Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 3, pp. 193-207, June 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.3.193

LLM-RAG 성능 및 안정성 개선을 위한 화이트 노이즈 최적화

김다영[1] · 고영명[2†]

[1]포항공과대학교 소셜데이터사이언스학과 / [2]포항공과대학교 산업경영공학과

Optimizing White Noise for Stability and Strategic Performance Enhancement of LLM-RAG Systems

Dayoung Kim[1] · Young Myoung Ko[2]

[1]Department of Social Data Science, Pohang University of Science and Technology
[2]Department of Industrial and Management Engineering, Pohang University of Science and Technology

Retrieval-Augmented Generation (RAG) has emerged as a technique for improving the performance of language models (LLMs), where the inclusion of noisy documents is unavoidable. While prior studies have largely treated noise as detrimental and focused on its removal, this study demonstrates that certain types and combinations of noisy documents can improve LLM-based question answering performance. We categorize noisy documents into six types and systematically evaluate their individual and combined effects within an LLM-RAG framework. To analyze the mechanisms underlying performance changes, we employ quantitative metrics that capture improvement and degradation. We define beneficial noise as White Noise and harmful noise as Black Noise, framing noise along a brightness continuum. Furthermore, we propose a Pure White Noise selection method that provides practical guidance for leveraging beneficial noise to enhance the robustness of RAG systems. Our results suggest that noise can be strategically utilized as a resource for robust RAG design.

Keywords: Retrieval-Augmented Generation, Large Language Models, Information Retrieval, Noise-aware Retrieval, Robustness

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a powerful framework for enhancing large language models (LLMs) (Lewis *et al.*, 2020; Guu *et al.*, 2020; Zhao *et al.*, 2024b). By integrating external knowledge, RAG enhances the reasoning capabilities and factual accuracy of LLMs, demonstrating strong performance across a diverse range of question answering and reasoning tasks (Lewis *et al.*, 2020; Izacard *et al.*, 2023). Furthermore, by supplying relevant documents, RAG mitigates common LLM limitations such as hallucinations and factual inaccuracies, thereby improving the reliability and accuracy of the system's responses (Asai *et al.*, 2024; Yan *et al.*, 2024). However, real-world retrieval pipelines inevitably include noisy documents (Fang *et al.*, 2024). The presence of noise presents a fundamental challenge to the robustness of RAG systems, as it cannot be fully eliminated in practice (Cuconasu *et al.*, 2024a; Yoran *et al.*, 2023). Thus, handling noise has become a critical issue for RAG performance and stability.

Prior studies have mostly treated noisy documents as harmful, focusing on filtering or mitigating their effects (Islam *et al.*, 2024; Zhu *et al.*, 2024; Yu *et al.*, 2024). Although these approaches can stabilize performance in some cases (Chang *et al.*, 2024; Ling *et al.*, 2025), they are based on the assumption that all noise degrades

† Corresponding author: Professor, Young Myoung Ko, Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Korea, Tel : +82-54-279-2373, Fax : +82-54-279-2870, E-mail: youngko@postech.ac.kr
Received January 20, 2026; Accepted March 16, 2026.

outcomes. In contrast, recent findings suggest that moderate levels of distractors can, in certain contexts, sharpen model attention or promote more robust decision making, thereby contributing positively to performance (Cuconasu *et al.*, 2024b; Shim *et al.*, 2025). However, despite these emerging insights, systematic analyses of noise in RAG remain limited, and the conditions under which noise becomes beneficial are not well understood (Gan *et al.*, 2025; Sharma, 2025). Therefore, existing research often treats noise as a peripheral issue, leaving its potential utility underexplored under the prevailing assumption that it is exclusively harmful.

In this paper, we investigate noise as a factor that can be either beneficial or detrimental depending on its type and composition. We categorize documents into six types (e.g., Relevant, Supportive, Ambiguous, Irrelevant, Misleading, and Out-of-Domain) and define the latter four as noise in our analysis. Using biomedical QA and fact-verification datasets, we conduct a systematic study of both individual noise types and their combinations. Our findings reveal that Irrelevant documents, unlike other noise types, consistently yield positive effects and can mitigate the negative impact of harmful noise when used in combinations. This provides empirical evidence that Irrelevant documents can serve as beneficial noise, and we further analyze the mechanisms behind these effects. Building on these insights, we propose a method to selectively identify and exploit beneficial noise to enhance the robustness of RAG systems. While prior approaches for constructing noise pools often relied on random sampling after excluding relevant documents (Cuconasu *et al.*, 2024a; Yoran *et al.*, 2023), which can lead to unstable outcomes, we introduce a more controlled approach based on a cosine-similarity band analysis between queries and irrelevant document candidates. This approach filters out harmful bands that degrade performance while retaining bands that contribute to stable improvements. Our results demonstrate that noise can be transformed from a risk factor into a strategically usable resource.

To conceptualize this phenomenon, we introduce a novel framework inspired by the intuitive notion of white noise, a type of ambient sound widely recognized for its concentration enhancing properties. Analogously, we redefine the retrieval noise along a brightness continuum that distinguishes between White Noise (beneficial noise that improves performance or reduces variance) and Black Noise (harmful noise that degrades outcomes). This conceptual framework is illustrated in <Figure 1>. Within this continuum, we further identify Pure White Noise as the optimal subset of beneficial noise, yielding the most consistent improvements. Based on this framework, we propose a selection method that strategically identifies and leverages Pure White Noise while filtering out harmful components to enhance RAG robustness. This perspective extends beyond conventional noise elimination, reframing noise as a strategic asset and a design principle for RAG systems.

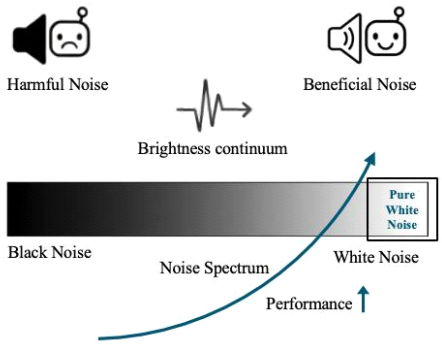


Figure 1. Conceptual framework of noise along a brightness continuum from Black Noise (harmful) to White Noise (beneficial), highlighting the Pure White Noise region.

Ultimately, this work challenges the conventional view of noise in RAG. We reframe noise not as an inherent limitation to be eliminated, but as a potential asset that can be strategically harnessed. This perspective offers new insights for designing more robust RAG systems and provides practical guidelines for future research on noise-aware retrieval and document composition.

The main contributions of this work are as follows:

- A systematic taxonomy and empirical evaluation of noise: We define six document types and analyze their individual and combined effects on RAG performance.
- A novel conceptual framework for noise: We introduce the concepts of White Noise (beneficial) and Black Noise (harmful) along a brightness continuum to reinterpret the role of noise.
- A strategic method for beneficial noise selection: We propose the Pure White Noise method, a band-based filtering approach that selectively leverages beneficial noise to improve system robustness and stability.
- Practical design implications for robust RAG systems: Our findings shift the paradigm from noise elimination to strategic utilization, offering actionable guidelines for noise-aware RAG applications.

2. Related Work

2.1 Robustness of Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as a fundamental methodology for enhancing the accuracy and reliability of Large Language Models (LLMs) by leveraging external knowl-

edge sources. However, the retrieval process inevitably incorporates noise documents that are irrelevant, ambiguous, or even misleading with respect to the query. Ensuring robustness, that is, the ability to effectively handle such noise while maintaining stable performance, has therefore become a critical challenge in RAG research.

Various approaches have been proposed to address this issue. First, several studies have focused on controlling the model reasoning process through prompt engineering. Yu *et al.* (2023) proposed a structured prompting technique called Chain-of-Note (CoN), which guides the model to sequentially evaluate retrieved documents and generate intermediate summary notes, thereby inducing stable reasoning even in the presence of noise. Similarly, Chen *et al.* (2024) introduced a risk-aware framework using counterfactual prompting to proactively manage potential risks during retrieval and enhance robustness. Second, strategies that directly inject noise into training data to enable model adaptation have been actively explored. Fang *et al.* (2024) employed adaptive adversarial training to incorporate noisy documents into the learning process, thereby training the model to adaptively respond when exposed to noise similar to that in real-world environments. Particularly in specialized domains such as medicine, where retrieval precision is critical, SearchRAG (Shi *et al.*, 2025) optimizes selective retrieval by automatically generating expert-level search queries to minimize the influx of unnecessary noise. Recently, approaches that actively evaluate and correct the reliability of retrieved information have also emerged. The Corrective Retrieval Augmented Generation (CRAG) framework (Yan *et al.*, 2024) employs a lightweight retrieval evaluator to classify the quality of retrieved documents as Correct, Incorrect, or Ambiguous. Based on this evaluation, CRAG implements differentiated strategies, such as correcting incorrect information through web searches and refining ambiguous information by combining multiple sources, to further strengthen robustness.

These studies commonly operate under the premise that completely eliminating noise from RAG systems is impractical. Instead, they represent meaningful attempts to minimize its adverse effects and ensure RAG robustness through sophisticated prompt design, specialized training, and retrieval optimization with post-correction strategies.

2.2 The Discovery of Beneficial Noise

Traditionally, noise in RAG research has been regarded as a detrimental factor that degrades the quality of responses. However, recent studies have challenged this paradigm by demonstrating that certain types of noise can positively contribute to model performance enhancement.

The first experimental evidence of this possibility was presented by Cuconasu *et al.* (2024a). They demonstrated that some irrelevant documents can act as distractors that, paradoxically, guide the model to focus more intently on the core information related to the correct answer, thereby improving reasoning performance. However, this study had limitations, including the potential for bias caused by a predefined, fixed noise pool and the specific positioning of documents in its experimental design. Moreover, it did not fully account for the effects of highly distracting noise. A more in-depth analysis of the dual nature of noise was conducted by Wu *et al.* (2024). They systematically classified noise into seven types based on linguistic perspectives and proposed the NoiserBench benchmark for evaluation. Through this work, they demonstrated that noise can be practically distinguished into beneficial and harmful categories, though they did not clearly elucidate the quantitative mechanisms underlying these effects.

Similar phenomena have been observed in other studies. Zhao *et al.* (2024a) reported instances where noise enhanced performance but did not provide an elaborate explanation of the underlying causes. Abdolazimi *et al.* (2024) conducted a comprehensive review across various domains, arguing that noise can serve as a potential resource for improving system performance rather than merely functioning as an obstacle. Conversely, Leto *et al.* (2024) presented a contrasting perspective, asserting that all noise ultimately leads to performance degradation. These recent studies have challenged the conventional wisdom that noise is harmful and confirmed the potential existence of beneficial noise. However, the fundamental mechanisms explaining why and under what conditions these positive effects manifest remain unclear.

To bridge this gap, we aim to investigate the operational principles of beneficial noise using metrics that measure LLM attention distribution and contextual reasoning. Furthermore, our research differs from prior work in two key aspects. First, to overcome the intrinsic biases of limited scenarios, we systematically analyze the impact of various noise types, both individually and in combination, under conditions similar to real-world environments. Second, unlike previous studies confined to open-domain contexts using open-domain QA datasets, we extend our experiments to biomedical QA and fact-verification datasets. This expansion holds significant importance as it allows us to validate whether the positive effects of beneficial noise can be generalized beyond specific domains.

2.3 Noise Filtering and Selection

Another major research stream for enhancing RAG robustness

centers on filtering and selection techniques that either remove noisy documents after retrieval or selectively retain only useful information.

Among document-level filtering approaches, FineFilter (Zhang *et al.*, 2025) effectively filters noise through document-level fine-tuning, and Rationale-Guided RAG (Sohn *et al.*, 2024) evaluates whether each retrieved document contributes to the final answer, retaining helpful documents while discarding obstructive ones. Research has also been conducted focusing on more granular sentence-level filtering. Wang *et al.* (2023) assessed the utility of each sentence by comprehensively evaluating multiple criteria, including whether it contained the answer, its similarity to the query, and its impact on the probability of generating the correct answer. By using a trained filtering model, they automatically refined the context at test time, thereby improving RAG accuracy. Furthermore, active approaches where the generation model itself acts as the filtering agent have been proposed. Self-RAG (Asai *et al.*, 2024) trains LLMs to evaluate and critique the relevance and completeness of retrieved documents. The model employs reflection tokens to judge the utility of retrieved information and selectively uses only the necessary data to generate answers. Although these methods have significantly contributed to improving RAG performance through precise document and sentence-level filtering, they typically make a binary judgment between usefulness and irrelevance. This poses the potential risk of inadvertently eliminating the beneficial noise discussed in Section 2.2.

To address this limitation, we propose a novel strategy that preserves good noise while selectively removing harmful portions of the noise spectrum. Specifically, the band-based filtering (Pure White Noise selection) introduced in this research aims to simultaneously maximize both performance and stability of RAG systems by removing noise from bands exhibiting harmful characteristics while selectively retaining noise corresponding to beneficial noise bands.

3. Methodology

3.1 Framework Overview

As illustrated in <Figure 2>, our proposed research framework modifies the baseline Retrieval-Augmented Generation (RAG) pipeline through two primary stages: Step 1 and Step 2. The conventional RAG process involves retrieving relevant documents from an external knowledge base in response to a user query and using them as context for an LLM to generate a response. Our work intervenes at the context injection step, systematically introducing various types of noise documents alongside relevant ones to analyze their impact on RAG performance.

In Step 1, we conduct experiments with predefined noise types and their combinations to distinguish between White Noise, which is beneficial to performance, and Black Noise, which is detrimental to it. To this end, we generate random noise pools and observe performance changes as each noise type is injected into the LLM prompt with the query and relevant documents. This step specifically examines whether the beneficial noise effect reported in prior studies can be generalized from domain-specific QA datasets to fact-verification tasks.

The initial irrelevant pool, a candidate for White Noise, is constructed following the methodology of prior studies (Cuconasu *et*

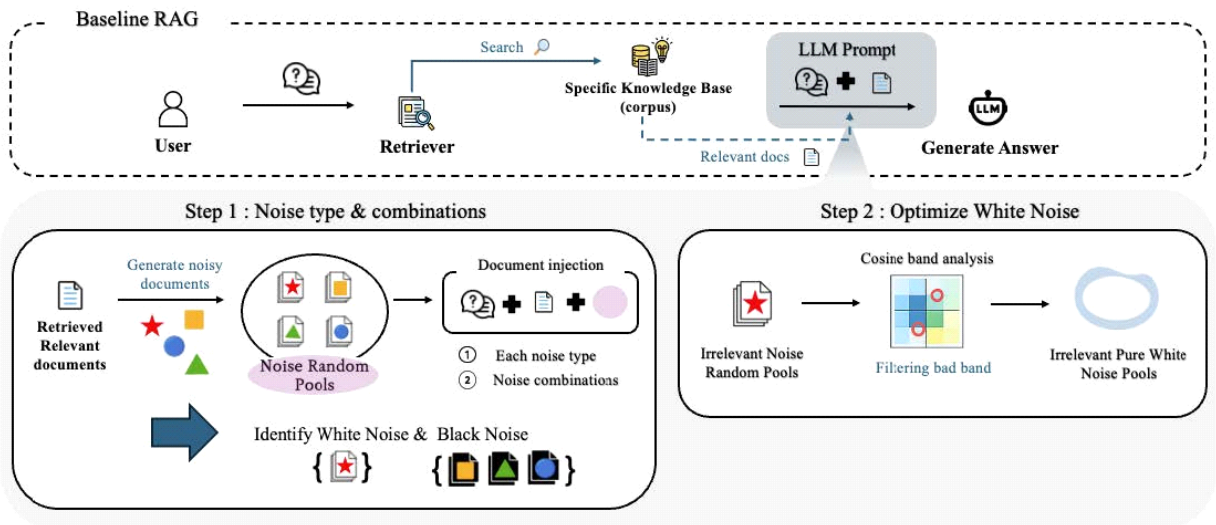


Figure 2. Overall Pipeline Framework of the Proposed Method.

al., 2024a; Zhao *et al.*, 2024a), where irrelevant documents are randomly sampled from the external database, excluding the retrieved relevant documents. However, such random pools may contain other noise types, such as ambiguous documents, making it difficult to ascertain whether observed performance improvements are purely attributable to the irrelevant documents themselves. Furthermore, these impure pools can introduce high variance in performance outcomes.

To address this limitation, Step 2 introduces an optimization procedure to selectively extract Pure White Noise from the documents that demonstrated a positive effect in Step 1. Specifically, we compute the cosine similarity between queries and documents within the irrelevant random noise pool. We then apply a cosine band analysis to filter out bands associated with performance degradation and high variance. This process yields an Irrelevant Pure White Noise pool, free from the interference of other noise types, enabling a stable analysis focused solely on the effects of genuinely irrelevant noise. Through these two steps, our study not only validates the beneficial noise effect in a new domain but also proposes a novel method for selecting Pure White Noise for its stable utilization. This perspective reframes noise from a purely detrimental factor to a strategically exploitable resource, highlighting its potential for building more robust LLM-RAG systems. The specific definitions of each noise document type and the detailed experimental procedures are further elaborated in Sections 3.4 and 4.

3.2 Dataset and Corpora

We employed two types of datasets and their corresponding external corpora to evaluate the domain generalization performance of our proposed RAG-based noise analysis and optimization method. Specifically, we used a domain-specific biomedical QA dataset and an open-domain fact-verification dataset. This dual design enabled us to examine performance variations across distinct domains and tasks, thereby assessing the generalizability of our approach.

(1) Datasets

For the biomedical domain, we adopted PubMedQA (Jin *et al.*, 2019), a representative dataset in which each question is paired with an answer labeled as yes, no, or maybe. Performance was quantitatively assessed using the official 500-sample test set. For the open-domain setting, we used the FEVER dataset (Thorne *et al.*, 2018), a large-scale fact verification benchmark derived from Wikipedia. Each claim is assigned one of three labels: SUPPORTS, REFUTES, or NOT ENOUGH INFO (NEI). To unify the evaluation schema, these labels were mapped to the same three-class format as PubMedQA (yes, no, maybe). This mapping enabled a direct comparison of our method under consistent evaluation criteria. The FEVER development set contains 37,566 claims, all of which were used in our experiments. Importantly, while we retained the original labels, the official evidence annotations were excluded from model inputs, as our primary focus was to analyze how the injection of noisy documents affects answer accuracy. Both datasets were processed under the same preprocessing and retrieval pipelines to ensure comparability. We selected these datasets with explicit class labels to overcome the limitations of open-ended QA, where response variability can hinder objective evaluation. This approach ensures that our performance comparisons are objective and quantitatively reliable. All data processing, including preprocessing and retrieval pipelines, was kept consistent across both tasks.

(2) Corpus

We constructed external corpora aligned with each dataset. The PubMedQA experiments employed a PubMed corpus, whereas the FEVER experiments utilized a Wikipedia corpus. Both served as external knowledge bases for retrieval in the RAG pipeline. The PubMed corpus was derived from PubMed metadata containing abstracts. From approximately 30,000 articles, we retained 15,377 documents, which were segmented into snippets of up to 1,000 characters, yielding 22,456 snippet level entries. The Wikipedia corpus was created from 5,416,537 original articles, segmented into 6,347,967 snippets up to 1,000 characters. To align the corpus scale with PubMed and ensure experimental efficiency, we sampled a random subset of approximately 75,000 snippets for the experiments. All corpora were standardized through the same segmentation, preprocessing, and indexing procedures, ensuring fair comparison across domains. Consequently, both PubMedQA and FEVER tasks were executed using an identical RAG pipeline, with each corpus functioning as a domain-specific external knowledge base.

3.3 Retrieval Methods

(1) Hybrid Retrieval

Traditional information retrieval approaches primarily rely on keyword-based methods (Karpukhin *et al.*, 2020), which identify relevant documents based on the lexical similarity between a query and a document. While this conventional approach offers efficiency and interpretability, they are limited in capturing deeper semantic or conceptual relationships beyond surface-level word overlap (Sawarkar *et al.*, 2024).

To address this limitation, we adopted a hybrid retrieval strategy that combines sparse-based keyword retrieval with dense-based semantic retrieval. This approach aims to leverage both the high precision of keyword-based retrieval and the contextual understanding of semantic search, thereby enabling a more sophisticated and robust document retrieval framework that considers both lexical and semantic relevance.

Specifically, the proposed framework integrates the BM25 model (Robertson *et al.*, 2009) and the FAISS based dense retriever (Johnson *et al.*, 2019) to simultaneously capture both lexical and semantic similarity between a query and candidate documents. The two models independently build indexes on the corpus, and their respective normalized relevance scores are linearly weighted and combined to determine the final retrieval ranking.

(2) Sparse Retrieval based on BM25

For the sparse retrieval component, we employed BM25, a representative probabilistic ranking function widely used in information retrieval. BM25 is an extension of the TF-IDF model that calculates the relevance between a query and a document by normalizing for term frequency and document length.

Recognizing that BM25 performance can be sensitive to term saturation and document length, we optimized its hyperparameters based on empirical evidence from prior research (Viswanathan *et al.*, 2023). The parameter k1, which controls the term frequency saturation, was set to 0.8, and b, which modulates the document length normalization, was set to 0.4. These values were fixed across all experiments to ensure consistency and comparability. To enhance retrieval efficiency, the preprocessed corpus described in Section 3.2 was segmented into approximately 300-token chunks, which served as the retrieval unit for indexing. This chunking strategy improves retrieval granularity and reduces computational overhead, allowing fine-grained document selection while maintaining scalability.

(3) Dense Retrieval based on FAISS

Semantic-based dense retrieval was performed by calculating the semantic similarity between sentence embeddings using FAISS. For this purpose, we utilized a pre-trained Sentence-BERT model, which features a dual-encoder architecture. This model independently encodes the query and the document into dense vectors, and their semantic relevance is then computed as the cosine similarity between these embeddings in the vector space.

Sentence-BERT is trained using contrastive learning (Karpukhin *et al.*, 2020), which aims to minimize the distance between positive pairs (a query and its relevant document) while maximizing the distance between negative pairs (a query and an irrelevant document) in the embedding space. To further enhance the model's discriminative power, we applied a Hard Negative Mining strategy by incorporating incorrectly top-ranked BM25 results as hard negatives during training. This encourages the dense retriever to better distinguish between lexically similar but semantically dissimilar documents, strengthening its ability to capture nuanced semantic relationships. The FAISS index was also constructed from the preprocessed corpus (Section 3.2), with precomputed Sentence-BERT embeddings to enable efficient approximate nearest neighbor (ANN) search over large-scale corpora.

(4) Hybrid Score Optimization

The final hybrid retrieval score S_{hybrid} was computed using a weighted linear combination of the normalized scores from the BM25 and FAISS retrievers as follows:

$$S_{hybrid} = (1-\alpha) \cdot S_{BM25} + \alpha \cdot S_{FAISS} \quad (1)$$

where S_{BM25} and S_{FAISS} represent the min-max normalized scores from BM25 and FAISS, respectively (Kim *et al.*, 2025). The hyperparameter α, which controls the contribution of the dense retriever, was tuned by searching in the range of $0 \leq \alpha \leq 1$ with a step of 0.1. For each value of α, we evaluated the performance of the RAG model using Accuracy@20. The highest accuracy was achieved at $\alpha = 0.3$, which was subsequently used for all experiments.

The proposed hybrid retrieval strategy effectively compensates for the shortcomings of single-model approaches. By combining the lexical precision of sparse retrieval with the semantic coverage of dense retrieval, the hybrid model captures higher-order contextual relevance beyond simple keyword matching. As a result, this approach improves recall of semantically relevant documents while minimizing precision loss from the inclusion of irrelevant ones, demonstrating strong robustness particularly in linguistically diverse domains such as biomedical question answering and fact verification.

3.4 Noise Document Types

To systematically simulate the diverse noise conditions encountered in real-world retrieval environments, we categorize documents into six distinct types. These are grouped into four noisy document types (Irrelevant, Ambiguous, Misleading, and Out-of-Domain) and two reference types (Relevant and Supportive) that constitute the ground-truth set.

While Irrelevant and Ambiguous documents have been com-

monly explored in prior studies on RAG robustness and noise analysis, this work introduces two novel noise categories: Misleading and Out-of-Domain. This expanded taxonomy enables a more comprehensive and realistic evaluation of noise impacts on RAG systems. In particular, the Out-of-Domain category addresses a key limitation of existing research, which often examines noise only within a fixed in-domain corpus. By including Out-of-Domain documents, we can better capture the external noise that naturally arises in open-world retrieval settings.

The detailed definitions of each document type are as follows:

- **Relevant**: Documents retrieved by the RAG system that are semantically pertinent to the given query.
- **Supportive**: The top 50% subset of Relevant documents most semantically aligned with the correct answer, representing the most reliable evidence for the model.
- **Irrelevant**: Documents randomly sampled from the corpus, excluding any Relevant documents, that bear no semantic relation to the query. This type models random retrieval noise.
- **Ambiguous**: Documents that receive high retrieval scores despite low semantic similarity to the query. They often appear superficially related, potentially causing contextual confusion or reasoning drift in the LLM.
- **Misleading**: Documents created by intentionally modifying Supportive documents using a GPT model to reverse correct statements or distort factual content. These documents represent "informationally corrupted" noise containing plausible but incorrect content, which can misguide reasoning.
- **Out-of-Domain**: Documents synthetically generated by a GPT model on topics entirely external to the source corpus (e.g., philosophy, astronomy, or economics). Unlike Irrelevant documents, which are sampled from within the domain, Out-of-Domain documents represent external noise from disparate knowledge sources.

Our novel inclusion of Misleading and Out-of-Domain documents extends the analysis beyond simple lexical or semantic irrelevance. This allows for a deeper investigation of noise arising from factual corruption, domain shift, and semantic confusion. Consequently, our extended taxonomy facilitates a more fine-grained examination of how these heterogeneous noise characteristics affect the performance and reasoning stability of RAG-based large language models.

3.5 Pure White Noise selection

This subsection introduces the Pure White Noise selection strategy to enhance the robustness of the RAG system. This method, constituting Step 2 of our framework, builds upon the Random Irrelevant White Noise pool generated in Step 1. The objective is to selectively leverage White Noise (beneficial noise), documents that contribute positively to model performance, thereby reducing performance variance and achieving

more stable reasoning outcomes.

To systematically identify this beneficial noise, we first computed the cosine similarity for each query-document pair within the Irrelevant document set. We then partitioned the similarity spectrum into 0.1-wide intervals (e.g., 0.0-0.1, 0.1-0.2, ...) to construct cosine bands. Subsequently, we conducted band-wise injection experiments, where documents from each band were exclusively injected into the LLM prompt alongside the relevant documents. We measured the resulting RAG performance using Accuracy, Macro-F1, and Weighted-F1, enabling a fine-grained analysis of performance variation across similarity bands.

As shown in <Figure 3>, this analysis revealed that in both the

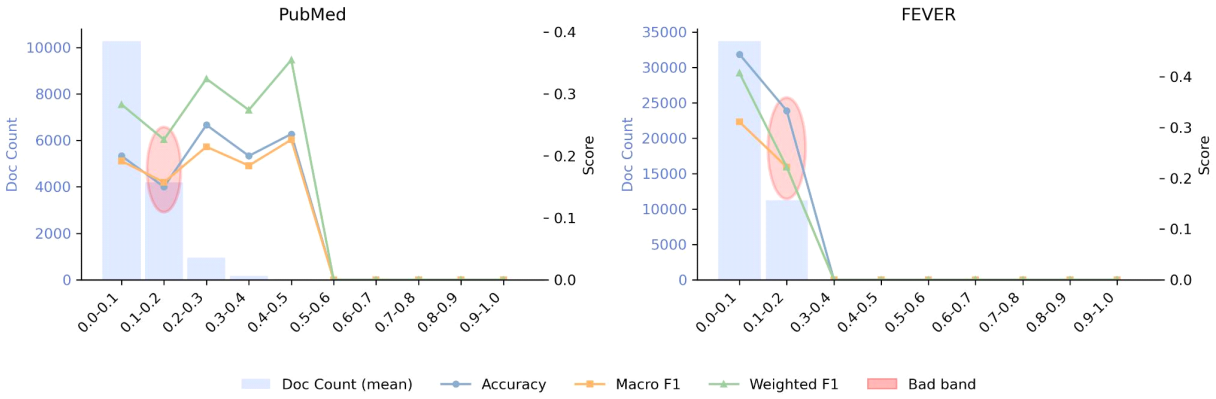


Figure 3. Cosine Band Analysis on PubMed (left) and FEVER (right). The bad band (0.1-0.2) shows the lowest accuracy and F1 scores, indicating performance degradation and increased variance.

PubMed and FEVER datasets, the 0.1-0.2 similarity band consistently exhibited the lowest accuracy and F1 scores, along with the highest performance variance. Consequently, we identified this interval as a "bad band" that adversely affects RAG stability. Accordingly, documents within this bad band were filtered out, and the remaining Irrelevant documents were retained to construct the Pure White Noise pool. Through this filtering process, Step 2 yields a refined noise set compared to the Random Irrelevant Noise pool from Step 1. To validate the efficacy of our strategy, we evaluated three configurations: (1) Baseline (only relevant documents), (2) Random White Noise (the random irrelevant pool from Step 1), and (3) Pure White Noise (the filtered pool from Step 2). This comparative setup empirically validates the proposed filtering mechanism.

One possible explanation for this phenomenon is that documents in the 0.1-0.2 similarity band function as hard distractors for the model. Unlike clearly irrelevant documents with very low similarity (e.g., cosine similarity below 0.1), which share little semantic overlap with the query and can be easily ignored by the model, documents in the 0.1-0.2 range often exhibit partial lexical or topical overlap with the query while lacking genuine evidential relevance. Such weak semantic similarity may create misleading contextual signals that compete with truly relevant evidence during generation. As a result, the model may allocate attention to these partially related but uninformative documents, leading to increased instability in reasoning and performance variance. From a retrieval perspective, this region can be interpreted as a semantic boundary zone between clearly irrelevant and meaningfully relevant documents. Documents located in this intermediate band may appear weakly related in embedding space but fail to provide useful evidence for answering the query, thereby introducing contextual ambiguity rather than beneficial informational diversity.

To further verify the reliability of this cosine-band filtering, we conducted a band-sweep experiment using 500 samples from the PubMedQA and FEVER development sets across five random seeds. As visualized in <Figure 4>, the results show the mean accuracy and standard deviation across similarity bands, consistently confirming that the bad band is located within the 0.1-0.2 range. These findings demonstrate that removing this band mitigates performance degradation and variance, thereby yielding a more stable and robust RAG system compared to the Random White Noise condition.

3.6 Implementation Details

(1) LLM Setup

We utilized the OpenAI GPT-4o-mini model as the primary large language model (LLM) for both answer and document generation. The model was configured to generate responses mapped to three discrete classes: yes, no, and maybe. To quantitatively investigate the underlying causes of performance variations and improvements, we additionally utilized the LLaMA-2-7B model. This model was used to conduct a fine-grained analysis of the attention mechanism and to evaluate the model's contextual reasoning through entropy-based attention analysis.

For the LLM input configuration, both the query and the retrieved relevant documents, along with various noise document types, were jointly injected into the prompt. This design enabled a systematic evaluation of the LLM's performance across different noise conditions. The prompt templates were adapted from formats used in prior RAG studies (Xiong *et al.*, 2024), with minor modifications tailored to the experimental objectives of this work.

(2) Evaluation Metrics

To comprehensively evaluate the performance of the RAG sys-

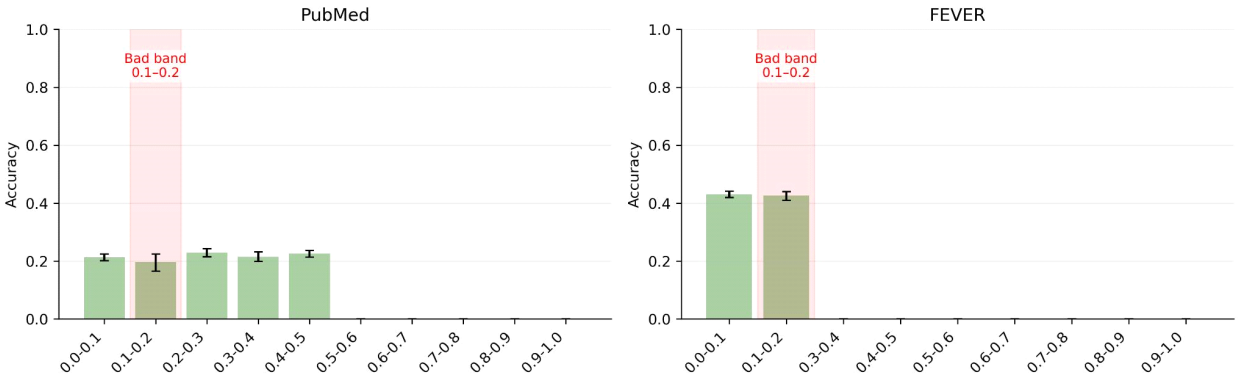


Figure 4. Band-sweep results across five random seeds for PubMed (left) and FEVER (right), validating the reliability of the Pure White Noise selection.

tem, we adopted Accuracy as the primary metric and employed Macro-F1 and Weighted-F1 as complementary measures. To account for the stochastic nature of LLM outputs, we implemented a multi-run evaluation strategy to mitigate response variance. Specifically, 20 queries were randomly sampled from the dataset, and the LLM was executed five times per query, yielding 100 response samples for each experimental condition. For each query, the mode of the five responses was taken as the final prediction, and the average of these mode values across all queries was used to compute the final Accuracy and F1 scores.

In addition to conventional performance metrics, we incorporated attention-based and reasoning-based analytical indicators to elucidate the internal mechanisms driving performance changes. The Attention Entropy metric was employed to quantify how the inclusion of irrelevant documents alters the LLM's attention distribution. In particular, we measured the degree of concentration or dispersion in attention weights when irrelevant documents were injected into the prompt. This analysis was designed to determine whether certain noisy documents could act as White Noise by enhancing the model's focus and reasoning stability.

Conversely, the Contextual Reasoning metric was designed to assess how Black Noise, including Misleading, Ambiguous, and Out-of-Domain documents, disrupts the LLM's reasoning process. To quantify this disruption, we measured two indicators, Answer Flip Rate (AFR) and Maybe Rate. These indicators are defined as:

$$Answer\ Flip\ Rate\ (AFR) = \frac{number\ of\ answer\ flips\ after\ noise\ injection}{number\ of\ correct\ answers\ in\ the\ baseline} \tag{2a}$$

$$Maybe\ Rate = \frac{number\ of\ maybe\ responses\ that\ are\ incorrect}{total\ number\ of\ samples} \tag{2b}$$

The Answer Flip Rate (AFR) captures reasoning instability by measuring how often the model's prediction flips from correct to incorrect after noise is introduced and the Maybe Rate assesses model uncertainty by quantifying the proportion of incorrect answers that the model flags as maybe. Based on these indicators, we provide a quantitative measure of the reasoning instability and inconsistency induced by noise. This multi-faceted evaluation framework allowed us not only to measure performance changes but also to uncover the fundamental influence of noise on the LLM's internal attention mechanisms and reasoning dynamics within the RAG framework.

4. Experiments

4.1 Experimental Design

We designed this study to systematically investigate the impact of noisy documents on the performance of the LLM-RAG framework and to explore how noise can be strategically leveraged as a beneficial factor. To this end, we established two baseline models and structured a two-stage experimental procedure. To ensure a fair and consistent comparison, we maintained an identical retrieval-generation pipeline and evaluation protocol across all experiments.

(1) Baselines

- Baseline RAG : This model represents the standard RAG configuration, which utilizes the top-10 retrieved relevant documents as context for generating responses. It establishes the baseline performance without any injected noise and serves as the control group for our experiments.
- Random White Noise RAG : The Random White Noise RAG extends the baseline configuration by appending randomly sampled irrelevant documents drawn from the noise pool constructed in Step 1. It serves as a comparative benchmark to evaluate the effectiveness of the Pure White Noise approach introduced in Step 2, thereby enabling a direct comparison between unfiltered random White Noise and systematically filtered White Noise.

(2) Experimental Procedures

Our experiments were conducted in two main stages:

- Step 1: Analysis of Noise Types and Combinations

Using the Baseline RAG as a reference, we incrementally introduced four distinct types of noise documents (e.g., Irrelevant, Ambiguous, Misleading, and Out-of-Domain) in quantities ranging from 0 to 20 documents per query. This setup allowed us to measure the marginal effect of each noise type on RAG performance. Subsequently, we conducted combination experiments to analyze the interaction effects when multiple noise types were introduced simultaneously. To further interpret the underlying mechanisms, we measured both Attention Entropy and Contextual Reasoning metrics to examine how different noises influence the model's attention distribution and contextual reasoning capabilities.

- Step 2: Optimization of White Noise

The second stage aimed to validate the potential benefits of irrelevant noise by evaluating the proposed band-based filtering

strategy, which is designed to isolate a refined subset of noise we term Pure White Noise. We compared three models under identical experimental conditions:

1. Baseline RAG : The control model without noise injection
2. Random White Noise RAG : The model augmented with unfiltered, randomly selected irrelevant noise.
3. Pure White Noise RAG : The model augmented with Pure White Noise, refined by our proposed filtering strategy

This comparative evaluation aimed to empirically demonstrate that strategically refined noise can lead to performance stabilization (i.e., variance reduction) and accuracy enhancement in RAG systems.

4.2 Results

(1) Step 1 : Analysis of Noise Types

In this study, we analyzed how different types of noise affect the performance of the Baseline RAG model by progressively increasing the number of noisy documents from 0 to 20. Four different types of noise (e.g., Irrelevant, Ambiguous, Misleading, and Out-of-Domain) were injected individually into the prompt, and the resulting performance variations were systematically evaluated. The results are summarized in <Figure 5>, where the y axis shows the relative score against the baseline. A positive score indicates improvement, while a negative score indicates degradation. Our analysis reveals two distinct patterns of noise effects.

As shown in <Figure 5>, across both the PubMed and FEVER datasets, Irrelevant documents functioned as White Noise, either maintaining or slightly enhancing the performance of the LLM-RAG model. This observation suggests that the inclusion of weakly related information can, in certain cases, improve the model's robustness by promoting more stable reasoning behavior. In contrast, the Ambiguous, Misleading, and Out-of-Domain document types consistently degraded model performance, thereby acting as Black Noise. Among these, Misleading caused the most pronounced decline, as it contains superficially plausible yet factually incorrect information that can easily mislead the model during reasoning.

(2) Step 1 : Analysis of Noise Combinations

In this study, we extended our analysis beyond individual noise types to simulate more realistic retrieval environments with composite noise. To this end, we constructed all possible combinations of the predefined noise categories and systematically evaluated their effects on RAG performance.

As shown in <Figure 6>, the experimental results revealed that combinations including Irrelevant documents consistently outperformed all other noise configurations. This finding confirms that the Irrelevant type robustly acts as White Noise, demonstrating beneficial effects both in isolation and when coexisting with other noise types.

Notably, when Irrelevant documents were combined with Ambiguous or Misleading types, classified as Black Noise, they exhibited a complementary mitigating effect, alleviating the performance degradation typically caused by these harmful noise types. In other words, the presence of Irrelevant documents partially counterbalanced the adverse influence of disruptive information. This stabilizing property of White Noise was con-

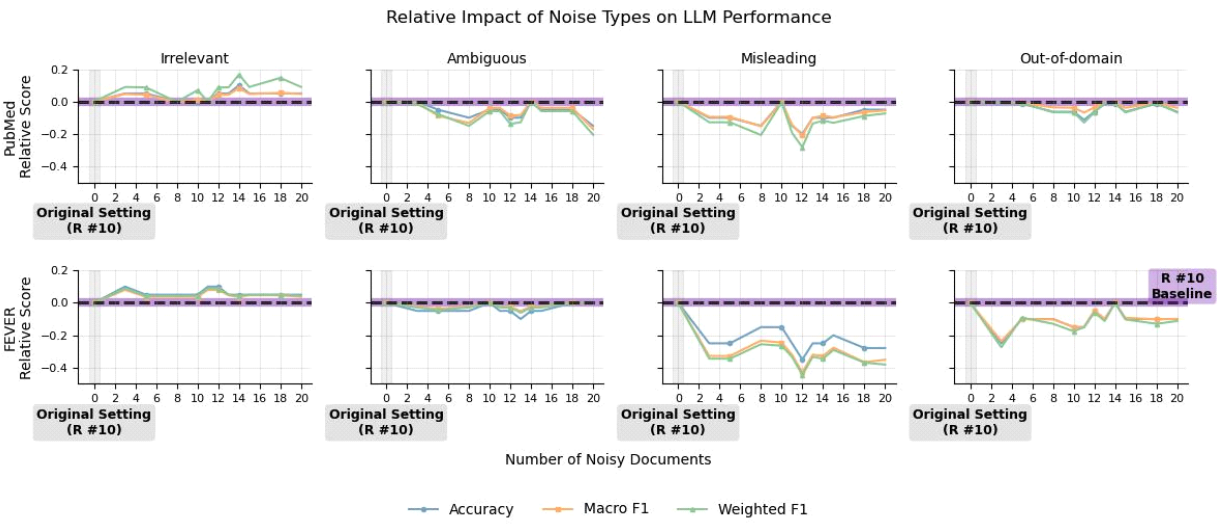


Figure 5. The Relative Impact of Different Noise Types and Quantities on LLM-RAG Performance.

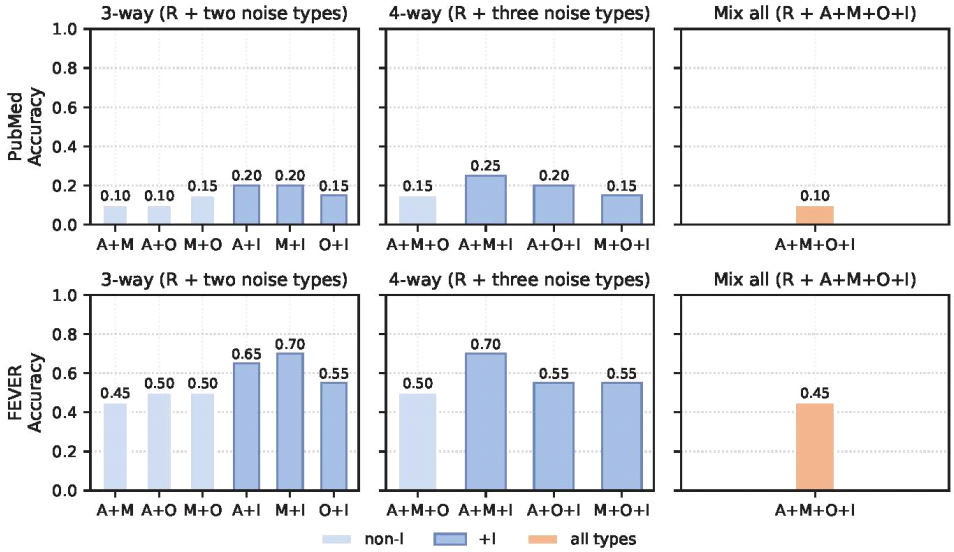


Figure 6. LLM-RAG Performance Accuracy Across Various Noise Document Combinations. The presence of Irrelevant (I) documents (White Noise) consistently enhances model performance against combinations of other noise types (Black Noise). (R: Relevant; A: Ambiguous; M: Misleading; O: Out-of-Domain; I: Irrelevant).

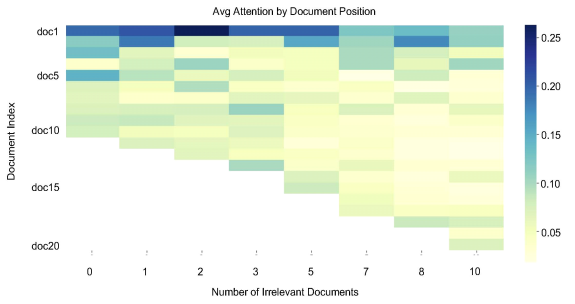
sistently observed across both the PubMed and FEVER datasets, suggesting its beneficial effect is a generalizable phenomenon rather than dataset-specific.

Conversely, when all noise types were injected simultaneously, the model showed the largest decline in performance. We attribute this to information overload and semantic conflicts arising from the high diversity of the documents, which likely disrupted the model's internal reasoning process. These observations highlight the need for future research on the context processing limits of LLM-RAG systems and the development of robust information selection mechanisms to maintain stability under complex noise conditions.

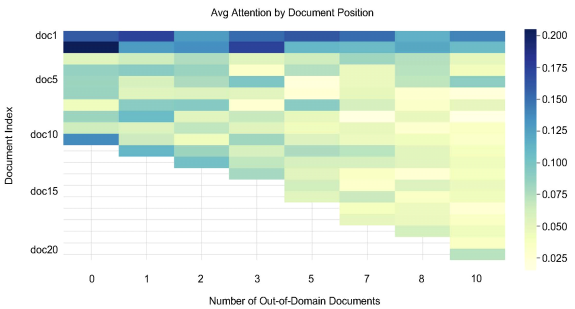
(3) Step 1 : Validation of Noise Effects

To investigate why Irrelevant documents exhibit beneficial effects as White Noise, we analyzed the underlying mechanism through the attention entropy of the LLM. <Figure 7> displays representative attention heatmaps for the White Noise (Irrelevant) and Black Noise (Out-of-Domain) conditions. In these visualizations, documents 1-10 correspond to relevant sources, while documents beyond 11 represent injected noise documents.

Although overall entropy tended to increase as the total number of documents grew in both conditions, the attention distributions diverged significantly. Under the White Noise condition, attention remained concentrated on the initial relevant documents, whereas under the Black Noise condition, attention became broadly dispersed toward the subsequent noise documents. This observation indicates that White Noise enables the model to maintain a stable focus on key evidence (relevant documents), thereby strengthening its grounding for reasoning and ultimately enhancing performance.



(a) White Noise (Irrelevant)



(b) Black Noise (Out-of-Domain)

Figure 7. Average Attention Distribution Across Document Positions for (a) White Noise (Irrelevant) and (b) Black Noise (Out-of-Domain) conditions, as the number of injected noise documents increases. Documents 1-10 represent relevant source documents, while documents after 11 are the injected noise. Darker regions indicate higher attention.

To quantitatively validate these observations, we measured the model's contextual reasoning stability using two metrics: Answer Flip Rate (AFR) and Maybe Rate (MR), as defined in Equations (2a) and (2b). As shown in <Table 1>, the two conditions yielded opposing results. Under the White Noise condition, both AFR and MR decreased as more noise documents were added, relative to the baseline. Conversely, under the Black Noise condition, these metrics consistently increased. These findings clearly demonstrate that Black Noise disrupts the model's contextual reasoning stability, inducing frequent answer reversals and greater uncertainty (i.e., "maybe" responses), ultimately leading to significant performance degradation.

In essence, White Noise appears to function as a form of attentional regularizer, sharpening the model's focus on relevant evidence to enhance reasoning. Conversely, Black Noise acts as a distractant, dispersing attention, which undermines contextual grounding and ultimately degrades performance.

(4) Step 2 : Optimization of White Noise

Building on the beneficial effects of White Noise identified in Step 1, this subsequent experiment, Step 2, explored a more refined approach for leveraging informative noise through the Pure White Noise selection procedure introduced in Section 3.5. We compared three configurations: the Baseline RAG (ten relevant documents only), the Random White Noise (Random WN) RAG (using the Irrelevant pool from Step 1), and the Pure White Noise (Pure WN) RAG (constructed from the Step 2 selection procedure). The results are summarized in <Table 2>.

Table 1. Performance comparison under different noise conditions. Condition indicates the number of noise documents added to the baseline. AFR and MR denote Answer Flip Rate and Maybe Rate, respectively. Flipped represents the absolute count of flipped answers. The arrows (↓, ↑) indicate the general decreasing or increasing trend of the metrics as more noise documents are added.

Condition	AFR	MR		Flipped	
White Noise (Irrelevant)					
Baseline	0.00	0.80		0	
+ 1 Noise	0.05	0.84		1	
+ 2 Noise	0.05	0.85		1	
+ 3 Noise	0.00	0.80		0	
+ 5 Noise	0.05	0.81	↓	1	↑
+ 7 Noise	0.00	0.77		0	
+ 8 Noise	0.00	0.75		0	
+ 10 Noise	0.00	0.75		0	
Black Noise (Out-of-Domain)					
Baseline	0.00	0.71		0	
+ 1 Noise	0.08	0.77		2	
+ 2 Noise	0.04	0.72		1	
+ 3 Noise	0.21	0.80		5	
+ 5 Noise	0.17	0.80	↓	4	↑
+ 7 Noise	0.08	0.78		2	
+ 8 Noise	0.17	0.80		4	
+ 10 Noise	0.25	0.80		6	

Table 2. Random WN vs. Pure WN on PubMed and FEVER. Pure WN consistently reduces σ (bold), indicating improved robustness. Accuracy and Macro-F1 show minor fluctuations between Random WN and Pure WN, but both methods outperform the Baseline across all noise levels, reflecting the beneficial effect of irrelevant noise. ∗ Dramatic reductions in σ are observed for PubMed (q = 5: 0.0106 → 0.0014) and FEVER (q = 4: 0.0125 → 0.0046). Bold in σ highlights Pure WN's lower variance, while ↑ marks improvements over Baseline in Accuracy and Macro-F1.

Dataset	Noise Level	σ		Accuracy		Macro-F1	
		Random WN	Pure WN	Random WN	Pure WN	Random WN	Pure WN
PubMed	Baseline	-		0.2152		0.2035	
	2	0.0055	**0.0054**	0.2380	0.2356	0.2262	0.2235
	3	0.0050	**0.0041**	0.2348	0.2453 ↑	0.2226	0.2322 ↑
	4	0.0081	**0.0071**	0.2380	0.2444 ↑	0.2256	0.2320 ↑
	5	0.0106	**0.0014***	0.2437	0.2400	0.2286	0.2279
FEVER	Baseline	-		0.3944		0.1886	
	2	0.0078	**0.0054**	0.4188	0.4184	0.1968	0.1966
	3	0.0065	**0.0034**	0.4155	0.4147	0.1957	0.1954
	4	0.0125	**0.0046***	0.4148	0.4152 ↑	0.1954	0.1956 ↑
	5	0.0091	**0.0059**	0.4100	0.4108 ↑	0.1938	0.1941 ↑

To visually illustrate these trends, <Figure 8> presents accuracy and variance across the two datasets. In both cases, Pure WN exhibits markedly lower variance than Random WN, while both methods surpass the baseline in accuracy. Although the absolute accuracy gap between the two White Noise methods remains modest, the significant reduction in variance achieved by Pure WN underscores its value as a stability-oriented enhancement.

Overall, these findings demonstrate that the Pure White Noise selection method not only improves model robustness but also effectively mitigates performance fluctuations. Rather than treating noise merely as a detrimental factor, carefully filtering and curating noise can serve as a strategic mechanism for stabilizing and enhancing LLM-RAG performance, offering practical guidance for the design of future noise-aware, robustness-oriented retrieval systems.

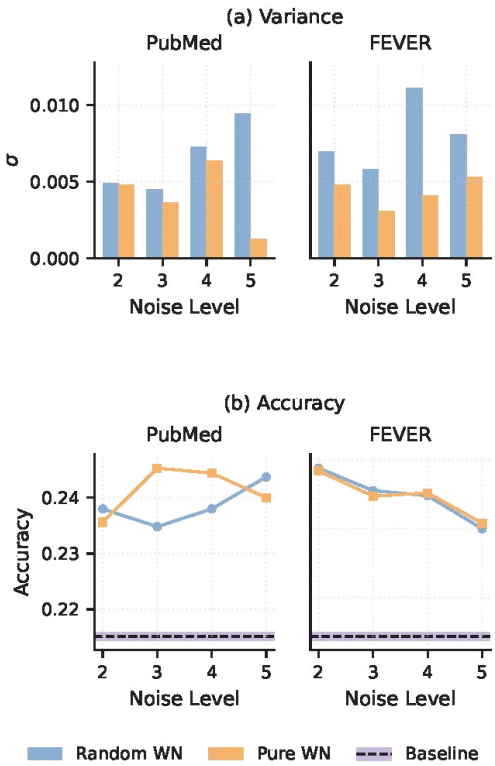


Figure 8. Comparison between the Random White Noise (WN) and Pure White Noise (WN) methods across the PubMed and FEVER datasets. Subfigure (a) illustrates the output variance (σ) , while subfigure (b) presents model accuracy. The results indicate that the Pure WN method consistently achieves markedly lower variance than the Random WN method, demonstrating enhanced stability. Both methods maintain accuracy levels superior to the baseline.

5. Conclusion

This study systematically investigates the impact of various noisy documents on the performance of Retrieval-Augmented Generation (RAG) systems, exploring the potential to strategically leverage such noise for performance enhancement. Our experimental results reveal that Irrelevant documents can be characterized as White Noise; they enhance the model's reasoning stability rather than degrading its performance. Specifically, injecting Irrelevant documents allows the LLM to maintain its focus on the core evidence within relevant documents. In contrast, Black Noise types (e.g., Ambiguous, Misleading, and Out-of-Domain) disrupted the model's contextual reasoning, leading to performance deterioration.

Building on this observation, we propose the Pure White Noise selection method to selectively exploit these beneficial signals. By partitioning documents into predefined cosine bands based on their similarity scores and injecting only those from specific ranges, we identified a subset of noise, the Pure White Noise pool that maintains mean performance while significantly reducing its variance. This finding demonstrates that the robustness of RAG systems can be enhanced not by the passive exclusion of noise, but through the active selection and strategic utilization of beneficial noise.

Ultimately, this work challenges the conventional paradigm that all noise should be eliminated. Instead, we introduce a new perspective that the selective integration of beneficial noise can enhance the robustness and reliability of RAG systems. Our findings contribute both theoretically and practically, theoretically, by redefining the role of noise in retrieval as a distinction between harmful and useful signals; and practically, by proposing a lightweight and efficient solution that stabilizes system behavior through simple similarity-based sampling, without requiring complex filtering mechanisms.

While the proposed framework demonstrates consistent improvements across two benchmark datasets and two LLM configurations, the generalization of the identified similarity bands may depend on the characteristics of the underlying language model. Differences in model scale, training data, reasoning capability, and context-processing architecture may influence how the model interprets partially related documents and allocates attention across retrieved contexts. Therefore, the similarity intervals identified in this study should not be interpreted as universally optimal thresholds across all LLMs, but rather as empirically observed patterns under the experimental settings considered in this work.

Furthermore, the cosine-band analysis identified the 0.1-0.2 similarity interval as a harmful region associated with perform-

ance degradation and increased variance. However, the precise location of such harmful similarity bands may vary depending on dataset characteristics, retrieval distributions, embedding models, and LLM architectures. Future research will extend the current fixed-band selection approach toward an adaptive band selection strategy that dynamically determines optimal noise bands according to dataset characteristics and document distributions. Such an approach aims to establish a noise-adaptive RAG framework capable of maintaining robustness under dynamically changing retrieval conditions in real-world environments. Nevertheless, increasing the number of injected noise documents inevitably expands the input context length, introducing a practical trade-off between robustness improvements and token cost or inference latency in real-world deployments.

References

Abdolazimi, R., Jin, S., Varshney, P. K., and Zafarani, R. (2024), Harnessing the power of noise: A survey of techniques and applications, *arXiv preprint* arXiv:2410.06348.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024), Self-RAG: Learning to retrieve, generate, and critique through self-reflection, *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chang, C. Y., Jiang, Z., Rakesh, V., Pan, M., Yeh, C. C. M., Wang, G., Hu, M., Xu, Z., Zheng, Y., Das, M., and Zou, N. (2025), MAIN-RAG: Multi-agent filtering retrieval-augmented generation, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 2607-2622.

Chen, L., Zhang, R., Guo, J., Fan, Y., and Cheng, X. (2024), Controlling risk of retrieval-augmented generation: A counterfactual prompting framework, *arXiv preprint* arXiv:2409.16146.

Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., and Silvestri, F. (2024a), The power of noise: Redefining retrieval for RAG systems, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 719-729.

Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F., et al. (2024b), Rethinking relevance: How noise and distractors impact retrieval-augmented generation, *CEUR Workshop Proceedings*, CEUR-WS, 95-98.

Fang, F., Bai, Y., Ni, S., Yang, M., Chen, X., and Xu, R. (2024), Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training, *arXiv preprint* arXiv:2405.20978.

Gan, A., Yu, H., Zhang, K., Liu, Q., Yan, W., Huang, Z., Tong, S., and Hu, G. (2025), Retrieval-augmented generation evaluation in the era of large language models: A comprehensive survey, *arXiv preprint* arXiv:2504.14891.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020), Retrieval-augmented language model pre-training, *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 3929-3938.

Islam, S. B., Rahman, M. A., Hossain, K. S. M. T., Hoque, E., Joty, S., and Parvez, M. R. (2024), Open-RAG: Enhanced retrieval-augmented reasoning with open-source large language models, *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14231-14244.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. (2023), Atlas: Few-shot learning with retrieval-augmented language models, *Journal of Machine Learning Research*, **24**, 1-43.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019), PubMedQA: A dataset for biomedical research question answering, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567-2577.

Johnson, J., Douze, M., and Jegou, H. (2019), Billion-scale similarity search with GPUs, *IEEE Transactions on Big Data*, **7**, 535-547.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., Chen, D., and Yih, W. T. (2020), Dense passage retrieval for open-domain question answering, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769-6781.

Kim, S., Song, H., Seo, H., and Kim, H. (2025), Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems, *arXiv preprint* arXiv:2503.15191.

Leto, A., Aguerrebere, C., Bhati, I., Willke, T., Tepper, M., and Vo, V. A. (2024), Toward optimal search and retrieval for RAG, *arXiv preprint* arXiv:2411.07396.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., *et al*. (2020), Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, **33**, 9459-9474.

Ling, Z., Guo, Z., Huang, Y., An, Y., Xiao, S., Lan, J., Zhu, X., and Zheng, B. (2025), MMKB-RAG: A multi-modal knowledge-based retrieval-augmented generation framework, *arXiv preprint* arXiv:2504.10074.

Robertson, S. and Zaragoza, H. (2009), The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends in Information Retrieval*, **3**, 333-389.

Sawarkar, K., Mangal, A., and Solanki, S. R. (2024), Blended RAG: Improving retrieval-augmented generation accuracy with semantic search and hybrid query-based retrievers, *Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 155-161.

Sharma, C. (2025), Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers, *arXiv preprint* arXiv:2506.00054.

Shi, Y., Yang, T., Chen, C., Li, Q., Liu, T., Li, X., and Liu, N. (2025), SearchRAG: Can search engines be helpful for LLM-based medical question answering?, *arXiv preprint* arXiv:2502.13233.

Shim, S. W., Kim, M., Cho, J. H., and Lee, B. J. (2025), Beyond RAG vs. long-context: Learning distraction-aware retrieval for efficient knowledge grounding, *arXiv preprint* arXiv:2509.21865.

Sohn, J., Park, Y., Yoon, C., Park, S., Hwang, H., Sung, M., Kim, H., and Kang, J. (2025), Rationale-guided retrieval augmented generation for medical question answering, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 12739-12753.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018), FEVER: A large-scale dataset for fact extraction and verification, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, 809-819.

Viswanathan, V., Gao, L., Wu, T., Liu, P., and Neubig, G. (2023), DataFinder: Scientific dataset recommendation from natural language descriptions, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 10288-10303.

Wang, Z., Araki, J., Jiang, Z., Parvez, M. R., and Neubig, G. (2023), Learning to filter context for retrieval-augmented generation, *arXiv preprint* arXiv:2311.08377.

Wu, J., Zhang, S., Che, F., Feng, M., Shao, P., and Tao, J. (2025), Pandora's box or Aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 5019-5039.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024), Benchmarking retrieval-augmented generation for medicine, *Findings of the Association for Computational Linguistics: ACL 2024*, 6233-6251.

Yan, S. Q., Gu, J. C., Zhu, Y., and Ling, Z. H. (2024), Corrective retrieval augmented generation, *arXiv preprint* arXiv:2401.15884.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. (2023), Making retrieval-augmented language models robust to irrelevant context, *arXiv preprint* arXiv:2310.01558.

Yu, W., Zhang, H., Pan, X., Ma, K., Wang, H., and Yu, D. (2023), Chain-of-note: Enhancing robustness in retrieval-augmented language models, *arXiv preprint* arXiv:2311.09210.

Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., Shoeybi, M., and Catanzaro, B. (2024), RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs, *Advances in Neural Information Processing Systems*, **37**, 121156-121184.

Zhang, Q., Zhang, H., Pang, L., Zheng, H., Tong, Y., and Zheng, Z. (2025), FineFilter: A fine-grained noise filtering mechanism for retrieval-augmented large language models, *arXiv preprint* arXiv:2502.11811.

Zhao, S., Huang, Y., Song, J., Wang, Z., Wan, C., and Ma, L. (2024a), Towards understanding retrieval accuracy and prompt quality in RAG systems, *arXiv preprint* arXiv:2411.19463.

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., and Qiu, L. (2024b), Retrieval-augmented generation (RAG) and beyond: A comprehensive survey on how to make your LLMs use external data more wisely, *arXiv preprint* arXiv:2409.14924.

Zhu, K., Feng, X., Du, X., Gu, Y., Yu, W., Wang, H., Chen, Q., Chu, Z., Chen, J., and Qin, B. (2024), An information bottleneck perspective for effective noise filtering on retrieval-augmented generation, *arXiv preprint* arXiv:2406.01549.

Author Profile

Dayoung Kim: She received her B.S. degree from Stony Brook University and is currently pursuing an M.S. degree. Her research interests include Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs).

Young Myoung Ko: He received his B.S. and M.S. degrees in Industrial Engineering from Seoul National University in 1998 and 2000, respectively, and his Ph.D. degree in Industrial Engineering from Texas A&M University in 2011. He is currently a Professor in the Department of Industrial and Management Engineering at Pohang University of Science and Technology. He serves as an Associate Editor of IISE Transactions and as a Vice Chair of the INFORMS Telecommunications and Network Analytics Section.