Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

A Comparative Evaluation of Dense, Sparse, and Hybrid Retrieval for Improving Retrieval-Augmented Generation Reliability

Aamir Shaikh, Aditya Kumar Pandey, Rohan Arjun Sherkar, Pragya Sinha
Department of Information Technology
Bharati Vidyapeeth (Deemed to be University) College of Engineering
Pune, Maharashtra, India

Abstract **-** Grounding large language model (LLM) outputs in external knowledge, and thereby curbing hallucination, is commonly achieved through Retrieval-Augmented Generation (RAG). Yet the retrieval stage of a RAG pipeline is itself often a weak link: dense retrievers capture semantic similarity but can miss exact lexical matches; sparse retrievers capture exact lexical matches but can miss semantic similarity; and hybrid retrievers combine both, though their effects on the reliability of generated outputs have received comparatively less evaluation. In this paper, we propose a comprehensive evaluation of dense, sparse and hybrid retrieval strategies within a unified retrieval and generation (RAG) pipeline, with a specific focus on how these retrieval methods affect generation reliability (faithfulness, hallucination rate, and answer relevance). We build a retriever-swappable RAG pipeline backed by a Llama-3-8B-Instruct generator and evaluate it on 300 held-out queries drawn from a technical support and product-documentation corpus. We compare four retrieval configurations—dense-only, sparse-only (BM25), static hybrid, and dynamic hybrid—using retrieval accuracy and generation-reliability metrics. Our results demonstrate that static hybrid retrieval achieves an 11-percentage-point increase in faithfulness over dense-only retrieval, and a 7-percentage-point decrease in the hallucination rate over sparse-only retrieval. An ablation study further examines the contributions of the dense and sparse components to the reliability improvement of the hybrid configuration. The results presented here indicate that the reliability gain from hybrid retrieval is driven mainly by a reduction in retrieval-failure hallucinations, while the remaining errors are increasingly attributable to the generator misusing correctly retrieved evidence, a limitation that retrieval strategy alone does not resolve. The findings provide practical insights for reliability-sensitive RAG deployments, while broader validation is needed across domains such as legal and medical question answering.

Keywords - Retrieval Augmented Generation; Dense Retrieval; Sparse Retrieval; Hybrid Retrieval; BM25; Semantic Search; Hallucination; Faithfulness; Large Language Models;

I. INTRODUCTION

Large language models (LLMs) are highly capable of producing fluent text and reasoning about natural-language problems, but their knowledge is limited by the information encoded in their model parameters during training. These limitations can lead to factually incorrect statements, which can be problematic when LLMs are applied in high-risk domains such as business, medicine, and law.

RAG addresses this limitation by pairing a non-parametric retrieval component with a parametric generator, so that model outputs draw not only on weights learned during training but also on evidence pulled in from an external source [1]. In a typical RAG setup, a query is first used to pull a set of relevant passages from a large document collection, and the generator then conditions its answer on that retrieved material. Because output quality depends so directly on what gets retrieved, the retrieval step itself is considered to be the key factor controlling a RAG system's performance.

Two broad forms of retrieval have historically been used for this stage. The first, dense retrieval, represents queries and documents as continuous vector embeddings and locates passages through nearest-neighbour search over that embedding space, which lets it capture richer relationships than just word similarity [2]. The second type is sparse retrieval, which represents queries and documents using weighted term vectors, as in the classic BM25 ranking function [3] and learned sparse models such as SPLADE [4], and retrieves passages through lexical matching. However, each retrieval paradigm has complementary limitations: dense retrieval may miss exact lexical matches, while sparse retrieval may fail to capture semantic similarity.

Hybrid retrieval emerged from these complementary limitations by combining the strengths of dense and sparse representations. Hybrid systems commonly use rank-based fusion such as Reciprocal Rank Fusion [14] or score-based weighted fusion [5] – [8]. Several empirical studies have reported that hybrid retrieval can outperform single-mode retrievers on standard information-retrieval measures such as Recall@k, MRR, and nDCG@k, and in some settings a well-tuned sparse model can even beat a commercial dense-embedding system outright [8].

Nonetheless, another line of work, largely orthogonal to retrieval-comparison studies, focuses on assessing the reliability of RAG systems themselves, particularly how faithfully generated answers reflect retrieved evidence. The line of research includes self-critiquing generation models [9], corrective retrieval strategies that detect and correct retrieval failures [10], reference-less automated frameworks for faithfulness and answer relevance assessment [11], hand-annotated corpora for assessing hallucination errors [12] and detailed diagnostic frameworks to verify whether generated claims are truly grounded in retrieved information for each reasoning facet [13]. Importantly, the reliability-focused studies considered in this review evaluate RAG systems under

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

a fixed retrieval context; the retrieval system is therefore not treated as an experimental variable.

This separation between optimization experiments (accuracy-based metrics) and reliability experiments (where retrieval is fixed) brings up an important question: Does the choice of retrieval paradigm (dense, sparse, or hybrid) affect the reliability of the generated output, and if so, how? This study addresses three important gaps identified in the existing literature, as described below.

- Gap 1 - Reliability Blindspot: retrieval-comparison papers benchmark dense, sparse, and hybrid retrieval systems only based on retrieval accuracy metrics, and reliability papers benchmark faithfulness and hallucination based only on one fixed retriever.
- Gap 2 - Component Attribution Gap: hybrid retrieval systems are generally benchmarked as a combined system, while limited work directly examines how the relative contributions of dense and sparse components affect reliability.
- Gap 3 - Causal Mechanism Gap: previous studies demonstrated that high retrieval quality does not necessarily lead to high faithfulness, because hallucinations can persist when correctly retrieved evidence is not faithfully used by the generator; however, this was only observed with one retrieval setup, and it remains unclear whether different retrieval setups would provide evidence that would be used differently by the generator.

A. Objectives and Contributions

Contributions based on the gaps highlighted include:

- An integrated, retriever-swappable RAG pipeline that works with dense-only, sparse-only, static hybrid, and dynamic hybrid retrieval.
- A joint evaluation methodology that presents results with retrieval accuracy metrics and RAG reliability metrics (faithfulness, hallucination rate, and answer relevance) for each retrieval mode using the same query set, which is a direct solution to Gap 1.
- An ablation study that examines the relative contribution of dense and sparse components to the reliability gains observed under hybrid retrieval, directly addressing Gap 2.
- An evidence-usage diagnostic methodology that looks into whether the hallucination difference between the different retrieval modes is due to the quality of retrieval itself or evidence misuse by the generator, which is a direct solution to Gap 3.
- A deployable interface that can help compare retrieval modes with the same query, as discussed in Section VI.

B. Paper Organization

The remainder of the paper is organized as follows. Section II reviews prior work on dense, sparse, and hybrid retrieval alongside RAG reliability evaluation and pinpoints the gap this study addresses. Section III describes the system architecture. Section IV lays out the experimental methodology, covering the datasets, retrieval setups, and evaluation metrics used. Section V reports the results. Section VI presents the user-interface design and deployment. Section VII discusses the findings and limitations. Section VIII concludes the paper, and Section IX outlines future work.

II. LITERATURE REVIEW

This review starts with a discussion on previous works based on two themes crucial to this paper's context: (i) retrieval methods in RAG, including dense, sparse, and hybrid retrieval, and (ii) evaluation of RAG reliability and hallucination. Thirteen research papers from 2009 to 2026, references [1]–[13], form the basis of this literature review, presented in narrative form before being categorized into subsections A–E; references [14]–[18] are supporting technical and implementation citations introduced later in the paper.

Lewis et al. introduced the core RAG architecture in 2020 [1], the same year Karpukhin et al. established dense retrieval through Dense Passage Retrieval (DPR) [2]. Sparse retrieval traces back further still, to Robertson and Zaragoza's 2009 probabilistic relevance framework [3], the origin of the BM25 ranking function still in use today; this classical lexical approach was later recast in 2021 as a learned, neural process by the SPLADE sparse lexical and expansion model [4].

Three notable hybrid approaches combine dense and sparse signals: Blended RAG (2024), Dynamic Alpha Tuning (2025), and the 2026 retrieval benchmark by Akarsu et al. [8]. Blended RAG pairs dense vector indexes with sparse encoder indexes and issues hybrid queries against both [5]. Dynamic Alpha Tuning, proposed by Hsu and Tzeng, instead computes a query-specific weighting between dense and sparse scores rather than relying on a fixed α coefficient [7]. The third, Akarsu et al.'s large-scale benchmark of retrieval strategies for text-and-table documents, compares ten retrieval methods spanning sparse, dense, hybrid, and corrective approaches [8]. That same year, Wang et al. separately compared dense and sparse retrieval performance in RAG pipelines across a range of computational budgets [6].

A distinct but related body of work looks at RAG reliability and hallucination rather than at retrieval comparison itself. Asai et al.'s Self-RAG (2023, presented at ICLR 2024) builds in self-reflection so the model critiques its own generations as it produces them [9]. Yan et al. followed in 2024 with Corrective Retrieval-Augmented Generation (CRAG), which tackles retrieval failure through a lightweight evaluation step [10]. Also in 2023 (published at EACL 2024), Es et al. put forward RAGAs, a framework providing automated metrics for assessing faithfulness, answer relevance, context precision, and context recall, reducing reliance on human-labeled evaluation for several RAG assessment tasks [11]. Niu et al.'s 2024 RAGTruth supplies a hand-annotated hallucination corpus spanning several LLMs under standard RAG conditions [12]. Most recently, Elchafei et al.'s 2026 facet-level tracing framework diagnoses retrieval-generation misalignment and shows that better retrieval does not guarantee better faithfulness [13].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

A. *Foundations of Retrieval-Augmented Generation*

Lewis et al.'s original RAG design coupled a pretrained dense retriever with a sequence-to-sequence generator and optimized both jointly for knowledge-intensive tasks, fixing retrieval, augmentation, and generation as the three stages that later work has built on [1]. That three-stage architecture underlies every system discussed in this review, including the pipeline proposed in Section III.

B. *Dense Retrieval*

In 2020, Karpukhin et al. presented Dense Passage Retrieval (DPR), demonstrating the feasibility of using a simple dual-encoder architecture with contrastive learning for open-domain question answering, which surpassed a BM25 baseline substantially, thereby demonstrating dense vectors' applicability, sometimes even superiority, compared with traditional lexical retrieval approaches [2]. Although subsequent research further improved the dual encoder architecture by employing more sophisticated techniques such as negative sampling, distillation, and better-pretrained encoders, the basic problem has remained the same: DPR models may perform poorly on rare terms, exact matches, and unknown entities in the query.

C. *Sparse and Learned-Sparse Retrieval*

BM25, which Robertson and Zaragoza introduced within the probabilistic relevance framework, remains the dominant classical sparse-retrieval method, scoring passages from term-frequency and inverse-document-frequency statistics with a length-normalization adjustment [3]. Its strength lies in matching exact lexemes, making it particularly effective for domain-specific terms, numeric values, and named entities. Formal et al. proposed the SPLADE method, which learns sparse lexical expansion weights using a BERT masked-language-model head with a sparsity regularizer, achieving much of the same performance in comparison to dense retrieval while maintaining the efficiency and transparency of inverted index structure [4].

D. *Hybrid Retrieval Strategies*

Blended RAG was introduced by Sawarkar et al., blending together dense and sparse indexes, and presented new benchmark results for NQ and TREC-COVID as well as increased accuracy of generative QA on SQuAD [5]. Wang et al. conducted a direct comparison between the algorithms used for sparse and dense retrieval, within a RAG framework, under different conditions of available resources, but without a hybrid algorithm condition in their comparisons [6]. This problem was pointed out by Hsu and Tzeng, who noted that the static weight hybrid fusion method does not have a way to adjust the weighting coefficients to match the changing requirements of each query and suggested Dynamic Alpha Tuning (DAT), where an LLM assigns scores to the top retrieval from each retriever for each query and computes the appropriate weight for the query [7]. Most recently, Akarsu et al. compared the performance of ten retrieval methods, including sparse and dense retrievals, hybrid fusion, reranking, and adaptive/corrective retrievals, on a large-scale financial text-and-table question answering task and found that BM25 retrieval performed better than even a commercial dense retrieval system on most metrics, while hybrid retrieval with cross-encoder reranking achieved much of the difference from the ideal oracle context ceiling [8].

E. *RAG Reliability and Hallucination Evaluation*

A separate strand of work tackles reliability directly. Asai et al.'s Self-RAG trains a model to judge for itself whether retrieval is even needed and to critique its own output through three reflection tokens — relevance, support, and usefulness — without a separate verifier [9]. Yan et al.'s CRAG instead takes corrective action once a retrieved document has been scored as correct, incorrect, or ambiguous [10]. Es et al.'s RAGAs framework provides automated metrics for faithfulness, answer relevance, context precision, and context recall, with the required inputs differing across metrics [11]. Niu et al.'s RAGTruth provides nearly 18,000 RAG-generated responses with manual hallucination annotations at both the case and word levels across multiple LLMs [12]. Most recently, Elchafei et al. used a facet-based diagnostic approach — decomposing questions into individual reasoning facets — to show that stronger retrieval does not automatically yield more faithful generation, since hallucination often stems from the generator departing from or overriding the evidence it was given [13]. Across all of these, the reliability tooling was built and validated against one specific, unchanging retrieval setup.

III. SYSTEM ARCHITECTURE

The proposed system is built around a retriever-swappable RAG pipeline that enables dense, sparse, and hybrid retrieval to be compared under otherwise identical generation and evaluation conditions. Rather than a diagram, the architecture is described in detail below, first through an end-to-end walkthrough and then through the individual components A–F.

The pipeline begins with a corpus of documents which undergo cleaning and chunking during the pre-processing phase. Each of these chunks is indexed in two ways – once using dense vector indexes derived from the semantic embeddings, and once using sparse lexical indexes based on the statistics of terms. During retrieval, the query is directed to one of four interchangeable retrieval modes – dense-only, sparse-only, static hybrid, and dynamic hybrid retrieval. For the hybrid configurations, the dense and sparse results are combined through the fusion stage. For dense-only and sparse-only configurations, the corresponding single-index results are passed directly to the generation stage. In all configurations, the final top-5 retrieved passages were supplied to the generator. The resulting top-5 passages are packed into a single prompt and handed to the generation stage, where the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

language model drafts an answer conditioned on that retrieved context. The draft is then scored for faithfulness, answer relevance, context precision/recall, and hallucination rate, and the answer, its supporting passages, and the reliability score are all surfaced to the user, as Section VI describes.

A. Document Corpus and Preprocessing

The corpus consists of 1,480 technical support articles and product-documentation pages drawn from a mid-size SaaS company's public knowledge base, totaling approximately 2.3 million tokens with an average document length of around 1,550 tokens. Documents in the dataset are divided into chunks of 300 tokens with an overlap of 50 tokens to preserve context across chunk boundaries. The pre-processing steps include HTML and markup stripping, de-duplication of near-identical articles, normalization of product names and error codes, and removal of boilerplate navigation and footer text.

B. Indexing Layer

Two indexes are created in parallel on the preprocessed corpus:

- Dense Index — each chunk is embedded with the all-mpnet-base-v2 model from Sentence Transformers [16], and the resulting vectors are stored in a FAISS index built on hierarchical navigable small world (HNSW) graphs [15] for fast approximate nearest-neighbour lookup.
- Sparse Index — chunks are indexed using BM25 (k1 = 1.2, b = 0.75) via an Elasticsearch index.

C. Retrieval and Fusion Layer

The system exposes four interchangeable retrieval configurations evaluated in this study: dense-only, sparse-only, static hybrid (fixed fusion coefficient $\alpha = 0.5$), and a dynamic, query-adaptive hybrid whose α is set per query as $\alpha = 0.3 + 0.4 \times (1 - \text{overlap\_ratio})$, clipped to [0.3, 0.7], where overlap_ratio is the fraction of query terms (after stopword removal) that appear verbatim in the corpus vocabulary. Queries with high lexical overlap — typically those containing exact product names or error codes — receive a lower α and are weighted toward the sparse index, while queries with little exact-term overlap, such as paraphrased or conceptual questions, receive a higher α and are weighted toward the dense index. Queries with fewer than four tokens after stopword removal default to $\alpha = 0.5$, since overlap_ratio is unstable at that length. Unlike DAT [7], which derives its weighting through LLM-based evaluation of the top retrieved results, this dynamic hybrid uses a lightweight, query-adaptive heuristic that requires no additional model calls. Hybrid fusion combines ranked lists from the dense and sparse indexes using a weighted linear combination of min-max normalized scores, with normalization performed independently for each query over the candidates retrieved by that index; a passage retrieved by only one index is assigned a score of 0 on the other index before the weighted sum is computed.

D. Generation Layer

Each of the top-5 passages is inserted into a fixed prompt template, numbered and listed beneath the user's question, and the whole prompt is sent to the generator (Llama-3-8B-Instruct [17], run locally through vLLM [18]) with temperature 0.2, top-p 0.9, and a 400-token cap on the response.

E. Reliability Evaluation Layer

Each generated answer is scored for faithfulness, answer relevance, and context precision/recall using the RAGAs definitions [11], and for hallucination rate using the annotation categories from RAGTruth [12]; the full procedure is given in Section IV-B.

F. User Interface / Deployment Layer

The system is exposed through a Streamlit dashboard, described further in Section VI.

IV. METHODOLOGY

A. Retrieval Configurations Compared

Four retrieval setups are considered in the entire experiment. The first setup is the dense-only setup, where retrieval is performed using only the dense vector index and the embedding model mentioned in Section III-B. The second setup is the sparse-only setup, where retrieval is performed using only the BM25 sparse index described in Section III-B. The third setup is the static hybrid setup, where dense and sparse scores are fused using a constant coefficient of $\alpha = 0.5$, giving equal weight to normalized dense and sparse scores. The fourth setup is the dynamic hybrid setup, in which α is adjusted per query between 0.3 and 0.7 as described in Section III-C.

B. Evaluation Metrics

In this study, retrieval accuracy refers to the quality of retrieved passages as measured by Recall@5, MRR, and nDCG@5, whereas generation reliability refers to faithfulness, answer relevance, the context-level metrics, and hallucination rate. Both sets of metrics are computed for every retrieval system under test on the identical query set, directly addressing Gap 1. Retrieval accuracy is captured by three measures: Recall@k (k = 5), the share of queries for which the ground-truth relevant passage appears among the top-k results; Mean Reciprocal Rank (MRR), the average reciprocal rank of the first relevant hit; and nDCG@k, a ranking-sensitive measure of the position of the relevant result within the retrieved ranking. Retrieval relevance judgments: each evaluation query was associated with its corresponding ground-truth source document or passage from the evaluation corpus, and a retrieved passage was considered relevant according to this predefined query-to-source mapping. These relevance judgments were used consistently to calculate Recall@5, MRR, and nDCG@5 across all retrieval configurations. Reliability is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

measured using faithfulness, defined as the extent to which the retrieved context supports the generated claims, and answer relevance, defined as how closely the generated answer addresses the original query, together with context precision and context recall — respectively, the share of retrieved passages that are relevant and the share of relevant passages that were retrieved — as defined by the RAGAs framework [11]. Context recall was computed using the reference answer associated with each evaluation query, following the RAGAs formulation [11]. Hallucination rate, the share of answers containing unsupported or contradicted claims, is measured using the annotation categories from RAGTruth [12].

C. Experimental Protocol

All four retrieval configurations were evaluated once each over the same fixed set of 300 held-out queries, so that every configuration is compared against an identical query set as required by Gap 1. The 300 evaluation queries were held out from the development process and were not used for retriever configuration or parameter selection. The generation temperature was fixed at 0.2 across all configurations to maintain identical generation settings and improve comparability between configurations. Statistical significance between configurations was assessed using paired bootstrap resampling with 10,000 resamples over the per-query metric differences, following the practice used by Akarsu et al. [8]. Statistical significance was assessed using a significance level of 0.05.

D. Ablation Design (Gap 2)

To examine the relative influence of the dense and sparse components within the hybrid fusion, a sweep was performed on the fusion weight parameter for the static hybrid fusion, evaluating α values of {0.0, 0.25, 0.5, 0.75, 1.0}, with retrieval-accuracy and reliability measurements taken at each value. $\alpha = 0$ corresponds to sparse-only and $\alpha = 1$ to dense-only, so the sweep also nests both single-mode baselines within the same analysis.

E. Evidence-Usage Diagnostic (Gap 3)

To investigate whether hallucinations stem from poor retrieval or from misapplication of correctly retrieved evidence, a randomly selected subset of 150 held-out queries was evaluated across all four retrieval configurations, producing 600 generated answers for the evidence-usage diagnostic; each hallucination was manually classified as either a "retrieval failure," where no retrieved passage supported the claim, or "evidence mismatching," where a supporting passage was present among the top-5 results, but the generator failed to use it faithfully. This classification, reported in Section V-C, was applied uniformly across all four retrieval configurations to support comparison of how the source of unreliability shifts across retrieval configurations.

V. RESULTS

A. Retrieval Accuracy Results

TABLE I. RETRIEVAL ACCURACY RESULTS

Configuration	Recall@5	MRR	nDCG@5
Dense-only	0.71	0.58	0.63
Sparse-only	0.68	0.55	0.60
Static Hybrid	0.79	0.66	0.70
Dynamic Hybrid	0.81	0.68	0.72

Table II reports the corresponding reliability metrics for the same four configurations, following the RAGAs definitions given in Section IV-B.

TABLE II. RELIABILITY RESULTS

Configuration	Faithfulness	Answer Relevance	Context Precision	Context Recall	Hallucination
Dense-only	0.74	0.88	0.69	0.72	18%
Sparse-only	0.77	0.87	0.71	0.68	16%
Static Hybrid	0.85	0.90	0.80	0.79	9%
Dynamic Hybrid	0.86	0.90	0.82	0.81	8%

B. Ablation Results

Following the ablation design described in Section IV-D, we swept the static fusion weight α over {0.0, 0.25, 0.5, 0.75, 1.0}, where $\alpha = 0$ reduces to sparse-only and $\alpha = 1$ reduces to dense-only. Faithfulness rose from 0.77 at $\alpha = 0$ to a peak of 0.85 at $\alpha = 0.5$–0.75 before dropping back to 0.74 at $\alpha = 1$, while the hallucination rate followed the inverse pattern: 16% at $\alpha = 0$ (sparse-only), 18% at $\alpha = 1$ (dense-only), and bottoming out at 9% in the same 0.5–0.75 range, a 7-percentage-point reduction from the sparse-only rate and a 9-percentage-point reduction from the dense-only rate. The ablation results suggest that incorporating the sparse component contributed substantially to the observed reliability improvement. In particular, the highest faithfulness and lowest hallucination rates occurred at intermediate fusion weights, indicating that combining lexical and semantic retrieval signals was more reliable than either single-mode configuration on this dataset.

C. Evidence-Usage Diagnostic Results (Gap 3)

Following the diagnostic method described in Section IV-E, manually classifying the hallucinations from the 150-query, 600-answer annotated subset into "retrieval failure" (the supporting evidence was never retrieved) versus "evidence mismatching" (the evidence was retrieved but the generator ignored or contradicted it) revealed a clear shift across configurations. Within the annotated hallucination cases,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

retrieval failures accounted for 62% of dense-only hallucinations, consistent with its weaker performance on exact-match queries, compared with 24% of the pooled hallucinations from the static and dynamic hybrid configurations. The hybrid configurations produced a substantially smaller overall set of hallucination cases, among which evidence mismatching accounted for the remaining 76%. This suggests that once retrieval quality is improved through hybrid fusion, the dominant remaining bottleneck for reliability shifts from what is retrieved to how the generator uses what it is given, a limitation that retrieval-side improvements alone cannot fully address.

D. Statistical Significance

Using the paired bootstrap procedure described in Section IV-C, the faithfulness difference between static hybrid and dense-only retrieval was statistically significant ($p < 0.01$) over 10,000 resamples, as was the difference between static hybrid and sparse-only retrieval ($p < 0.01$). The difference between static hybrid and dynamic hybrid faithfulness, however, was not statistically significant ($p = 0.18$), indicating that in this dataset the query-adaptive α weighting produced a numerically higher faithfulness score, but the difference from fixed-weight hybrid retrieval was not statistically significant ($p = 0.18$).

VI. USER INTERFACE AND DEPLOYMENT

The system is exposed to end users through a Streamlit web application. The interface enables an end user to provide a natural language query along with the selection of a retrieval method (dense, sparse, static hybrid, or dynamic hybrid) and view the following information side by side: (i) top-5 passages retrieved using the selected method, (ii) generated answer and (iii) reliability scores (faithfulness, relevance, hallucination) for the answer.

A. Deployment Environment

The system was containerized using Docker and deployed on a single cloud virtual machine with one NVIDIA A10 GPU used for the generation model; the dense and sparse indexes are small enough at this corpus size to serve retrieval from CPU alone. Measured from the moment a query is submitted to the point a complete answer is returned, average latency came to 2.1 seconds for dense-only, 1.4 seconds for sparse-only, and 2.6 seconds for either hybrid configuration — the extra time in the hybrid case coming from having to query and fuse two indexes per request instead of just one.

B. Usability Considerations

The retrieval mode is presented to the user as a simple dropdown selector above the query box, with the selected mode's Recall@5, faithfulness, and hallucination rate shown as a small reference panel so users can weigh accuracy against reliability before reading the answer. Generated answers and retrieved passages are cached by a hash of the query text and retrieval mode, which reduces repeated latency for common support questions to under 200 milliseconds. Accessibility was addressed through full keyboard navigation, screen-reader labels on all interactive elements, and a high-contrast display option for the retrieved-passage panel.

VII. DISCUSSION

A. Interpretation Relative to the Identified Gaps

The joint retrieval-accuracy and reliability evaluation in Section V challenges the assumption, implicit in much of the literature discussed in Section II-D, that better retrieval accuracy automatically translates into better generation reliability. Static and dynamic hybrid retrieval improved Recall@5 by 8 and 10 percentage points over dense-only retrieval, respectively, and by 11 and 13 points over sparse-only retrieval; however, the increase in retrieval accuracy was not accompanied by a proportional improvement in faithfulness, particularly between the static and dynamic hybrid configurations, which differed by only two percentage points on each reported retrieval metric, while their faithfulness difference was not statistically significant. These results provide evidence in support of Gap 1 within the evaluated dataset: retrieval-accuracy benchmarks and reliability benchmarks do not necessarily move in lockstep, and a system optimized purely for Recall@k or nDCG@k is not guaranteed to be optimized for faithfulness.

The ablation study conducted in Section V-B addresses Gap 2: the results suggest that incorporating the sparse component contributed substantially to the hybrid configuration's reliability gain, as the highest faithfulness and lowest hallucination rates occurred at intermediate fusion weights rather than at either single-mode extreme. Notably, this component-level effect on reliability was more pronounced than the corresponding effect on retrieval-accuracy metrics, where the dense and sparse components contributed more evenly, reinforcing that accuracy and reliability respond differently to the same underlying retrieval components.

The evidence-usage diagnostic in Section V-C speaks directly to Gap 3. Once hybrid retrieval closed most of the retrieval-failure gap, the majority of remaining hallucinations were evidence mismatching cases, where correct evidence was retrieved but not faithfully used by the generator. This mirrors the retrieval-generation misalignment described by Elchafei et al. [13]. Within the manually annotated subset, the results suggest that further reliability gains may increasingly depend on generator-side behavior rather than retrieval strategy alone

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

— an argument for pairing hybrid retrieval with generator-side faithfulness interventions in future work.

B. Comparison with Prior Literature

The accuracy-only results in Table I agree directionally with Blended RAG [5], Wang et al.'s comparison [6], and DAT [7]: hybrid retrieval outperformed either single-mode retriever on Recall@5, MRR, and nDCG@5 in this corpus as well. Our results diverge from these accuracy-only studies, and to some extent from Akarsu et al.'s benchmarking [8], on the question of reliability: while DAT motivates dynamic α weighting as a meaningful improvement over a static coefficient, our results found no statistically significant reliability difference between the static and dynamic hybrid on this dataset, suggesting that the benefit of dynamic weighting may be more dataset- or query-distribution-dependent than previously reported. More importantly, none of the accuracy-only benchmarks surveyed in Section II report reliability metrics alongside accuracy, so the finding that higher retrieval accuracy does not necessarily produce a statistically significant improvement in generation reliability highlights the importance of evaluating both retrieval and generation reliability jointly — and is, to the best of our knowledge, a joint comparison of retrieval accuracy and generation reliability across dense, sparse, and hybrid configurations that the studies reviewed here do not make.

C. Limitations and Threats to Validity

- Domain/Dataset Scope: The results were obtained from one dataset/domain (SaaS technical support and product documentation) and cannot necessarily be generalized to other domains, given the domain-specific limitation mentioned in the reviewed literature [5]–[8].
- LLM-as-Judge Reliability: Metrics derived from LLM-judge grading (based on [11]) might be prone to being influenced by prompt formulation and the judge language model itself; a human-annotated validation of these automated reliability scores, as recommended in Section II-E, was not performed in this work and is left for future validation. (The manual annotation conducted for the evidence-usage diagnostic in Section IV-E was limited to classifying hallucinations as retrieval failures or evidence mismatches, and did not validate the RAGAs/LLM-judge scores themselves.)
- Retrieval Setup Scope: This paper considers four retrieval setups; other approaches such as learned-sparse retrieval (SPLADE [4]) and cross-encoder reranking were not included and constitute potential avenues for future research.
- Scale: The query set of the evaluation includes 300 queries; this is smaller than the large-scale benchmarks used by Akarsu et al. [8], and results may not fully hold at larger query volumes or across a wider range of query types.
- Single-Generator Evaluation: All configurations were evaluated with a single generation model (Llama-3-8B-Instruct); the reliability patterns observed here, particularly the balance between retrieval-failure and evidence mismatching hallucinations, may differ for larger or differently-tuned generators.
- Single-Run Generation: Because each configuration was evaluated in a single generation run, the paired bootstrap procedure in Section V-D quantifies uncertainty across queries but does not capture run-to-run stochastic variation in generation; repeated-generation variance was not separately quantified in this work.

VIII. CONCLUSION

This paper conducted a comparative study of dense, sparse, and hybrid approaches to retrieval for Retrieval-Augmented Generation, emphasizing generative reliability alongside retrieval accuracy. Guided by three gaps identified in the literature — Gap 1, concerning the separation between retrieval comparison and generation-reliability assessment; Gap 2, concerning component-level analysis of hybrid retrieval; and Gap 3, concerning the influence of retrieval type on evidence use during generation — we developed a retriever-swappable RAG pipeline and evaluated it using both information-retrieval and RAG reliability metrics.

Concretely, static hybrid retrieval reached 0.85 faithfulness and a 9% hallucination rate, compared with 0.74 and 18% for dense-only and 0.77 and 16% for sparse-only. Dynamic hybrid had slightly higher retrieval accuracy than static hybrid, but the corresponding difference in faithfulness was not statistically significant ($p = 0.18$). Together, these results indicate that the reliability improvement observed with hybrid retrieval was not explained solely by the corresponding gains in retrieval accuracy, and that within the manually annotated subset, most residual hallucination cases after hybrid retrieval were classified as evidence mismatching errors, suggesting that generation-side behavior becomes an important remaining reliability bottleneck rather than retrieval quality alone.

IX. FUTURE WORK

- Extend the comparison to additional domains and corpora to test the generalizability of the findings beyond SaaS technical support and product documentation.
- Include sparse retrieval techniques like SPLADE [4] along with classical BM25 as an additional configuration.
- Evaluate reliability using multiple LLMs from different model families and generations to determine whether the relationship between retrieval and reliability depends on the generator.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

- Scale up the evidence-usage diagnostic (Section IV-E) to a complete facet-level analysis as proposed in [13] across all retrieval configurations, rather than being limited to the current subset.
- Include human-based evaluation of faithfulness and hallucination along with LLM-judge-based evaluation to validate the reliability scores obtained in this work.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems 33 (NeurIPS), 2020, pp. 9459–9474.

[2] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.

[3] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[4] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse lexical and expansion model for first stage ranking," in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2021, pp. 2288–2292.

[5] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG: Improving RAG (Retriever-Augmented Generation) accuracy with semantic search and hybrid query-based retrievers," in Proc. IEEE 7th Int. Conf. Multimedia Information Processing and Retrieval (MIPR), 2024, pp. 155–161.

[6] Y. Wang et al., "Evaluating sparse and dense retrieval in retrieval-augmented generation systems: A study," in Proc. 2024 10th Int. Conf. Communication and Information Processing (ICCIP), 2024. doi: 10.1145/3708657.3708747.

[7] H.-L. Hsu and J. Tzeng, "DAT: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation," arXiv preprint arXiv:2503.23013, 2025.

[8] M. Akarsu, R. K. Karaman, and C. Mierbach, "From BM25 to corrective RAG: Benchmarking retrieval strategies for text-and-table documents," arXiv preprint arXiv:2604.01733, 2026.

[9] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[10] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," arXiv preprint arXiv:2401.15884, 2024.

[11] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL), 2024, pp. 150–158.

[12] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in Proc. 62nd Annu. Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 10862–10878.

[13] P. Elchafei, M. Swain, S. Masoudian, and M. Schedl, "Facet-level tracing of evidence uncertainty and hallucination in RAG," arXiv preprint arXiv:2604.09174, 2026.

[14] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2009, pp. 758–759.

[15] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, 2020.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[17] A. Grattafiori, A. Dubey, A. Jauhri et al., "The Llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[18] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in Proc. 29th ACM Symp. Operating Systems Principles (SOSP), 2023, pp. 611–626.