning/rag-evaluation

[10] David Kirchhoff, "Metrics for Evaluation of Retrieval in Retrieval-Augmented Generation (RAG) Systems," DeconvoluteAI, 2025. https://deconvoluteai.com/blog/rag/metrics-retrieval