Disentangling Faithfulness Hallucinations in Retrieval-Augmented Generation: A Systematic Benchmark and Analysis

Chandana Sree Mala[1,2*], Gizem Gezici[2] and Fosca Giannotti[2]

[1]University of Pisa, Pisa, Italy.
[2]Scuola Normale Superiore, Pisa, Italy.

*Corresponding author(s). E-mail(s): c.mala@studenti.unipi.it;
Contributing authors: gizem.gezici@sns.it; fosca.giannotti@sns.it;

Abstract

Faithfulness hallucinations, where large language models generate outputs unsupported by retrieved evidence, remain a central challenge for trustworthy AI. We present a systematic empirical evaluation of faithfulness in retrieval-augmented generation (RAG) systems using two benchmark datasets, HotpotQA and HaluBench, covering both multi-hop reasoning and single-hop hallucination detection. We analyze three small-to-mid-sized (2B–8B) open-weight LLMs in combination with multiple retrieval strategies, including sparse, dense, and hybrid approaches, as well as score-based and rank-based fusion techniques, enabling a comprehensive assessment of retrieval–generation interactions. By disentangling retrieval and generation errors, we characterize how different pipeline components contribute to hallucinations in RAG systems. Our analysis provides actionable insights and practical evaluation protocols, highlighting the critical role of robust retrieval and careful system design. These findings offer a benchmarking-oriented perspective for developing more reliable and faithful RAG systems within evaluated model scales.

Keywords: Retrieval Augmented Generation, Large Language Models, Faithfulness Hallucinations, Hallucination Mitigation, Hybrid Retrieval, Knowledge Grounding, Multi-hop Reasoning

1 Introduction

Large Language Models (LLMs) have transformed generative AI and NLP, yet their propensity to produce fluent but factually incorrect outputs so-called hallucinations remains a

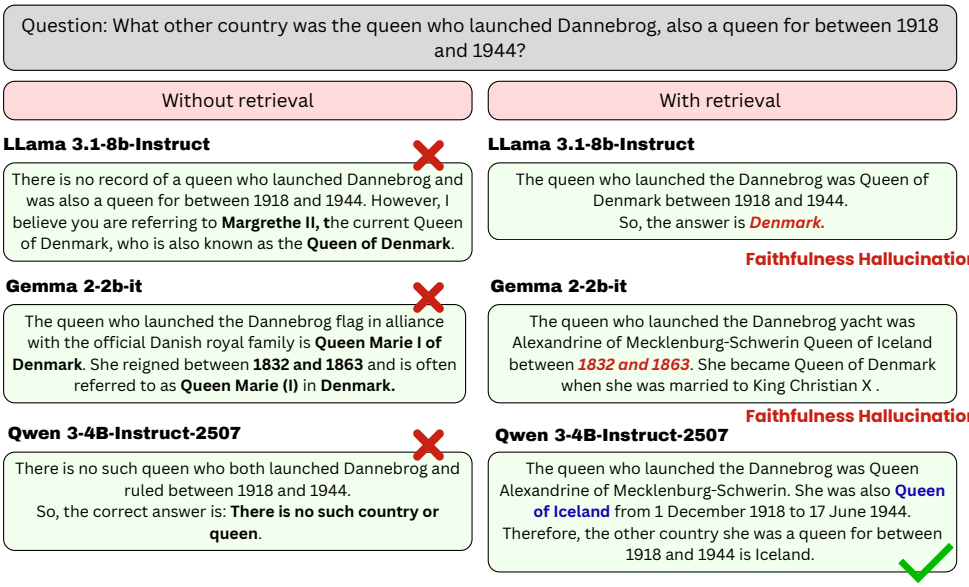

Fig. 1: One example from the HALUBENCH dataset illustrating LLM answer generation *with* and *without* retrieval (using the hybrid retrieval). The example demonstrates how grounding generation in retrieved evidence can reduce hallucinations, underscoring the importance of retrieval-augmented generation for producing faithful responses.

key barrier to reliability and real-world deployment. Hallucinations are typically categorized as *intrinsic* (also known as *faithfulness hallucinations*), when outputs deviate from the provided context, or *extrinsic*, when they contradict established facts. In this study, we particularly focus on intrinsic, or in other words *faithfulness hallucinations*.

RAG has emerged as a promising framework to mitigate hallucinations by grounding LLM outputs in external, verifiable sources. RAG systems combine a retrieval module which fetches relevant documents from knowledge bases with a generation module that conditions its output on both the user query and the retrieved evidence. While RAG is designed to improve factual alignment, *faithfulness hallucinations* still occur, as shown in Figure 1, revealing limitations in how models interpret and reason over retrieved content (Islam et al. 2024). RAG blends the encyclopedic memory of a search engine with generative models and consists of two main modules: the retrieval phase and the generation phase. Unlike traditional models that depend solely on internal knowledge, RAG enhances the generation process by incorporating relevant external documents retrieved from knowledge sources such as databases, search engines, or vector stores. RAG systems typically adopt a retrieve-then-read architecture, wherein a retriever first identifies relevant content, and a generator subsequently produces a response conditioned on both the user query and the retrieved documents (Guu et al. 2020; Lewis et al. 2020; Shuster et al. 2021; Karpukhin et al. 2020). Despite significant progress in retrieval and prompting strategies for RAG, the nuanced interactions between retrieval quality, fusion techniques, and generation remain insufficiently explored in current

research. Recent studies highlight that RAG systems, while designed to improve factual alignment, still struggle with faithfulness hallucinations, especially when the retrieved context conflicts with the model's internal knowledge or when fusion methods inadequately combine evidence sources. Existing approaches often focus on enforcing strict context adherence or modifying decoding strategies, but these can suppress the model's parametric knowledge and introduce new risks of misinterpretation (Zhang et al. 2025a; Gao et al. 2025; Wallat et al. 2025). While hallucinations in RAG systems primarily originate from the generation component, they are frequently amplified by retrieval errors that supply incomplete or misleading context. Empirical evidence from recent benchmarks including HotpotQA (Yang et al. 2018), HaluBench (Ravi et al. 2024), and MEGA-RAG (Xu et al. 2025) demonstrates that grounding generation on diverse and accurate retrieved evidence effectively constrains the generation space and reduces hallucinations. Moreover, prior work has identified key factors that degrade RAG performance and directly contribute to hallucinated outputs, such as insufficient contextual coverage and missing top-ranked documents. These findings underscore the importance of robust and contextually rich retrieval strategies for faithfulness in RAG-based LLMs (Barnett et al. 2024). Our study is motivated by persistent gaps in the literature regarding the interplay between retrieval quality, fusion techniques, and generation strategies in RAG systems.

Our main contributions are as follows.

- We conduct a systematic evaluation of faithfulness hallucinations in RAG-based LLMs, leveraging two comprehensive benchmark datasets HotpotQA and HaluBench to cover both multi-hop reasoning and single-hop hallucination detection tasks.
- We present a comprehensive study of three small to mid sized (2B–8B) open weight LLMs, five retrieval models spanning sparse and dense paradigms with hybrid variants, and three fusion techniques, enabling a systematic assessment of interactions between retrieval and generation components in RAG systems.
- We provide actionable insights and detailed evaluation protocols for future RAG system development, emphasizing the importance of disentangling retrieval and generation errors to advance trustworthy and faithful LLM applications within evaluated model scales.

The paper is structured as follows. Section 2 reviews related work. Section 3 presents the LLM-driven RAG architecture, detailing the pre-retrieval, retrieval, post-retrieval, and generation phases. Section 4 describes the experimental setup and presents the results. Section 5 discusses the limitations of this work. Finally, Section 6 concludes the paper and outlines potential future directions.

2 Related Work

Recent surveys (Huang et al. 2025b; Ji et al. 2023; Zhang et al. 2025b) offer comprehensive taxonomies and emphasize the persistent challenge of hallucinations in knowledge-intensive tasks, detailing definitions, causes, evaluation protocols, and mitigation strategies, with particular attention to faithfulness errors and the role of RAG systems. Hallucinations originate primarily from the generation component, yet they are often amplified by retrieval errors that

provide incomplete or misleading context (Alansari and Luqman 2026). While only a few studies (Hu et al. 2025chan; Lewis et al. 2020; Gumaan 2025) have explored the theoretical relationship between retrieval quality and hallucination mitigation, recent benchmarks offer strong empirical evidence that grounding generation on diverse and accurate retrieved evidence helps constrain the generation space and thereby reduces hallucinations (Xu et al. 2025). By decoupling external knowledge from the LLM and grounding its outputs in up-to-date, reliable information, RAG improves factual grounding and mitigates unsupported generations. This approach provides LLMs with access to current facts and enables responses to be attributed to verifiable sources, thereby reducing hallucinations and enhancing user trust in the generated outputs (Shuster et al. 2021; Ayala and Bechard 2024).

Prompting strategies especially Chain-of-Thought (CoT) prompting (Wei et al. 2022) have been shown to enhance reasoning and factual accuracy in LLMs, and are increasingly explored for hallucination mitigation in RAG systems. However, their integration with retrieval-augmented architectures remains an active research area, with recent work highlighting both the benefits and trade-offs of CoT for reducing hallucinations and for detection accuracy. In parallel, recent advances in fact-level conflict modeling and context-faithful retrieval have emerged as essential for building reliable RAG systems, as demonstrated by new frameworks that explicitly address knowledge conflicts and context alignment in generation (Kumar et al. 2024; Zhang et al. 2025a; Gao et al. 2025; Wallat et al. 2025).

Additionally, recent decoding methods (Chuang et al. 2024; Gema et al. 2024; Shi et al. 2024; Chen et al. 2024b) offer complementary mechanisms to RAG architectures for hallucination mitigation. These approaches enhance factuality and reasoning consistency at the decoding stage, providing synergy with retrieval-based contextual grounding (Xu et al. 2025), and further highlighting how retrieval quality and evidence consistency influence hallucination rates (Xiong et al. 2026). This synergy aligns with prior work on interleaving retrieval and reasoning (Trivedi et al. 2023), which dynamically integrates contextual evidence into intermediate reasoning steps rather than treating retrieval and generation as separate phases. Complementary findings in (Liu et al. 2024) reveal the challenges of maintaining factual focus in long-context reasoning, where models often ignore mid-context evidence, leading to partial hallucinations or fragmented context usage. Recent years have seen a surge of interest in faithfulness hallucinations, with several studies proposing new metrics, benchmarks, and modeling approaches for their detection and mitigation (Zhang et al. 2025a; Fadeeva et al. 2026; Tian et al. 2025). In parallel, recent work argues for improving reasoning faithfulness alignment by leveraging feedback signals from retrieval within RAG pipelines (Rathee et al. 2026). Together, these works deepen our understanding of how LLMs produce *faithfulness hallucinations* and establish more rigorous protocols for evaluating and improving faithfulness in generative models.

Our work builds on these insights by systematically evaluating the interplay between retrieval quality, fusion techniques, and generation strategies in RAG-based LLMs, using both multi-hop and single-hop hallucination benchmarks. This approach provides actionable insights and evaluation protocols for advancing faithful RAG applications.

3 Methodology

As introduced in Section 1, RAG frameworks aim to extend the capabilities of LLMs by grounding generated outputs in external, verifiable knowledge, thereby mitigating reliance on solely pre-trained knowledge. These frameworks incorporate curated knowledge bases to provide context during response generation. In this section, we detail the architecture of the proposed LLM-driven RAG system, which consists of two principal components: a retrieval module and a generation module. The end-to-end retrieval process unfolds across three distinct phases: pre-retrieval, retrieval, and post-retrieval.

3.1 Pre-retrieval Phase

In the *pre-retrieval phase*, the RAG framework processes a document collection denoted as $d_i \in D$, where $i = \{1, \ldots, n\}$. Each document d_i is encoded into a high-dimensional vector representation via an embedding function $\texttt{Enc}$, yielding:

$$\mathbf{d}_i = \texttt{Enc}(d_i)$$

The resulting embeddings $\{\mathbf{d}_i\}$ are stored in an indexed database *DB*, which serves as the external knowledge base. This database is organized and optimized for efficient vector similarity search and encapsulates domain-specific information relevant to anticipated user queries in the context of the target application.

3.2 Retrieval Phase

During the retrieval phase, we initially explored a query expansion (*QE*) module to improve coverage across heterogeneous document vocabularies. We implemented this using WordNet (Miller 1995), incorporating semantic relations such as synonyms, antonyms, and hypernyms to mitigate lexical mismatch. However, this approach yielded only marginal gains and occasionally reduced precision, due to noisy expansions from polysemy, weak semantic relations, and missing domain-specific terms. We also observed that lexical expansion disproportionately benefits sparse retrievers by increasing surface-form overlap, without consistent improvements for dense or hybrid methods. We therefore removed *QE* as a standalone module in the final framework. Instead, we adopt more controlled alternatives: retrieval methods that encode lexical and semantic variation in their representations, as well as adaptive retrieval strategies that incorporate query complexity information (Section 3.3.1). While *QE* may still be beneficial in specialized domains, it was not effective in our setting (Azad and Deepak 2019).

Retrieval Strategy

Retrieval methods can be broadly divided into two paradigms: *sparse* and *dense* retrievers (Lin et al. 2021). *Sparse retrievers* (Ret_S) represent queries and documents as high-dimensional, vocabulary-aligned vectors derived from discrete lexical tokens. Classical examples (e.g. BM25) (Robertson et al. 2009), estimate relevance via exact term matching on these sparse representations. While efficient and highly interpretable, traditional sparse models suffer from the well-known *lexical mismatch problem*, failing to capture synonymy,

polysemy, or implicit semantic relations. Recent *learned sparse retrievers*, most notably SPLADE (Formal et al. 2021), mitigate this limitation by leveraging contextualized language models to expand and re-weight terms in a differentiable manner, effectively bridging the *lexical–semantic gap* while retaining the scalability and interpretability advantages of sparse retrieval.

In contrast, dense retrievers (Karpukhin et al. 2020; Izacard et al. 2022), denoted as Ret_D, encode both queries and documents into continuous-valued representations within a shared embedding space. Contemporary variants such as MPNet- and Contriever-based models (Song et al. 2020; Izacard et al. 2022) employ transformer encoders to generate semantically rich embeddings that facilitate matching across lexically diverse yet semantically aligned inputs. This approach has shown improved recall and generalization, particularly in zero-shot and domain-transfer retrieval settings.

For the *sparse retrievers* (Ret_S), we use BM25 and SPLADE[1], and for the *dense retrievers* (Ret_D), we employ MPNet, MPNet-QA, and Contriever[2]. SPLADE is a sparse lexical and expansion retriever fine-tuned on the MS MARCO dataset for passage retrieval. MPNET is a general-purpose sentence embedding model fine-tuned using a contrastive learning objective on large-scale NLI and paraphrase data, but not directly on QA datasets. MPNET-QA is a QA-specialized variant fine-tuned on question, answer pairs from diverse sources including MS MARCO (Li et al. 2023). CONTRIEVER is an unsupervised dense retriever pre-trained via self-supervised contrastive learning on text pairs. CONTRIEVER, an unsupervised dense retriever trained via contrastive learning on unlabeled text corpora (Izacard et al. 2022), enables strong domain-agnostic generalization compared to supervised models such as DPR (Karpukhin et al. 2020) and COLBERT (Khattab and Zaharia 2020). MPNET serves as a general-purpose semantic baseline without retrieval-specific fine-tuning, whereas MPNET-QA represents its question-answering-oriented variant fine-tuned on QA-style pairs from diverse sources including MS MARCO. Together, these retrievers provide a balanced and comprehensive evaluation of retrieval–generation performance across multi-hop reasoning and domain-transfer settings.

3.3 Post-retrieval Phase

In the post-retrieval phase, the system employs a *hybrid retriever* and a *reranking module* to refine candidate selection. The hybrid retriever fuses results from sparse (Ret_S) and dense (Ret_D) retrieval models, combining their complementary lexical and semantic signals to produce an initial relevance ranking. Building upon this foundation, the reranking module further refines the retrieved candidates using fine-grained contextual features to better capture subtle semantic relationships. This two-stage refinement improves retrieval precision and helps reduce downstream issues such as semantic drift or hallucination during response generation.

3.3.1 Hybrid Retriever

Rather than committing to a single retrieval paradigm sparse (Ret_S) or dense (Ret_D) the system employs a *hybrid retriever* that integrates both retrieval streams within the post-retrieval stage. As shown in Figure 2, the hybrid retriever combines the outputs of the best-performing

[1]Sparse model identifiers: `pyserini-bm25`, `naver/splade-cocondenser-ensembledistil`.

[2]Dense model identifiers: `sentence-transformers/all-mpnet-base-v2`, `sentence-transformers/multi-qa-mpnet-base-dot-v1`, `facebook/contriever`.

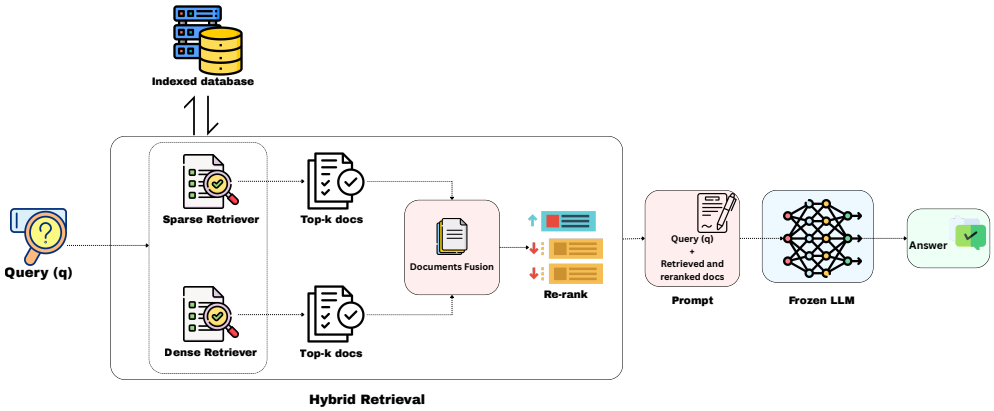

Fig. 2: Overview of the proposed pipeline combining hybrid retrieval, reranking, and frozen-LLM generation.

sparse and dense retrievers through several fusion strategies. These complementary fusion mechanisms balance lexical precision and semantic generalization by adaptively weighting or reordering signals from heterogeneous retrieval sources. This design enables the retrieval process to dynamically align with query-specific characteristics, achieving a more effective balance between recall and precision across diverse information-seeking scenarios.

We apply both *score-based* and *rank-based* fusion techniques, namely *linear interpolation* (Vogt and Cottrell 1999) and *CombMNZ* (Fox and Shaw 1994) for the former, and *vanilla Reciprocal Rank Fusion (RRF)* and its *query-specific weighting variants* for the latter (França et al. 2025; Spärck Jones 2004), combining evidence from multiple retrievers to enhance retrieval robustness and cross-system agreement.

Score-based fusion

Before applying any fusion, we normalize the document scores produced by each retriever to ensure comparability across different scoring scales (Manmatha and Sever 2002). Given a raw score $s_i(d)$, assigned to document d, by retriever i, its normalized form $\tilde{s}_i(d)$ is computed using min–max scaling:

$$\tilde{s}_i(d) = \frac{s_i(d) - \min(s_i)}{\max(s_i) - \min(s_i)} \tag{1}$$

where the minimum and maximum scores returned by retriever i are denoted by $\min(s_i)$ and $\max(s_i)$, respectively. This normalization ensures that all retrievers contribute proportionately within the same range $[0, 1]$.

We employ *linear interpolation* with static weights (Vogt and Cottrell 1999), where the final fused score for a document d is computed as:

$$\text{Score}_{\text{LI}}(d) = \sum_{i=1}^{n} w_i \tilde{s}_i(d) \tag{2}$$

where the normalized relevance scores assigned to document d by the sparse and dense retrievers are denoted by $\tilde{s}_{\text{Ret}_S}(d)$ and $\tilde{s}_{\text{Ret}_D}(d)$, respectively. The dense retriever computes its similarity-based score as:

$$s_{\text{Ret}_D}(d) = \text{sim}(q,d) \tag{3}$$

where $\text{sim}(q,d)$ typically represents either a dot product or cosine similarity between the query and document embeddings.

Additionally, we incorporate the *CombMNZ* method (Fox and Shaw 1994), which reinforces the influence of documents retrieved by multiple systems. For a document d retrieved across n retrievers, the final fused score is computed as:

$$\text{Score}_{\text{CombMNZ}}(d) = \left(\sum_{i=1}^{n} \tilde{s}_i(d) \right) \times |\{i \mid \tilde{s}_i(d) > 0\}| \tag{4}$$

where $\tilde{s}_i(d)$ denotes the normalized score assigned to document d by retriever i, and the term $|\{i \mid \tilde{s}_i(d) > 0\}|$ indicates the number of retrievers that retrieved the document. This formulation leverages cross-retriever agreement by assigning higher scores to documents supported by multiple systems effectively boosting those that obtain consistent non-zero evidence across retrievers through the multiplicative factor in *CombMNZ*, which amplifies the weighted sum of normalized scores by the number of contributing retrievers.

Rank-based fusion

We apply the *Vanilla RRF*, in which the contribution of each retriever depends on the reciprocal of its assigned rank position. To enhance adaptability, we further consider two RRF variants that incorporate *query-specific weighting*: an RRF version (i) modulated by term-level *tf*idf*-based query weights (Spärck Jones 2004), and (ii) where weights are dynamically adjusted through a *query complexity classifier* (França et al. 2025), allowing the fusion strategy to adaptively balance sparse and dense retrieval signals per query. The final fused score for a document d is computed as:

$$\text{Score}_{\text{RRF}}(d) = \sum_{i=1}^{n} \frac{1}{\varepsilon + r_i(d)} \tag{5}$$

where $r_i(d)$ denotes the rank of document d in the result list from retriever i, n is the total number of retrievers, and ε is a small positive constant introduced to both avoid division by zero and dampen the effect of large rank differences. For the weighted RRF, the equation becomes:

$$\text{Score}_{\text{wRRF}}(d) = \sum_{i=1}^{n} \frac{w_i}{\varepsilon + r_i(d)} \tag{6}$$

where w_i denotes a query-specific weight applied to retriever i. Unlike the standard RRF (Cormack et al. 2009), wRRF introduces adaptive weighting across sparse and dense retrievers, allowing the contribution of each retriever to vary according to query characteristics. In the dynamic variant used in this study, the weights w_i are determined through either (1) $tf \times idf$ term importance, capturing lexical specificity, or (2) a fine-tuned BERT-based query complexity classifier on a randomly sampled subset (3,000 instances) of the *QA Expert*

Multi Hop QA V1.0 dataset[3], inspired by (Jeong et al. 2024). In this dataset, single-hop and multi-hop labels are used as a proxy for query complexity, enabling the classifier to distinguish between simple factoid queries and complex reasoning questions. At inference time, the predicted complexity score modulates the weighting scheme, assigning greater emphasis to dense retrieval for multi-hop queries and relying more on sparse retrievers for simple, lexical queries.

For all fusion formulations, we fix $n = 2$, corresponding to the combination of the best-performing sparse retriever and the best-performing dense retriever in the hybrid system.

3.3.2 Reranker Module

After the fusion stage, we apply a reranker module based on the cross-encoder model (Wang et al. 2020)[4]. This cross-encoder jointly encodes query–document pairs to compute a contextual relevance score and it is relatively fast compared with other variants of cross encoders. Trained on the MS MARCO Passage Ranking dataset, it serves as a second-stage reranker that refines the top-k candidates from the hybrid retrieval output[5]. It reranks the hybrid retrieval results, i.e. the fused best performing sparse and dense retrieval results, improving retrieval effectiveness through fine-grained semantic scoring.

The comparison is shown in Table 7 in Appendix A.1. showing retrieved outputs from three retrieval strategies, namely SPLADE, MPNet, and Hybrid with reranker. This table illustrates how each strategy affects evidence grounding and contextual relevance.

3.4 Generation Phase

The generation phase comprises two essential components: the prompt p and a selected pre-trained language model M. During the retrieval phase, the given retriever Ret which may correspond to a sparse (Ret_S), dense (Ret_D), or hybrid configuration integrating fusion techniques identifies the top- k most relevant documents D_{Ret} from the document collection DB with respect to the query q. The resulting set $D_{\text{Ret}} = \{d_1, d_2, \ldots, d_k\}$ is subsequently used to provide contextual evidence for the generation stage of the frozen LLM M. To construct the prompt p, the query q is concatenated with D_{Ret}, and the combined input is then passed to the model M to generate the corresponding response y for the original query q.

We employ two prompting strategies: standard prompting with detailed task instructions under zero-shot settings, and CoT prompting (Wei et al. 2022), which facilitates step-by-step reasoning. The latter has been effectively incorporated into RAG frameworks to enhance multi-hop inference and contextual grounding (Wang et al. 2024b,a). A sample CoT prompt is provided in Box 3.1. The standard prompting used in our experiments follows the same structure but excludes explicit reasoning traces, differentiating it from the CoT setup. Few-shot prompting (Brown et al. 2020) is not considered in this study. The pre-trained LLM M is used as a frozen model without any fine-tuning; its parameters remain fixed throughout evaluation. We employ three instruction-tuned open models for the generation phase: GEMMA 2-2B-IT[6],

[3]Hugging Face repository: https://huggingface.co/datasets/khaimaitien/qa-expert-multi-hop-qa-V1.0.
[4]Model identifier: `cross-encoder/ms-marco-MiniLM-L-6-v2`.
[5]The reranker jointly encodes each query–document pair, enabling token-level alignment for more accurate contextual relevance estimation.
[6]https://huggingface.co/google/gemma-2-2b-it (released July 2024, cutoff June 2024)

QWEN 3-4B-INSTRUCT[7], and LLAMA 3.1-8B-INSTRUCT[8]. Generation is configured in a deterministic manner (`do_sample = false`), ensuring consistent decoding behavior across runs (Atıl et al. 2025).

CoT Prompting

`[SYSTEM PROMPT]` You are a helpful assistant. Follow the user's instructions carefully and answer the question in one or maximum two sentences.
Do not repeat any part of the prompt in your final answer and answer strictly based on provided contexts.
`Prompt:`
`Question:` {query}
`Context:` {context}

Your task is to answer the given question by thinking progressively following the steps:
Step 0: Process the answer within your memory and only provide the necessary answer.
Step 1: Carefully read and understand the question and given contexts which can support you to answer better.
Step 2: Analyze whether the given contexts provides sufficient information to answer the question.
- If the given contexts do not provide sufficient information, respond with: 'The context does not provide sufficient information to answer the question.'
- If the given contexts provide sufficient information, proceed to Step 3.
Step 3: Generate an accurate and grounded response strictly based on the provided contexts. Avoid guessing or providing incorrect/hallucinated responses.
Step 4: Only when you are sure of Step 3, Clearly state the final answer in one or two sentences.

`Answer:`

4 Experimental Setup

This section outlines the experimental framework, which is founded on the LLM-driven RAG architecture discussed in Section 3. We evaluate the presented architecture's ability to reduce hallucinations in two ways: first, by looking at the retrieval system's performance alone (Section 3.1, Section 3.2, and Section 3.3); second, by looking at the pipeline's overall effectiveness, which considers both the intermediate retrieval results and the final response produced for the query q (Section 3.4). This framework facilitates an in-depth assessment of how variations in retrieval quality impact the model's capacity to reduce faithfulness-related hallucinations.

Our full implementation, the annotated data files via LLM-as-a-judge with human verification as well as statistical significance tests are publicly available in our GitHub repository[9] for reproducibility purposes.

[7] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507 (released August 2025, cutoff ≈ early 2025)
[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (released July 2024, cutoff December 2023)
[9] https://github.com/chandanasreemala/Beyond_Retrieval

4.1 Dataset

We evaluate the proposed LLM-driven RAG system on two datasets: HOTPOTQA (Yang et al. 2018) and HALUBENCH (Ravi et al. 2024). Our objective is to analyze retrieval–faithfulness interactions rather than retrieval effectiveness alone. While our experiments focus on QA-style retrieval, the RAG framework is inherently task-agnostic, as it grounds generation in retrieved evidence whether the task involves question answering, summarization, or reasoning. QA-oriented datasets, such as HOTPOTQA and HALUBENCH, offer well-structured benchmarks with annotated evidence, enabling reliable faithfulness evaluation. Importantly, the underlying retrieval–generation mechanism generalizes naturally to other knowledge-intensive tasks. Early retrieval-augmented models, notably REALM (Guu et al. 2020), demonstrates that joint retriever–encoder pre-training enables factual grounding and knowledge-intensive reasoning, establishing the foundation for subsequent retrieval-augmented systems. Recent work highlights that RAG architectures have evolved into highly versatile frameworks applicable across domains ranging from QA to summarization and reasoning, as comprehensively reviewed by (Gupta et al. 2024). These studies conceptualize summarization and reasoning as sequences of evidence-grounded QA steps, where each generated segment (e.g., a summary sentence or reasoning argument) is validated against retrieved context. For example, T5-based RAG systems retrieve supporting passages and condition generation on this information, yielding more faithful outputs and achieving state-of-the-art performance across multiple benchmarks.

Accordingly, we conduct experiments with *the distractor subset* of the HOTPOTQA dataset[10] and with HaluBench[11]. While HALUBENCH targets single-hop hallucination detection, HOTPOTQA-DISTRACTOR requires multi-hop reasoning and includes distractor documents, thereby introducing controlled yet realistic retrieval noise. This setup closely mirrors real-world RAG conditions and enables a more precise examination of evidence faithfulness.

HOTPOTQA

We employ the HOTPOTQA dataset (Yang et al. 2018), specifically *the train split of the distractor subset* comprising approximately 90,000 instances. HOTPOTQA is a large-scale question answering benchmark constructed from Wikipedia, designed to evaluate multi-hop reasoning by requiring models to integrate information from multiple documents to answer a single question. Each example contains a question (q_i), ten candidate paragraphs drawn from Wikipedia two providing supporting evidence and eight serving as distractors and a corresponding ground-truth answer. Annotators provide the ground-truth answer in natural language, and it is based solely on the information in the two supporting paragraphs. We deliberately use *the distractor subset* of HOTPOTQA, which increases retrieval difficulty by intermixing relevant and irrelevant passages, creating controlled yet realistic retrieval noise.

Regarding potential data contamination, HOTPOTQA is constructed using crowdsourced multi-hop questions over Wikipedia. For each question, crowd workers identify two supporting paragraphs that together provide the information necessary to answer the query and select eight additional distractor paragraphs. This results in ten candidate paragraphs per question, of which only two are relevant. While HOTPOTQA may share factual content (e.g., dates or

[10] https://huggingface.co/datasets/hotpot_qa
[11] https://huggingface.co/datasets/PatronusAI/HaluBench

entity names) with the pre-training corpora of the LLMs used in our experiments, its human-authored questions and reasoning chains are unique, making dataset-specific memorization unlikely.

HALUBENCH

We also leverage the HALUBENCH dataset (Ravi et al. 2024). The original HALUBENCH corpus contains approximately 15,000 context–question–answer triplets labeled for faithfulness. To ensure data quality, we perform additional preprocessing by removing missing or incomplete entries and filtering out duplicate instances, resulting in a cleaner and more reliable subset used for our evaluation. HALUBENCH is constructed from six distinct benchmark sources: HALUEVAL (Li et al. 2023), DROP (Dua et al. 2019), RAGTRUTH (Niu et al. 2024), FINANCEBENCH (Islam et al. 2023), PUBMEDQA (Jin et al. 2019), and COVIDQA (Möller et al. 2020). The benchmark consists of context–question–answer triplets labeled for faithfulness, containing both hallucinated and accurate responses across domains that include reasoning, general knowledge, finance, and healthcare. Each sample in HALUBENCH includes a context passage, a question related to that context, an answer generated by an LLM, and a ground-truth label indicating whether the answer is a hallucination. The dataset is specifically designed to evaluate hallucination detection in LLMs by providing both faithful and hallucinated responses across various real-world domains.

To ensure that the LLM-generated samples in HALUBENCH are of high quality and that the applied perturbations effectively induce hallucinations, random subsets are drawn from each distinct data source for human annotation. Perturbations consist of systematically modified LLM-generated answers designed to introduce inaccuracies while maintaining surface plausibility, such as swapping entities, altering numbers, or introducing reasoning errors. HALUBENCH is a particularly valuable and challenging dataset for hallucination mitigation in LLMs, as it contains subtle, difficult-to-detect hallucinations answers that appear plausible yet are in fact incorrect. HALUBENCH, released in 2024, may appear minimally in the training data of recently released LLMs, i.e. QWEN 3-4B; however, its examples are derived from earlier public QA datasets (e.g., PUBMEDQA, DROP, RAGTRUTH) and augmented with hallucination annotations and perturbed answers that are absent from the pre-training corpora. Consequently, any potential contamination in the training data of recently released LLMs is likely minimal.

4.2 Evaluation Protocol

For evaluation, we employ different versions of HOTPOTQA and HALUBENCH to separately assess the retrieval phase and the overall performance of the RAG pipeline in mitigating hallucinations.

Retrieval Evaluation

Retriever performance is assessed in an automated manner by determining whether the supporting paragraphs are successfully retrieved, with retrieved documents labeled as relevant or irrelevant accordingly using the metrics in Section 4.3. To evaluate retrieval effectiveness on HOTPOTQA, we use the full train split of the distractor subset denoted as $HotpotQA_{\text{full}}$, which contains approximately 90,000 examples. For HALUBENCH, comprising approximately 14,000 instances and denoted as $HaluBench_{\text{full}}$.

We report the retrieval effectiveness by using the retrieval metrics introduced in Section 4.3.1. We then evaluate statistical significance using the Wilcoxon Signed-Rank Test (Wilcoxon 1945), a non-parametric paired test widely adopted in information-retrieval (IR) experiments for comparing matched observations. Since the distributions of the retrieval metrics do not generally satisfy the normality assumption confirmed using the Shapiro–Wilk test (Shapiro and Wilk 1965) we use the Wilcoxon Signed-Rank Test instead of the paired t-test. Following significance testing, we apply the False Discovery Rate (FDR) correction to adjust p-values for multiple comparisons. In addition to statistical significance, we use Cohen's d to quantify effect sizes. While p-values help determine whether observed differences are unlikely to be due to noise, effect sizes capture the magnitude of those differences; together, they provide complementary evidence for interpreting the results. We refer to Cohen's d as an effect size measure, using Cohen's classic guidelines of 0.2, 0.5, and 0.8 for small, medium, and large effects, respectively (Cohen 2013).

Overall Evaluation

We evaluate the responses generated by the complete RAG pipeline to assess the faithfulness of each output y_i through comparing the generated response y_i against the corresponding ground-truth answer $\hat{y}_i$ and assigns one of these three labels:

- **Hallucinated Answer** (✗): The generated response is not grounded in the given context.
- **Correct Answer** (✓): The generated response aligns with the ground-truth and is fully supported by the provided context.
- **Calibrated No Answer** (**?**): The model recognizes insufficient contextual support and, instead of hallucinating, decides not to answer.

Building on the reasoning in Kalai et al. (2025) who argue that hallucinations arise from misaligned incentives that penalize uncertainty and reward confident guessing we define the third label as *Calibrated No Answer*. This label is intended not to penalize abstention but to recognize epistemic humility: cases where the model identifies insufficient contextual support and expresses justified uncertainty by abstaining rather than producing an overconfident, unguided response.

For the automated component, we employ an *LLM-as-a-Judge* framework (Huang et al. 2025a) using CLAUDE-SONNET-4.5 (Anthropic 2025) via API. While the generation phase relies on three open-weight LLMs, we select a closed-source judge model because such models have been shown to outperform open-source alternatives in complex reasoning and faithfulness assessment[12]. The evaluation is conducted in two stages: first, automatically through the *LLM-as-a-Judge* framework, which assigns predefined faithfulness labels; and second, on a randomly sampled subset, where the assigned labels and reasoning are verified by two human annotators with doctoral-level expertise in machine learning. Due to resource constraints, we evaluate the overall performance of the proposed system on randomly sampled subsets of HOTPOTQA and HALUBENCH, denoted as $HotpotQA_{\text{small}}$ and $HaluBench_{\text{small}}$, each containing 1,200 instances. The subset $HaluBench_{\text{small}}$ is constructed via stratified uniform sampling across its six constituent sources (HALUEVAL, DROP, RAGTRUTH,

[12] WildBench leaderboard by the Allen Institute for AI: https://huggingface.co/spaces/allenai/WildBench

FINANCEBENCH, PUBMEDQA, and COVIDQA), ensuring equal representation and preserving dataset diversity. No additional stratification (e.g., by query type or difficulty) is applied. To ensure fair and consistent comparisons, we use the same sampled subset across all model retriever combinations.

For HALUBENCH, we do not rely solely on automated evaluation since its ground-truth labels are not obtained directly from human annotators. Instead, they are produced from model-based perturbations and subsequently verified by expert annotators to ensure labeling quality and consistency, whereas the ground-truth answers in HOTPOTQA are fully human-annotated. Consequently, we apply only the second stage of human verification to HALUBENCH, where a subset of model-generated labels undergoes expert review rather than full human annotation from the outset. Following the procedure established in HALUBENCH (Ravi et al. 2024), the overall evaluation thus combines automated assessment with targeted human verification, balancing scalability with annotation reliability. Specifically, a randomly selected subset of 120 instances, with 20 from each subdomain, undergoes expert review to validate the faithfulness labels assigned by the *LLM-as-a-Judge* to model-generated answers.

4.3 Evaluation Metrics

We employ two types of evaluation metrics to assess our system: one set for measuring retrieval performance and another for evaluating overall performance in mitigating hallucinations. Specifically, we use two metrics to evaluate retrieval effectiveness and four complementary metrics to assess hallucination mitigation.

4.3.1 Retrieval Metrics

To assess the retrieval performance, we employ commonly used order-aware metrics from the Information Retrieval domain, namely Mean Average Precision (MAP) (Salemi and Zamani 2024; Yue et al. 2007; Yu et al. 2024) and Normalized Discounted Cumulative Gain (NDCG) (Salemi and Zamani 2024; Järvelin and Kekäläinen 2002; Sawarkar et al. 2024), for this evaluation. As described in Section 3, each retriever returns a ranked list of k documents (top-k). We compute the metrics at different cut-off values $k = 3, 5, 10$, corresponding to the number of retrieved documents considered in the evaluation.

MAP averages the precision@k metric at each relevant item position in the retrieved ranked list of documents, where precision@k measures the proportion of relevant documents in a ranked list of size k. For a query q, the Average Precision (AP) is defined as:

$$\text{AP}(q) = \frac{1}{|R_q|} \sum_{k=1}^{|D_q|} P(k) \cdot \text{rel}(k) \tag{7}$$

where $|R_q|$ is the number of relevant documents for query q, $|D_q|$ is the total number of documents retrieved for query q, $P(k)$ is the precision at rank k, and $\text{rel}(k)$ is an indicator function that is 1 if the document at rank k is relevant, and 0 otherwise. The MAP of a retriever is computed as the mean of AP over all queries Q in the datasets $HotpotQA_{\text{full}}$ and $HaluBench_{\text{full}}$ as follows:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q) \tag{8}$$

This metric rewards retrieval approaches that place more relevant documents at the top of the ranked list. DCG has a stronger concept of ranking, which discounts the *value* of each relevant document based on its rank in a ranked list of size k, using a logarithmic discount function as follows:

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \tag{9}$$

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \tag{10}$$

where rel_i is the relevance grade of a document at rank i, and for the binary case, if a document is relevant, the relevance grade is assigned as 1, otherwise 0. NDCG@k normalizes DCG@k by the *ideal* ranked list (IDCG@k), where every relevant document is ranked at the top. For both MAP and NDCG metrics, higher scores represent better retrieval performance.

4.3.2 Overall Evaluation Metrics

To evaluate the overall performance of the presented LLM-driven RAG in mitigating hallucinations, we utilize the following metrics as defined in (Yu et al. 2024; Chen et al. 2024a).

- **Accuracy** (Li et al. 2023; Kalra et al. 2024): The proportion of correctly generated answers among all evaluated responses (higher values indicate better performance).
- **Hallucination Rate**: The proportion of generated answers identified as hallucinated (lower values are desirable).
- **Rejection Rate** (Fadeeva et al. 2026): The proportion of cases where the model rejects and abstains from answering (lower values are desirable).
- **Adjusted Accuracy** (Es et al. 2024; Min et al. 2022; Kryściński et al. 2020): The proportion of correct predictions among all instances where the model provides an answer, excluding *calibrated no answer* cases. This adjustment captures the model's precision conditional on attempted answers, rewarding epistemically calibrated behavior where the model abstains appropriately under uncertainty while isolating performance on confident responses. The metric is defined as:

$$\text{Adjusted Accuracy} = \frac{\text{Correct Answers}}{\text{Correct Answers} + \text{Hallucinated Answers}} \times 100 \tag{11}$$

4.4 Results & Discussion

Retrieval Performance

For the HotpotQA dataset, as shown in Table 1, SPLADE consistently outperforms both sparse and dense retrievers in terms of MAP and NDCG across all cutoff values. Among the dense models, MPNetQA the QA-specialized variant of MPNet exhibits the lowest overall performance. While MPNet slightly outperforms Contriever at *MAP*@3, Contriever

Metric	*BM25*	*SPLADE*	*MPNet*	*Contriever*	*MPNetQA*
MAP@3	0.2660	**0.2819**	0.2020	0.1998	0.1677
MAP@5	0.2776	**0.2945**	0.2127	0.2128	0.1765
MAP@10	0.2847	**0.3019**	0.2202	0.2209	0.1828
NDCG@3	0.3304	**0.3493**	0.2580	0.2583	0.2149
NDCG@5	0.3428	**0.3622**	0.2712	0.2743	0.2267
NDCG@10	0.3562	**0.3756**	0.2864	0.2903	0.2403

Table 1: Retrieval performance of sparse and dense retrievers on the HOTPOTQA dataset. The table reports *MAP* and *NDCG* at different cutoff values (k = 3, 5, 10). *BM25* and *SPLADE* represent sparse models, while *MPNet*, *Contriever*, and *MPNetQA* are dense retrievers. Higher values indicate better retrieval quality. The best results for each metric are shown in **bold**. All pairwise comparisons are statistically significant at $p < 0.05$ except for *MAP*@5, *MAP*@10, and *NDCG*@3 between *MPNet* and *Contriever*.

achieves higher scores on *NDCG*@5 and *NDCG*@10. All pairwise comparisons are statistically significant under a paired Wilcoxon signed-rank test with FDR correction at $p < 0.05$, except for *MAP*@5, *MAP*@10, and *NDCG*@3 in the MPNET–CONTRIEVER comparison, where no statistically significant difference is observed.

These results are further supported by effect size analysis. The MPNET–CONTRIEVER comparison exhibits negligible effect sizes (e.g., $r \approx 0.01$), indicating that the differences between these two models are not practically meaningful. In contrast, most other comparisons yield medium to large effect sizes, pointing to substantial performance differences. An exception is the comparison between the sparse models SPLADE and BM25, where the effect size remains small despite statistical significance. Although the effect size between SPLADE and BM25 is small, SPLADE consistently achieves higher scores across all evaluation metrics, indicating a modest but reliable advantage. As a learned sparse retriever, SPLADE can assign richer term weights and model contextual information beyond traditional lexical matching, which helps explain its stronger effectiveness relative to BM25. Based on these observations, we choose SPLADE as the sparse component of our hybrid retriever, owing to its more consistent performance and suitability for integration with dense representations. For the dense component, we adopt MPNET in all hybrid configurations, preferring it over CONTRIEVER for computational efficiency; the resulting hybrid configurations are reported in Table 3.

For the HALUBENCH dataset, as shown in Table 2, SPLADE consistently outperforms both sparse and dense retrievers in terms of *MAP* and *NDCG* across all cutoff values. Among the dense retrievers, MPNETQA consistently achieves higher MAP and NDCG scores than its dense counterparts across all cutoffs. All pairwise comparisons are statistically significant under a paired Wilcoxon signed-rank test with FDR correction at $p < 0.05$, except for *NDCG*@5, where the gap between BM25 and SPLADE remains insignificant. In terms of effect sizes, most comparisons yield medium to large effects, indicating substantial performance differences, except for the comparison between the sparse models SPLADE and BM25, and between MPNET and its dense counterparts, which show small effect sizes despite statistical significance. Based on these observations, we similarly choose SPLADE as the sparse component of our hybrid retriever and MPNETQA as the dense component in all hybrid configurations, as it consistently outperforms the other dense retrievers. The

Metric	*BM25*	*SPLADE*	*MPNet*	*Contriever*	*MPNetQA*
MAP@3	0.8234	**0.8303**	0.7672	0.7477	0.7794
MAP@5	0.8283	**0.8343**	0.7740	0.7524	0.7858
MAP@10	0.8319	**0.8378**	0.7793	0.7562	0.7902
NDCG@3	0.8317	**0.8377**	0.7779	0.7570	0.7892
NDCG@5	0.8407	**0.8449**	0.7903	0.7656	0.8007
NDCG@10	0.8495	**0.8535**	0.8028	0.7746	0.8113

Table 2: Retrieval performance of sparse and dense retrievers on the HALUBENCH dataset. The table reports *MAP* and *NDCG* at different cutoff values (k = 3, 5, 10). Higher values indicate better retrieval quality. The best results for each metric are shown in **bold**. All pairwise comparisons are statistically significant at $p < 0.05$ except for *NDCG*@5 between *BM25* and *SPLADE*.

Metric	SPLADE	$\text{Hybrid}_{\text{RRF}}$	$\text{Hybrid}_{\text{CombMNZ}}$	$\text{Hybrid}_{\text{LI}}$	$\text{Hybrid}_{\text{LI–Rerank}}$
MAP@3	0.2819	0.2566	0.2615	0.2822	**0.2954**
MAP@5	0.2945	0.2735	0.2763	0.2949	**0.3073**
MAP@10	0.3019	0.2824	0.2850	0.3022	**0.3139**
NDCG@3	0.3493	0.3225	0.3281	0.3501	**0.3628**
NDCG@5	0.3622	0.3425	0.3448	0.3629	**0.3743**
NDCG@10	0.3756	0.3589	0.3606	0.3762	**0.3863**

Table 3: Retrieval performance of the best-performing sparse retriever, SPLADE, and hybrid retrieval models on the HOTPOTQA dataset under different rank-based and score-based fusion strategies. Hybrid models combine the scores of the top sparse and dense retrievers (SPLADE and MPNET) using RRF, COMBMNZ, and linear interpolation (LI). We also report a variant that applies reranking on top of linearly interpolated results ($\text{Hybrid}_{\text{LI–Rerank}}$). Higher values indicate better retrieval effectiveness. Best results for each metric are highlighted in **bold**. All pairwise differences are statistically significant at $p < 0.05$ based on a paired Wilcoxon signed-rank test with FDR correction.

medium effect size between MPNETQA and CONTRIEVER, together with the small effect size between MPNET and CONTRIEVER, further supports this choice; the resulting hybrid configurations are reported in Table 4.

Hybrid retrieval results on the HOTPOTQA dataset are presented in Table 3. Among the evaluated fusion methods, vanilla RRF and COMBMNZ yield moderate improvements, but their performance remains below that of the best single retriever, SPLADE. Weighted RRF (WRRF) variants, which dynamically assign weights based on query dependent signals, namely $tf \times idf$ term importance or a fine tuned BERT based query complexity classifier (Section 3.3.1), perform worst overall and are therefore omitted from the table. The best hybrid configuration employs static linear interpolation ($\text{Hybrid}_{\text{LI}}$), outperforming alternative fusion techniques. Using linear interpolation, we evaluate multiple weighting schemes to combine sparse and dense retrievers. Among the tested configurations with sparse weights of 0.5, 0.6, and 0.7, a sparse weight of 0.7 yields the best overall performance. This result is consistent with the observation that SPLADE, the sparse retriever, consistently outperforms the

Metric	SPLADE	$\text{Hybrid}_{\text{LI}}$	$\text{Hybrid}_{\text{LI–Rerank}}$
MAP@3	0.8303	0.8371	**0.8472**
MAP@5	0.8343	0.8413	**0.8514**
MAP@10	0.8378	0.8446	**0.8546**
NDCG@3	0.8377	0.8442	**0.8538**
NDCG@5	0.8449	0.8519	**0.8613**
NDCG@10	0.8535	0.8599	**0.8690**

Table 4: Retrieval performance of SPLADE and hybrid retrievers on the HALUBENCH dataset using linear interpolation (LI). Hybrid models combine SPLADE and MPNETQA via LI, with an additional reranking variant ($\text{Hybrid}_{\text{LI–Rerank}}$). Higher values indicate better retrieval effectiveness. Best results are highlighted in **bold**. All pairwise differences are statistically significant at $p < 0.05$ under a paired Wilcoxon signed-rank test with FDR correction.

dense models (see Table 1); giving it greater weight improves both MAP and NDCG across all cutoffs. We denote this configuration as $\text{Hybrid}_{\text{LI}}$ in Table 3. When applying reranking on top of linear interpolation ($\text{Hybrid}_{\text{LI-Rerank}}$), we observe further improvements in both MAP and NDCG. All pairwise differences between fusion strategies are statistically significant under a paired Wilcoxon signed-rank test with FDR correction at $p < 0.05$. Effect size analysis shows that most comparisons are small, except for the wRRF and vanilla RRF comparison, which exhibits a large effect. For linear interpolation weights, comparisons between 0.5 and 0.6 and between 0.5 and 0.7 show large effects, while the comparison between 0.6 and 0.7 shows a medium effect.

Hybrid results on the HALUBENCH dataset are summarized in Table 4. As linear interpolation ($\text{Hybrid}_{\text{LI}}$) yields the strongest performance outperforming alternative fusion methods on HOTPOTQA (Table 3) we restrict our analysis to this fusion strategy. We evaluate multiple interpolation weights to balance sparse and dense retrievers. A sparse weight of 0.6 (and dense weight of 0.4) achieves the best overall effectiveness, with performance comparable to the 0.7 setting. This suggests that MPNETQA provides more robust retrieval on HALUBENCH, likely due to the dataset's cross-domain characteristics. As SPLADE consistently outperforms all other base retrievers on HALUBENCH (Table 2), we pair it with MPNETQA in all hybrid configurations. Applying reranking on top of linear interpolation ($\text{Hybrid}_{\text{LI–Rerank}}$) yields further improvements in both MAP and NDCG. All pairwise differences are statistically significant at $p < 0.05$, except for the comparison between interpolation weights of 0.6 and 0.7; we therefore select 0.7 for linear interpolation, in line with HOTPOTQA. Most comparisons show medium effect sizes, while the difference between sparse weights 0.5 and 0.7 is small, and the difference between 0.6 and 0.7 is negligible, consistent with the statistical significance results. Table 6 reports results obtained with the best interpolation weight (0.7).

Overall, the results indicate that retrievers perform worse on HOTPOTQA than on HALUBENCH because the two benchmarks differ fundamentally in retrieval complexity and structure. HotpotQA combines large-scale open-domain retrieval, multi-hop reasoning dependencies, and distractor-rich contexts, making it harder for retrievers to isolate all relevant evidence effectively. In contrast, HaluBench focuses on single-hop, semantically close factual contexts optimized for hallucination evaluation, where retrieval relies on surface-level

Retriever	Generation	Acc. (%)	Halluc. Rate (%)	Rej. Rate (%)	Adj. Acc. (%)
Contriever	GEMMA-2B	0.33	**0.42**	99.25	**44.44**
$\text{Hybrid}_{\text{LI–Rerank}}$		**9.05**	30.82	60.13	22.69
Contriever	QWEN-4B	37.78	**9.34**	52.88	**80.18**
$\text{Hybrid}_{\text{LI–Rerank}}$		**49.25**	16.00	34.75	75.48
Contriever	LLAMA-8B	37.92	**13.42**	48.67	**73.86**
$\text{Hybrid}_{\text{LI–Rerank}}$		**46.29**	16.68	37.03	73.51

Table 5: Overall performance on the HOTPOTQA dataset evaluated using the LLM-as-a-Judge framework. **Bold** indicates the best retriever per generation model. Acc. (%) reports raw answer accuracy over all evaluated instances, whereas Adj. Acc. (%) reports accuracy after accounting for abstentions, i.e., correct answers among the cases where the model actually responds.

semantic similarity rather than deep reasoning links. Moreover, HaluBench contains only one relevant document per query, while HotpotQA requires identifying multiple supporting documents intermixed with distractors, further amplifying retrieval difficulty and noise propagation to downstream generation.

Overall Performance

For the overall performance evaluation, we use two representative retrievers (see Tables 3 and 4): the best-performing hybrid configuration, $\text{Hybrid}_{\text{LI–Rerank}}$ (i.e., SPLADE+MPNET for HOTPOTQA and SPLADE+MPNETQA for HALUBENCH), and a contrasting dense baseline, CONTRIEVER, to assess the impact of retrieval on generation quality. For generation, we adopt CoT prompting in all experiments, as it consistently outperforms standard prompting in preliminary results. Table 5 summarizes the comparative performance of three instruction-tuned LLMs GEMMA2-2B, QWEN3-4B and LLAMA3.1-8B on the HOTPOTQA dataset using the LLM-as-a-Judge evaluation pipeline.

For GEMMA2-2B, the Contriever baseline yields minimal raw accuracy (0.33%) and extremely high rejection (99.25%), reflecting a conservative but largely unproductive generation regime. In contrast, $\text{Hybrid}_{\text{LI–Rerank}}$ increases raw accuracy to 9.05% but also raises the hallucination rate to 30.82% and lowers adjusted accuracy from 44.44% to 22.69%. This suggests that, for low-capacity models, the hybrid configuration may improve nominal performance while degrading grounded reliability. In contrast, QWEN3-4B benefits substantially from hybridization, with raw accuracy increasing from 37.78% to 49.25% and the rejection rate dropping from 52.88% to 34.75%, suggesting improved evidence utilization and fewer abstentions. Although the hallucinations rate also rises from 9.34% to 16.00%, adjusted accuracy remains high at 75%, indicating that the hybrid setup preserves strong overall faithfulness. This suggests that QWEN3-4B is better able to exploit hybrid retrieval signals than GEMMA2-2B.

For LLAMA3.1-8B, hybrid retrieval again improves raw performance: accuracy rises from 37.92% to 46.29 %, while rejection drops from 48.67% to 37.03%. Although the hallucination rate increases modestly from 13.42% to 16.68%, adjusted accuracy remains essentially unchanged (73.86% vs. 73.51%), suggesting that hybrid retrieval mainly shifts the balance

Dataset	Retrieval	Generation	Acc. (%)	Hallucination Rate (%)	Rejection Rate (%)	Adjusted Acc (%)
HaluEval	*Contriever*	GEMMA-2B	**4.50**	**6.00**	89.50	**42.86**
	$\text{Hybrid}_{\text{LI–Rerank}}$		4.00	**6.00**	90.00	40.00
	Contriever	QWEN-4B	56.00	29.50	14.50	**65.50**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**57.50**	**32.50**	10.00	63.89
	Contriever	LLAMA-8B	50.50	25.00	24.50	**66.89**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**54.50**	**28.50**	17.00	65.66
FinanceBench	*Contriever*	GEMMA-2B	**0.50**	10.00	89.50	4.76
	$\text{Hybrid}_{\text{LI–Rerank}}$		**0.50**	**8.00**	91.50	**5.88**
	Contriever	QWEN-4B	18.50	**18.50**	63.00	**50.00**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**30.00**	36.00	34.00	45.45
	Contriever	LLAMA-8B	14.50	**35.00**	50.50	29.29
	$\text{Hybrid}_{\text{LI–Rerank}}$		**29.15**	**38.69**	32.16	**42.96**
RAGTruth	*Contriever*	GEMMA-2B	**33.33**	**11.62**	55.05	**74.16**
	$\text{Hybrid}_{\text{LI–Rerank}}$		23.50	**20.00**	56.50	54.02
	Contriever	QWEN-4B	85.00	8.50	6.50	90.91
	$\text{Hybrid}_{\text{LI–Rerank}}$		**86.43**	**7.54**	6.03	**91.98**
	Contriever	LLAMA-8B	**72.50**	**8.50**	19.00	**89.51**
	$\text{Hybrid}_{\text{LI–Rerank}}$		65.50	11.00	23.50	85.62
DROP	*Contriever*	GEMMA-2B	**2.51**	**2.51**	94.97	**50.00**
	$\text{Hybrid}_{\text{LI–Rerank}}$		0.50	7.00	92.50	6.67
	Contriever	QWEN-4B	21.00	**35.50**	43.50	**37.17**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**24.50**	**60.50**	15.00	28.82
	Contriever	LLAMA-8B	16.50	**27.50**	56.00	**37.50**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**18.00**	40.50	41.50	30.77
PubMedQA	*Contriever*	GEMMA-2B	**4.00**	4.50	91.50	**47.06**
	$\text{Hybrid}_{\text{LI–Rerank}}$		1.00	**1.50**	97.50	40.00
	Contriever	QWEN-4B	60.80	**35.68**	3.52	63.02
	$\text{Hybrid}_{\text{LI–Rerank}}$		**61.81**	**35.68**	2.51	**63.40**
	Contriever	LLAMA-8B	50.25	30.15	19.60	**62.50**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**52.26**	**31.66**	16.08	62.28
CovidQA	*Contriever*	GEMMA-2B	**1.59**	4.76	93.65	25.00
	$\text{Hybrid}_{\text{LI–Rerank}}$		1.58	**3.16**	95.26	**33.33**
	Contriever	QWEN-4B	9.52	**35.98**	54.50	20.93
	$\text{Hybrid}_{\text{LI–Rerank}}$		**15.26**	**42.11**	42.63	**26.61**
	Contriever	LLAMA-8B	9.63	**35.83**	54.55	21.18
	$\text{Hybrid}_{\text{LI–Rerank}}$		**16.75**	37.17	46.07	**31.07**
HaluBench	*Contriever*	GEMMA-2B	**7.76**	**6.58**	85.67	**54.12**
	$\text{Hybrid}_{\text{LI–Rerank}}$		5.21	7.65	87.14	40.52
	Contriever	QWEN-4B	42.09	**27.19**	30.72	**60.75**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**46.13**	**35.69**	18.18	56.38
	Contriever	LLAMA-8B	35.92	**26.90**	37.18	**57.18**
	$\text{Hybrid}_{\text{LI–Rerank}}$		**39.53**	31.20	29.27	55.89

Table 6: Cross-domain performance on HALUBENCH using the LLM-as-a-judge framework. **Bold** indicates the best retriever per generation model; **green** highlights best Acc. and Adjusted Acc., while **red** highlights worst Hallucination Rate reported per subdomain and overall.

between confidence and faithfulness rather than substantially changing overall quality. This pattern indicates that, for larger-capacity models such as LLAMA3.1-8B, hybrid retrieval increases the model's willingness to answer while preserving broadly similar faithfulness levels.

Table 6 reports overall and per-subdomain performance on HALUBENCH. Among the evaluated LLMs, QWEN-4B exhibits the highest performance variance across datasets, with large fluctuations in Accuracy, Adjusted Accuracy, and Hallucination Rate, indicating strong sensitivity to domain characteristics. In contrast, GEMMA-2B, the smallest model in our setup, shows more consistent but generally lower performance, along with relatively stable hallucination rates. The consistently high rejection rates (often around or above 90%) observed for GEMMA-2B suggest a conservative generation strategy in which the model prioritizes faithfulness over content coverage. Because adjusted accuracy rewards calibrated abstention, the observed gains likely reflect effective hallucination avoidance rather than improved reasoning ability, consistent with Kalai et al. (2025).

Hybrid retrieval improves accuracy across models, particularly for QWEN-4B, but its effect on hallucination varies by domain. In reasoning-intensive benchmarks such as DROP, FINANCEBENCH, and COVIDQA, it often increases the hallucination rate, indicating a trade-off between accuracy and faithfulness. In contrast, in knowledge-grounded settings such as RAGTRUTH and PUBMEDQA, hybrid retrieval improves alignment between retrieved evidence and generated outputs. On RAGTRUTH, both accuracy and adjusted accuracy increase while the hallucination rate decreases, indicating improved faithfulness; on PUBMEDQA, both accuracy and adjusted accuracy improves without affecting the hallucination rate. For LLAMA-8B, hybrid retrieval improves accuracy across most domains (e.g., HALUEVAL, FINANCEBENCH, DROP, PUBMEDQA, and COVIDQA), as well as overall on HALUBENCH, but these gains are consistently accompanied by increased hallucination and reduced abstention. An exception is RAGTRUTH, where accuracy decreases while both hallucination and rejection increase, likely due to its emphasis on fine grained faithfulness and strict evidence attribution. GEMMA-2B shows limited and inconsistent gains from enhanced retrieval, suggesting that stronger generative capacity is required to effectively leverage retrieved evidence.

At the aggregate level, hybrid retrieval introduces a trade-off between accuracy and faithfulness. While it improves raw accuracy for larger models such as QWEN and LLAMA on HALUBENCH, these gains are typically accompanied by increased hallucination and reduced rejection, leading to lower adjusted accuracy. This pattern is primarily driven by answer oriented QA benchmarks (e.g., HALUEVAL, FINANCEBENCH, COVIDQA), whereas attribution grounded faithfulness benchmarks such as RAGTRUTH exhibit more stable or degraded performance. Overall, these results highlight the difficulty of jointly optimizing accuracy and faithfulness in retrieval augmented generation. Our analysis is limited to small-to-mid-sized (2B–8B) models, and faithfulness behavior may differ at larger scales, particularly under parametric–retrieval conflicts.

During automated evaluation, we observe that CLAUDE-SONNET-4.5, used as the judge model, occasionally refuses to respond, with approximately 5–6 instances for HOTPOTQA and 21–23 instances for HALUBENCH across retrievers and LLMs. These cases correspond to the API flag `stop_reason: "refusal"`, indicating that the model deliberately abstained from generating an output due to policy or content-safety constraints rather than technical

failure. In HALUBENCH, this corresponds to 21 refusals for GEMMA and 23 refusals for both QWEN and LLAMA, or approximately 2% of all cases, with a broadly consistent pattern across models. For GEMMA, refusals are concentrated in the COVIDQA subset, whereas for QWEN and LLAMA they are primarily observed in COVIDQA and, to a lesser extent, PubMedQA (2 cases). Overall, we do not observe a systematic correlation with retriever configuration (Contriever vs. Hybrid), although Contriever exhibits a slightly higher refusal rate. CLAUDE-SONNET-4.5 occasionally produces labels that are not part of the predefined prompt set. For instance, in a few cases involving both the Contriever and Hybrid retrievers with LLAMA, it outputs additional categories such as `Incorrect` or `Incorrect context interpretation`, or `Uncertain`. We discard these instances from the metric evaluation to ensure consistency in the reported results. According to Anthropic's documentation (Anthropic 2025), such refusals are typically triggered by (i) evaluative prompts involving sensitive or personal content, (ii) overly open-ended or potentially high-risk judgments, (iii) malformed or partial JSON structures that the safety filter misinterprets as injection attempts, or (iv) batch-processing without resetting the conversation state following a prior refusal.

Within the LLM-as-a-Judge framework, human verification is carried out through binary labeling marking each automatic judgment as correct or incorrect, effectively denoting whether the annotator agrees or disagrees with the model's decision. Empirically, the judge's assessments align with human annotations in approximately 95% of cases, indicating a high degree of reliability in its evaluation behavior. Each instance is verified by a primary annotator and then double-checked by a second annotator. Since a small number of cases are challenging, the annotators reach consensus to ensure agreement and label consistency.

Computational Costs

The approximate runtimes of the retrieval components in our RAG pipeline are as follows. On HOTPOTQA, for 90,000 queries, BM25 requires approximately 10 minutes in total (6.7 ms/query), SPLADE 190 minutes (126.7 ms/query), MPNET 26 minutes (17.3 ms/query), MPNETQA 40 minutes (26.7 ms/query), and CONTRIEVER 43 minutes (28.7 ms/query). The hybrid fusion step requires 15 minutes with linear interpolation (10.0 ms/query) and 10 minutes with RRF (6.7 ms/query), while the cross-encoder reranker requires 32 minutes (21.3 ms/query). For generation, GEMMA2-2B, QWEN3-4B, and LLAMA-8B require variable runtimes depending on model size and dataset, with larger models generally needing longer processing times.

5 Limitations

Despite promising results, this study has several limitations. First, our experiments are limited to small-to-mid-sized (2B–8B) open-weight LLMs. Existing work suggests that conflicts between parametric knowledge and retrieved/contextual evidence remain an important issue in larger LLMs, with effects that depend on model size, task requirements, and the nature of the conflict (Su et al. 2024; Xu et al. 2024). Evaluating these dynamics in larger models remains an important direction for future work. From a methodological standpoint, the retrieval evaluation was conducted on medium-scale benchmarks rather than web-scale corpora due to computational constraints. Incorporating larger corpora such as MS MARCO

would better assess scalability and real-world robustness, while some retrievers' fine-tuning on MS MARCO limits direct extrapolation to web-scale settings. Second, the evaluation is restricted to QA datasets (HotpotQA and HaluBench), leaving generalization to other generation tasks (e.g., summarization, dialogue, fact-checking) untested empirically. Similarly, although multiple retrievers and fusion approaches were compared, further inclusion of state-of-the-art rerankers, late-interaction models, and adaptive retrieval pipelines could further enhance the presented framework. Another limitation is that error analysis was qualitative and limited; a finer-grained categorization of failure types (e.g., retrieval noise, grounding drift, reasoning inconsistency) is needed to better understand when hallucinations occur and under which retrieval conditions. Finally, although combining LLM-as-a-Judge automatic scoring (1,200 samples per dataset) with limited expert verification provides a scalable approach for evaluating faithfulness, it may still under-represent subtle or nuanced hallucinations. Future work will extend evaluation across diverse tasks, investigate adaptive retrieval in adversarial and multilingual settings, incorporate advanced reranking architectures, support detailed error analysis, and expand human evaluation to strengthen faithfulness assessment.

6 Conclusion & Future Work

This study systematically evaluated faithfulness hallucinations in RAG systems, focusing on the interplay between retrieval quality, fusion techniques, and generation strategies. By leveraging two benchmark datasets HOTPOTQA and HALUBENCH and analyzing three open-weight LLMs, five retrieval models, and three fusion methods, we provided a broad assessment of how retrieval and generation components interact to affect faithfulness. Our findings highlight that robust, contextually rich retrieval and effective fusion are essential for reducing hallucinations and improving the reliability of LLM-driven RAG systems.

Future work will address several limitations of the current study. Our experiments are limited to small-to-mid-sized (2B–8B) open-weight LLMs, and evaluating whether the observed conclusions hold for larger models remains an important direction for future work. We also plan to extend evaluation to web-scale corpora to assess generalizability and robustness in real-world settings. Expanding beyond QA tasks, we aim to investigate the effectiveness of retrieval-augmented generation for other generation tasks such as summarization, dialogue, and fact-checking. Incorporating additional advanced retrievers, rerankers, and late interaction models, together with more sophisticated query expansion methods, would provide a more comprehensive understanding of hybrid retrieval performance across diverse query types. We also intend to perform finer-grained error analyses to systematically categorize hallucination types and better understand their causes. Finally, we will scale human evaluation within the LLM-as-a-Judge framework to capture subtle or nuanced hallucinations more effectively and obtain insights from the model's reasoning explanations, improving the reliability of faithfulness assessments.

7 Acknowledgments

This work has been supported by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", by the European Union under ERC-2018-ADG GA 834756 (XAI), and under grant agreement no. 101120763 ("TANGO").

References

Alansari A, Luqman H. Large language models hallucination: A comprehensive survey. Computer Science Review. 2026;61:100970.

Anthropic.: Claude Sonnet 4.5: Model Overview and Specifications; 2025. Accessed October 21, 2025. https://www.anthropic.com/claude/sonnet.

Atıl B, Aykent S, Chittams A, Fu L, Passonneau RJ, Radcliffe E, et al. Non-Determinism of "Deterministic" LLM System Settings in Hosted Environments. In: Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems; 2025. p. 135–148.

Ayala O, Bechard P. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track); 2024. p. 228–238.

Azad HK, Deepak A. Query Expansion Techniques for Information Retrieval: A Survey. Information Processing & Management. 2019;56(5):1698–1735. https://doi.org/10.1016/j.ipm.2019.05.009.

Barnett S, Kurniawan S, Thudumu S, Brannelly Z, Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system. In: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI; 2024. p. 194–199.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877–1901.

Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38; 2024. p. 17754–17762.

Chen S, Xiong M, Liu J, Wu Z, Xiao T, Gao S, et al. In-context sharpness as alerts: an inner representation perspective for hallucination mitigation. In: Proceedings of the 41st International Conference on Machine Learning; 2024. p. 7553–7567.

Chuang YS, Xie Y, Luo H, Kim Y, Glass JR, He P. Dola: Decoding by contrasting layers improves factuality in large language models. In: International Conference on Learning Representations, vol. 2024; 2024. p. 54158–54183.

Cohen J. Statistical power analysis for the behavioral sciences. routledge; 2013.

Cormack GV, Clarke CLA, Büttcher S. Reciprocal Rank Fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval ACM; 2009. p. 758–759.

Dua D, Wang Y, Dasigi P, Stanovsky G, Singh S, Gardner M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 2368–2378.

Es S, James J, Anke LE, Schockaert S. Ragas: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations; 2024. p. 150–158.

Fadeeva E, Rubashevskii A, Piatrashyn D, Vashurin R, Dhuliawala S, Shelmanov A, et al. Faithfulness-Aware Uncertainty Quantification for Fact-Checking the Output of Retrieval-Augmented Generation. In: Findings of the Association for Computational Linguistics: ACL 2026; 2026. p. 6814–6836.

Formal T, Piwowarski B, Clinchant S. SPLADE: Sparse lexical and expansion model for first stage ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021. p. 2288–2292.

Fox EA, Shaw JA. Combination of multiple searches. NIST special publication SP. 1994;243.

França C, Rabbi G, Salles T, Cunha W, Rocha L, André Gonçalves M. Ranking-based Fusion Algorithms for Extreme Multi-label Text Classification (XMTC). arXiv e-prints. 2025;p. arXiv–2507.

Gao Y, Zhang W, Du B, Xia GS. Probing Latent Knowledge Conflict for Faithful Retrieval-Augmented Generation. arXiv preprint arXiv:251012460. 2025;https://arxiv.org/html/2510.12460v1.

Gema AP, Jin C, Abdulaal A, Diethe T, Teare P, Alex B, et al. DeCoRe: Decoding by Contrasting Retrieval Heads to Mitigate Hallucinations. arXiv preprint arXiv:241018860. 2024;https://arxiv.org/abs/2410.18860, https://doi.org/10.48550/arXiv.2410.18860.

Gumaan E. Theoretical foundations and mitigation of hallucination in large language models. arXiv preprint arXiv:250722915. 2025;.

Gupta S, Ranjan R, Singh SN. A Comprehensive Review of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape, and Future Directions. arXiv preprint arXiv:241012837. 2024;https://arxiv.org/pdf/2410.12837.pdf.

Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval-augmented language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning; 2020. p. 3929–3938.

Hu H, He C, Xie X, Zhang Q. Lrp4rag: Detecting Hallucinations in Retrieval-Augmented Generation Via Layer-Wise Relevance Propagation. Available at SSRN 5199254. 2025chan;https://arxiv.org/abs/2408.15533.

Huang H, Bu X, Zhou H, Qu Y, Liu J, Yang M, et al. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Findings of the Association for Computational Linguistics: ACL 2025 Vienna, Austria: Association for Computational Linguistics; 2025. p. 5880–5895. https://aclanthology.org/2025.findings-acl.306/.

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2025;43(2):1–55.

Islam P, Kannappan A, Kiela D, Qian R, Scherrer N, Vidgen B. Financebench: A new benchmark for financial question answering. arXiv preprint arXiv:231111944. 2023;.

Islam SMT, Zaman SMM, Jain V, Rani A, Rawte V, Chadha A, et al. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv preprint arXiv:240101313. 2024;https://arxiv.org/abs/2401.01313.

Izacard G, Caron M, Hosseini L, Riedel S, Bojanowski P, Joulin A, et al. Unsupervised Dense Information Retrieval with Contrastive Learning. Transactions on Machine Learning Research. 2022;.

Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS). 2002;20(4):422–446.

Jeong S, Baek J, Cho S, Hwang SJ, Park J. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 7036–7050. https://aclanthology.org/2024.naacl-long.389/.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Computing Surveys. 2023;55(12):1–38.

Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. Pubmedqa: A dataset for biomedical research question answering. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP); 2019. p. 2567–2577.

Kalai AT, Nachum O, Vempala SS, Zhang E. Why language models hallucinate. arXiv preprint arXiv:250904664. 2025;.

Kalra R, Wu Z, Gulley A, Hilliard A, Guan X, Koshiyama A, et al. HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. In: Proceedings of the 1st workshop on customizable nlp: Progress and challenges in customizing nlp for a domain, application, group, or individual (customnlp4u); 2024. p. 237–256.

Karpukhin V, Oguz B, Min S, Lewis PS, Wu L, Edunov S, et al. Dense Passage Retrieval for Open-Domain Question Answering. In: EMNLP (1); 2020. p. 6769–6781.

Khattab O, Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval New York, NY, USA: Association for Computing Machinery; 2020. p. 39–48. https://doi.org/10.1145/3397271.3401075.

Kryściński W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP); 2020. p. 9332–9346.

Kumar A, Kim H, Nathani JS, Roy N. Improving the Reliability of LLMs: Combining Chain-of-Thought Reasoning and Retrieval-Augmented Generation. arXiv preprint arXiv:250509031. 2024;https://arxiv.org/abs/2505.09031.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459–9474.

Li J, Cheng X, Zhao X, Nie JY, Wen JR. Halueval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing; 2023. p. 6449–6464.

Lin J, Ma X, Lin SC, Yang JH, Pradeep R, Nogueira R. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021. p. 2356–2362.

Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the middle: How language models use long contexts. Transactions of the association for computational linguistics. 2024;12:157–173.

Manmatha R, Sever H. A formal approach to score normalization for meta-search. In: Proceedings of HLT, vol. 2; 2002. p. 88–93.

Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995;38(11):39–41.

Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al.: Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?; 2022. https://arxiv.org/abs/2202.12837.

Möller T, Reina A, Jayakumar R, Pietsch M. COVID-QA: A question answering dataset for COVID-19. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020; 2020. .

Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, et al. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024. p. 10862–10878.

Rathee M, Venktesh V, MacAvaney S, Anand A. Test-time Corpus Feedback: From Retrieval to RAG. In: Findings of the Association for Computational Linguistics: EACL 2026; 2026. p. 5637–5656.

Ravi SS, Mielczarek B, Kannappan A, Kiela D, Qian R. Lynx: An open source hallucination evaluation model. arXiv preprint arXiv:240708488. 2024;.

Robertson S, Zaragoza H, et al. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval. 2009;3(4):333–389.

Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024. p. 2395–2400.

Sawarkar K, Mangal A, Solanki SR. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In: 2024 IEEE 7th international conference on multimedia information processing and retrieval (MIPR) IEEE; 2024. p. 155–161.

Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). Biometrika. 1965;52(3-4):591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shi W, Han X, Lewis M, Tsvetkov Y, Zettlemoyer L, Yih Wt. Trusting your evidence: Hallucinate less with context-aware decoding. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers); 2024. p. 783–791.

Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval Augmentation Reduces Hallucination in Conversation. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021. p. 3784–3803.

Song K, Tan X, Qin T, Lu J, Liu TY. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems. 2020;33:16857–16867.

Spärck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation. 2004;60(5):493–502.

Su Z, Zhang J, Qu X, Zhu T, Li Y, Sun J, et al. CONFLICTBANK: a benchmark for evaluating knowledge conflicts in large language models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024. p. 103242–103268.

Tian F, Ganguly D, Macdonald C. Is Relevance Propagated from Retriever to Generator in RAG? In: European Conference on Information Retrieval Springer; 2025. p. 32–48.

Trivedi H, Balasubramanian N, Khot T, Sabharwal A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Toronto, Canada: Association for Computational Linguistics; 2023. p. 10014–10037. https://aclanthology.org/2023.acl-long.557/.

Vogt CC, Cottrell GW. Fusion via a linear combination of scores. Information retrieval. 1999;1(3):151–173.

Wallat M, et al. Correctness is not Faithfulness in Retrieval Augmented Generation. ACM Transactions on Information Systems. 2025;.

Wang L, Chen H, Yang N, Huang X, Dou Z, Wei F. CoRAG: Chain-of-Retrieval Augmented Generation. arXiv preprint arXiv:250114342. 2024;.

Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: Advances in Neural Information Processing Systems (NeurIPS); 2020. p. 5776–5788.

Wang Z, Liu A, Lin H, Li J, Ma X, Liang Y. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. arXiv preprint arXiv:240305313. 2024;.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022;35:24824–24837.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945;1(6):80–83.

Xiong G, He Z, Liu B, Sinha S, Zhang A. Toward Faithful Retrieval-Augmented Generation with Sparse Autoencoders. In: The Fourteenth International Conference on Learning Representations; 2026. https://openreview.net/forum?id=hgBZP67BkP.

Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, et al. Knowledge conflicts for llms: A survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024. p. 8541–8565.

Xu S, Yan Z, Dai C, Wu F. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. Frontiers in Public Health. 2025;13:1635381.

Yang Z, Qi P, Zhang S, Bengio Y, Cohen W, Salakhutdinov R, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018. p. 2369–2380.

Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z. Evaluation of retrieval-augmented generation: A survey. In: CCF Conference on Big Data Springer; 2024. p. 102–120.

Yue Y, Finley T, Radlinski F, Joachims T. A support vector method for optimizing average precision. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval New York, NY, USA: Association for Computing Machinery; 2007. https://doi.org/10.1145/1277741.1277790.

Zhang Q, Xiang Z, Xiao Y, Wang L, Li J, Wang X, et al. FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Vienna, Austria: Association for Computational Linguistics; 2025. p. 21863–21882. https://aclanthology.org/2025.acl-long.1062/.

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics. 2025;51(4):1373–1418.

A Appendix

A.1 Different Retrievers Retrieved Documents

Model	Retrieved Documents
SPLADE	{'doc id': '64_0', 'score': 22.4232, 'full text': 'iqaluit airport iata: yfb, icao: cyfb serves iqaluit, nunavut, canada and is located adjacent to the town.'}, {'doc id': '39763_2', 'score': 19.4198, 'full text': 'iqaluit inuktitut, meaning place of fish, is the capital of the canadian territory of nunavut; its largest community, and its only city.'}, {'doc id': '64_5', 'score': 17.4174, full text': 'canadian north inc. is an airline headquartered in calgary, alberta, canada.'}
MPNet	{'doc id': '39763_1', 'score': 0.743383, 'full text': 'it is located in iqaluit, nunavut.'}, {'doc id': '64_0', 'score': 0.743217, 'full text': 'iqaluit airport iata: yfb, icao: cyfb serves iqaluit, nunavut, canada and is located adjacent to the town.'}, {'doc id': '64_8', 'score': 0.680773, 'full text': 'its main base is edmonton airport.'}
Hybrid with Reranker	{'doc id': '64_0', 'score': 0.995698, 'rank': 1, 'full text': 'iqaluit airport iata: yfb, icao: cyfb serves iqaluit, nunavut, canada and is located adjacent to the town.'}, {'doc id': '64_5', 'score': 0.656949, 'rank': 2, 'full text': 'canadian north inc. is an airline headquartered in calgary, alberta, canada.'}, {'doc id': '39763_2', 'score': 0.341790, 'rank': 3, 'full text': 'iqaluit inuktitut, meaning place of fish, is the capital of the canadian territory of nunavut; its largest community, and its only city.'}

Table 7: An example from HotpotQA showing different retrieval models' outputs for the question: "Iqaluit Airport and Canadian North are based out of what country?"