Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) for domain-specific question-answering (QA) tasks by leveraging external knowledge sources. However, traditional RAG systems primarily focus on relevance-based retrieval and often struggle with redundancy, especially when reasoning requires connecting information from multiple sources. This paper introduces **Vendi-RAG**, a framework based on an iterative process that jointly optimizes retrieval diversity and answer quality. This joint optimization leads to significantly higher accuracy for multi-hop QA tasks. Vendi-RAG leverages the Vendi Score (VS), a flexible similarity-based diversity metric, to promote semantic diversity in document retrieval. It then uses an LLM judge that evaluates candidate answers, generated after a reasoning step, and outputs a score that the retriever uses to balance relevance and diversity among the retrieved documents during each iteration. Experiments on three challenging datasets—HotpotQA, MuSiQue, and 2WikiMultiHopQA—demonstrate Vendi-RAG's effectiveness in multi-hop reasoning tasks. The framework achieves significant accuracy improvements over traditional single-step or multi-step RAG approaches, with accuracy increases reaching +4.2% on HotpotQA, +4.1% on 2WikiMultiHopQA, and +1.3% on MuSiQue compared to Adaptive-RAG, the current best baseline. The benefits of Vendi-RAG are even more pronounced as the number of retrieved documents increases. Finally, we evaluated Vendi-RAG across different LLM backbones, including GPT-3.5, GPT-4, and GPT-4o-mini, and observed consistent improvements, demonstrating that the framework's advantages are model-agnostic.

1 Introduction

Retrieval-augmented generation (RAG) has emerged as a transformative framework for enhancing the performance of large language models (LLMs) in domain-specific tasks such as question-answering (QA). By retrieving relevant information from external sources beyond the training set, RAG enables LLMs to answer specialized queries more effectively (Achiam et al., 2023; Team et al., 2023; Jiang et al., 2024). This approach has been particularly successful in single-hop QA, where a question can be answered using information from a single document (Raiaan et al., 2024; Kwiatkowski et al., 2019). For instance, answering a question such as "Who wrote the novel Frankenstein?" only requires retrieving relevant information from a single document containing this fact.

However, multi-hop QA introduces significantly greater complexity. Finding the correct answer to queries in multi-hop QA requires reasoning across multiple sources (Press et al., 2022; Tang and Yang, 2024). For instance, answering "Which city is the capital of the African country where Mount Kilimanjaro is located?" necessitates first identifying that Mount Kilimanjaro is in Tanzania, and then determining that Dodoma is the capital of Tanzania. This process involves not only retrieving information from multiple documents but also synthesizing these different sources effectively to form an accurate answer, which greatly increases the complexity of both retrieval and reasoning and leads to redundancy.

To address these challenges, iterative RAG pipelines have been developed. These pipelines refine the retrieval process through repeated modifications and re-querying of retrieved documents, aiming to resolve ambiguities and improve relevance. Notable examples include Adaptive-RAG (Lewis et al., 2020), which controls the number of iterations of the pipeline including the retrieval process and modifying the queries based on a classification model's assessment of the input query, Self-RAG(Asai et al., 2023), which incorporates iterative self-reasoning, and IROC (Trivedi et al., 2022),

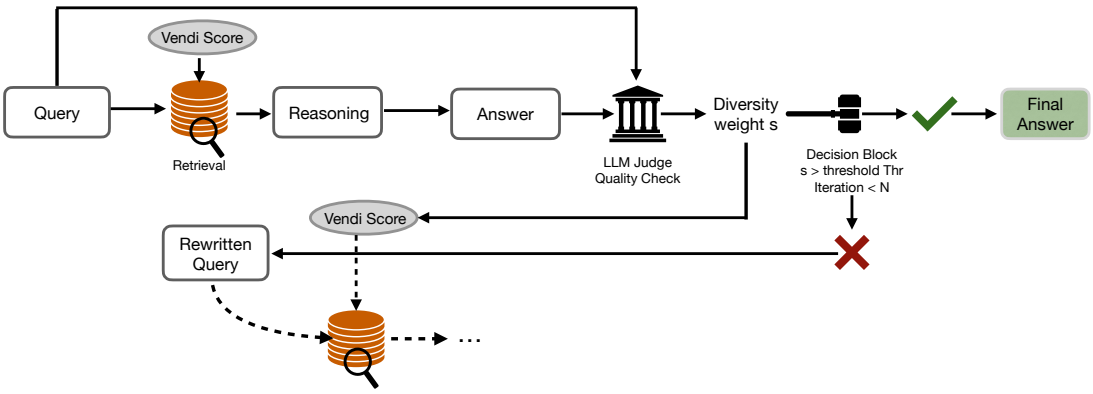

Figure 1: The process begins with an initial retrieval step, where a diverse set of documents is retrieved using the Vendi Score, ensuring broad semantic coverage. Next, leveraging a reasoning step to construct a coherent path to the final answer, the LLM generates an answer, which then undergoes quality assessment by an LLM judge. Based on the answer quality, the retriever is adjusted to balance diversity and relevance: high-quality answers limit the emphasis on diversity, while low-quality answers prompt the retriever to prioritize diversity more heavily. This adjustment is controlled by an adaptive parameter, s, which is updated over iterations. The process continues until the answer quality reaches an optimal threshold, denoted by Thr. Finally, the highest-quality responses and documents are selected, ensuring both diversity and accuracy.

which progressively refines retrieval to optimize the final answer (Wei et al., 2022; Wang and Zhou, 2024).

Despite their success, iterative RAG methods typically rely solely on relevance-based retrieval, which focuses on the similarity between the query and dataset entries. This approach presents a fundamental limitation: it does not actively manage the diversity and quality of the retrieved information to properly address the query. More complex queries require diverse retrieval. We therefore propose a novel retrieval method called *Vendi retrieval* to address the limitation of existing retrieval pipelines. Vendi retrieval leverages the Vendi Score (VS) to enhance the diversity of retrieved documents while accounting for retrieval quality through a simple weighting mechanism.

Building on Vendi retrieval, we propose an iterative RAG pipeline called Vendi-RAG that balances diversity and quality. More specifically, the pipeline is as follows: an initial set of candidate documents is retrieved. Based on these retrieved documents, the system generates chain-of-thought (CoT) reasoning steps. Using these reasoning steps and retrieved documents, the LLM then generates candidate answers. An LLM-based evaluator then assesses these candidates for relevance, coherence, and completeness. The highest-scoring answer is selected as the final response. If the answer does not meet the quality threshold, the Vendi retrieval process dynamically adjusts the balance between diversity and relevance in document selection, ensuring broader semantic exploration or increased specificity as needed. This iterative refinement continues until a high-quality response is achieved. Figure 1 provides a detailed overview of the Vendi-RAG framework.

We evaluated the Vendi retrieval process and Vendi-RAG on three challenging multi-hop QA datasets, HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020). To assess the Vendi retrieval method we measured the diversity of retrieved documents on the three datasets using two different diversity metrics, the VS and the max pairwise distance (MPD). We found that the Vendi retrieval process yields more diverse documents compared to the baselines according to both metrics. Second, we evaluated Vendi-RAG in terms of several performance metrics, looking at both accuracy and diversity. The results showed that Vendi-RAG substantially improves response accuracy, outperforming existing RAG approaches. Using GPT-3.5 as the LLM backbone, Vendi-RAG demonstrated significant accuracy gains across all datasets, with accuracy increases reaching +4.2% on HotpotQA, +4.1% on 2WikiMultiHopQA, and +1.3% on MuSiQue compared to Adaptive-RAG, the best baseline. Notably, the accuracy improvement remained consistent across different LLM backbones—GPT-4o, GPT-4o-mini, and GPT-3.5—indicating that Vendi-RAG's advantages are model-agnostic. Additionally, our experiments with varying numbers of retrieved documents—beyond the standard two-document setting—showed that Vendi-RAG maintained its superior performance,

especially as the number of retrieved documents increased. This underscores the critical role of the Vendi retrieval process in handling complex retrieval scenarios. For instance, when retrieving ten documents from HotpotQA, Vendi-RAG outperformed Adaptive-RAG by 7.8% in accuracy using GPT-4o-mini as the backbone LLM.

This work introduces a diversity-guided retrieval approach that optimizes both diversity and quality to address the challenges of multi-step reasoning in multi-hop QA. Our experimental results highlight the effectiveness of Vendi-RAG in enhancing retrieval diversity and response accuracy, underscoring its potential as a robust solution for complex multi-hop QA tasks.

2 Related Work

There are three main approaches to QA: non-retrieval-based methods (Petroni et al., 2019), single-step RAG (Lewis et al., 2020), and multi-step RAG (Asai et al., 2023). Non-retrieval-based QA methods pass queries directly to an LLM and use its generated output as the answer, without consulting external sources. While efficient, these methods struggle with queries requiring external or up-to-date information and suffer from hallucinations on out-of-distribution queries (Shuster et al., 2021). Single-step RAG methods integrate external knowledge retrieved from a knowledge base (e.g., Wikipedia). These methods improve factual accuracy but are limited by retrieval noise and perform poorly in complex reasoning tasks (Trivedi et al., 2022). Multi-step RAG methods are designed for complex multi-hop queries (Jeong et al., 2024; Asai et al., 2023; Tang and Yang, 2024). They iteratively retrieve documents and refine answers until they converge on a final response. This iterative refinement approach enables reasoning across multiple sources but introduces computational overhead and is prone to error accumulation (Jeong et al., 2024).

Advances in multi-hop QA. Recent improvements in multi-hop QA focus on question decomposition (Radhakrishnan et al., 2023), chain-of-thought reasoning (Wei et al., 2022; Liu et al., 2024a), and iterative retrieval (Jeong et al., 2024; Shao et al., 2023; Yu et al., 2024). Methods like ReCite (Sun et al., 2022) and IRCoT (Trivedi et al., 2022) refine retrieval with progressive reasoning, while Self-RAG (Asai et al., 2023) adapts retrieval strategies based on query complexity. Decomposed prompting (Khot et al., 2022) further enhances retrieval for complex queries (Zhang et al., 2024). MultiHop-RAG (Tang and Yang, 2024) integrates decomposition and retrieval pipelines but remains constrained by relevance-based retrieval, leading to redundancy and limited synthesis of diverse information.

Vendi scoring. The Vendi Score (VS) (Friedman and Dieng, 2023) is a similarity-based diversity metric applied in machine learning (Berns et al., 2023; Pasarkar and Dieng, 2023; Mousavi and Khalili, 2025; Nguyen and Dieng, 2024; Kannen et al., 2024; Jalali et al., 2024; Askari Hemmat et al., 2024; Rezaei and Dieng, 2025; Bhardwaj et al., 2025), chemistry (Pasarkar et al., 2023), materials science (Liu et al., 2024b), and biology (Pasarkar and Dieng, 2025). Vendi-RAG integrates VS into retrieval, balancing diversity and quality beyond conventional ranking systems (Carbonell and Goldstein, 1998; Slivkins et al., 2010). Unlike standard relevance-based retrieval (Guu et al., 2020), this approach enhances robustness and accuracy in multi-hop QA by incorporating semantic diversity into document retrieval.

3 Method

We now describe Vendi-RAG, including the novel retrieval process it uses.

3.1 Vendi Retrieval

Diversity in retrieved documents is essential for multi-hop QA, as it ensures broad semantic coverage, reduces redundancy, and incorporates multiple perspectives (Sun et al., 2022; Carbonell and Goldstein, 1998; Thakur et al., 2021). The most commonly used techniques for information retrieval are similarity search (SS)(Thakur et al., 2021) and maximal marginal relevance (MMR)(Carbonell and Goldstein, 1998). While SS retrieves documents based on their relevance to the query, it often results in redundant documents with high similarity. MMR attempts to balance relevance and novelty using pairwise comparisons, but it still struggles to capture global semantic diversity.

To address these limitations, we adopt a retrieval approach based on the Vendi Score (VS) (Friedman and Dieng, 2023), which explicitly quantifies semantic diversity in a set of documents. Let $\mathcal{D} = d_1, \ldots, d_n$ be a set of retrieved documents, and let $k(\cdot, \cdot)$ be a positive semi-definite similarity kernel such that $k(d_i, d_i) = 1$ for all i. Let K be the similarity matrix with entries $K_{ij} = k(d_i, d_j)$.

The Vendi Score is defined as:

$$\mathrm{VS}_k(\mathcal{D}) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right), \quad (1)$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the normalized kernel matrix K/n. As shown by Friedman and Dieng (2023), $\mathrm{VS}_k(\mathcal{D})$ reflects the effective number of unique documents in $\mathcal{D}$, attaining its maximum value n when all documents are orthogonal (fully diverse) and its minimum value 1 when all documents are identical.

While optimizing for diversity is important—especially for complex, multi-faceted queries—it must be balanced with query relevance. To achieve this, we define the Vendi Retrieval Score (VRS) as a convex combination of semantic diversity and similarity-based relevance:

$$\mathrm{VRS} = s \cdot \mathrm{VS}_k(\mathcal{D}) + (1 - s) \cdot \mathrm{SS}(q, \mathcal{D}), \quad (2)$$

where $s \in [0, 1]$ is a tunable parameter that controls the trade-off between diversity and relevance. The similarity score $\mathrm{SS}(q, \mathcal{D})$ is computed using dense vector representations of the query q and the documents in $\mathcal{D}$, typically obtained from transformer-based encoders. This ensures semantic matching beyond surface-level lexical overlap.

It is important to note that while the Vendi Score $\mathrm{VS}_k(\mathcal{D})$ is computed solely based on the retrieved documents and their pairwise similarities, query relevance is introduced in the initial retrieval step: the candidate set $\mathcal{D}$ is selected using similarity search with respect to the query q. Thus, the formulation in Equation (2) balances document-level diversity and query-level relevance, where a higher value of s favors diverse content, and a lower value prioritizes semantic similarity to the query. In this way, the VSR addresses the dual objectives of reducing redundancy and maintaining relevance, providing a principled and flexible framework for multi-hop document selection.

3.2 Vendi-RAG

We integrate the Vendi retrieval process into a flexible RAG pipeline that balances diversity and relevance for improved performance on multi-hop QA.

1. Initial retrieval. The process begins by retrieving a set of documents using Vendi retrieval. This first step prioritizes broad semantic coverage (we set $s = 0.8$ initially in all our experiments), ensuring that the retrieved documents capture multiple perspectives and to prevent recovering semantically redundant documents. This initial diversity is particularly critical for multi-hop QA, where synthesizing information from varied sources is essential to accurately answering the query.

2. Reasoning generation. Based on the retrieved documents, the system generates CoT reasoning steps. These intermediate reasoning steps help contextualize the retrieved information, building a coherent pathway to the final answer.

3. Candidate answer generation. Using the reasoning steps and retrieved documents, the LLM generates candidate answers. These proposed answers are evaluated to determine their quality and completeness.

4. Quality evaluation. An LLM judge assesses the candidate answers. This evaluation considers factors such as coherence, relevance, and alignment with the query. A quality score Q_t is produced at the end of this quality-check. Here t is used to indicate the iteration step.

5. Dynamic adjustment of the VRS. Based on the quality score Q_t, the parameter s is adjusted dynamically. We denote by s_t the value of the parameter s at the t^{th} iteration. It controls the trade-off between diversity (via VS) and relevance (via similarity search). If Q_t is low, s_t should be increased, to prioritize greater diversity in the subsequent retrievals. This ensures broader semantic exploration, which is beneficial for refining answers in cases where the retrieved information is already relevant but lacks coverage. Conversely, if Q_t is high, s_t should be decreased to focus more on relevance, retrieving documents that are closely aligned with the query to address potential gaps in specificity. We therefore define s_t as

$$s_t = f(Q_{t-1}) = 1 - \frac{Q_{t-1}}{\max(Q_{t-1})}, \quad (3)$$

where f is a simple linear function that maps Q_{t-1} to the interval $[0, 1]$, ensuring that higher quality scores correspond to lower diversity scores.

6. Iterative refinement. The retrieval and reasoning steps are repeated iteratively, with adjustments to s dynamically balancing diversity and relevance at each stage. This process continues until the desired answer quality is reached, ensuring that the system converges on an optimal set of documents and reasoning steps.

7. Final answer selection. Once the iterative refinement process is complete, the final set of documents and answers are selected based on their quality scores. This ensures that the output reflects both broad semantic coverage and high-quality, relevant information. Algorithm 1 summarizes the procedure.

Why Adjusting s Matters: The dynamic adjustment of s is essential for balancing diversity and relevance during retrieval. High diversity enables exploration of different facets of complex queries, especially in multi-hop QA, where synthesizing information from multiple sources is crucial. However, too much diversity can introduce noise, while excessive focus on relevance risks redundancy and limits comprehensive reasoning.

Vendi-RAG addresses this by adapting s based on retrieval quality: when the quality score is high, it reduces s to promote exploration of additional, semantically diverse documents; when quality is low, it increases s to prioritize more directly relevant documents. This adaptive retrieval strategy allows Vendi-RAG to dynamically adjust to the needs of each query and reasoning step, improving both the breadth and precision of generated answers. Unlike traditional RAG systems with fixed retrieval policies, Vendi-RAG's flexibility ensures richer, more contextually appropriate responses.

Performance characteristics. In practice, Vendi-RAG exhibits distinctive performance patterns that reflect its sophisticated design. The system naturally adapts its computational effort to query complexity, requiring more iterations for intricate multi-hop queries while converging quickly for simpler ones. Though the computational overhead exceeds that of basic RAG systems, the improved retrieval quality often results in better final answers. The system maintains reasonable scalability with document corpus size, as the primary computational bottleneck—eigenvalue computation—depends on the number of retrieved documents rather than the total corpus size. These characteristics make Vendi-RAG particularly suitable for complex tasks such as multi-hop QA.

4 Experiments

In this section, we present a comprehensive evaluation of Vendi-RAG on multi-hop QA tasks. We begin by analyzing the effectiveness of the Vendi retrieval strategy in enhancing retrieval diversity. We then evaluate the full Vendi-RAG pipeline, highlighting its ability to handle complex queries requiring multi-step reasoning, and compare its performance against several strong baselines. All experiments are conducted on three challenging multi-hop QA benchmark datasets: MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and 2WikiMultiHopQA (Ho et al., 2020) (see Appendix A for additional dataset details).

Algorithm 1 Vendi-RAG Inference Pipeline

Require: Query q, Knowledge base $\mathcal{D}$, Max iterations N, Quality threshold τ
Ensure: Final answer $\hat{a}^*$
  Initialize context: $q_1 \leftarrow q$, set initial parameter: $s_1 \leftarrow 0.8$
  for $i = 1$ to N **do**
    $\text{VRS}_i \leftarrow s_i \cdot \text{VS}_k(q_i, \mathcal{D}) + (1 - s_i) \cdot \text{SS}(q_i, \mathcal{D})$
    $D_i \leftarrow \text{Vendi-Retrieval}(q_i, \text{scores}_i; \mathcal{D})$
    Generate reasoning steps: $r_i \leftarrow \text{CoT}(q_i, D_i)$
    Produce answer: $\hat{a}_i \leftarrow \text{LLM}(q, D_i, r_{1:i})$
    if LLM-Judge$(\hat{a}_i) \geq \tau$ **then**
      return $\hat{a}_i$
    end if
    Update query: $q_{i+1} \leftarrow \text{RewriteQuery}(q_i, \hat{a}_i, r_i)$
    Update weight parameter: $s_{i+1} \leftarrow f(Q_i)$
  end for
  return $\hat{a}_N$

Dataset	Method	VS	MPD
MuSiQue	Adaptive Retrieval	6.13	1.25
	Adaptive-RAG	6.55	1.42
	Vendi-RAG	**7.12**	**1.95**
HotpotQA	Adaptive Retrieval	4.95	1.10
	Adaptive-RAG	5.21	1.31
	Vendi-RAG	**6.82**	**1.89**
2WikiMHQA	Adaptive Retrieval	5.34	1.32
	Adaptive-RAG	5.81	1.45
	Vendi-RAG	**6.68**	**1.78**

Table 1: Retrieval diversity (Vendi Score (VS) and Max Pairwise Distance (MPD)) across datasets and methods. Vendi-RAG achieves higher diversity than baselines.

Sensitivity Analysis of the VSR Process. To evaluate the robustness of the VSR process and understand its impact on retrieval diversity, we conducted a sensitivity analysis focusing on how varying the parameter s affects document ranking order within a vector database. This analysis helps elucidate the trade-off between retrieval precision and diversity, which is crucial for enhancing multi-hop reasoning performance. The sensitivity analysis was performed using 100 randomly sampled queries from the dataset to ensure a comprehensive evaluation covering a diverse range of query types and complexity levels. Our primary objective was to investigate how different values of s influence

Parameter (s)	τ	ρ
0.0	1.00	1.00
0.2	0.797	0.828
0.4	0.688	0.742
0.6	0.485	0.528
0.8	0.265	0.316
1.0	0.074	0.078

Table 2: Sensitivity Analysis of the VSR. Higher s values indicate a greater degree of diversity introduced in the ranking by the retrieval process.

retrieval diversity and ranking consistency.

We evaluated the retrieval pipeline across multiple s values ranging from 0.0 to 1.0, incremented in small steps to capture granular variations in retrieval performance. Setting $s = 0.0$ serves as a baseline representing a pure similarity search scenario, where retrieval relies exclusively on cosine similarity or dot product between embeddings, without any emphasis on diversity. This baseline provides a reference point for measuring the impact of increasing s on retrieval diversity. To quantify deviations from the baseline, we employed two complementary ranking comparison metrics:

- **Kendall's τ:** Measures the rank order similarity between two lists by evaluating concordant and discordant pairs. Higher τ values indicate stronger similarity to the baseline, while lower values reflect greater diversity introduced by increasing s.

- **Spearman's Rank Correlation ρ:** Assesses the monotonic relationship between two rankings by considering both orderings and positional shifts. Lower ρ values signal substantial deviation from the baseline, indicating increased diversity through higher s values.

The results of the sensitivity analysis are presented in Table 2. As s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively, demonstrating that higher s values promote retrieval diversity by prioritizing documents that may be less similar in their embeddings but more relevant from a broader perspective.

Vendi retrieval improves document retrieval diversity. To assess the impact of the Vendi retrieval process on retrieval diversity, we compared the diversity of outputs from Vendi-RAG against Adaptive-RAG and Adaptive Retrieval. We measured diversity using two different metrics, the VS and the max pairwise distance (MPD). Table 1 summarizes the results. Vendi-RAG achieves higher diversity compared to Adaptive Retrieval and Adaptive-RAG on all dataset, demonstrating its ability to retrieve documents that capture multiple perspectives relevant to the query. This is a crucial advancement, as increased diversity in retrieval directly correlates with improved robustness in multi-hop reasoning (see Table 3). Adaptive-RAG, which incorporates iterative refinement but lacks explicit diversity control, shows moderate retrieval diversity improvement over Adaptive Retrieval.

Accuracy on multi-hop QA tasks. We further evaluated the performance of the Vendi-RAG pipeline to assess its ability to reason across multiple documents. The results in Table 3 indicate that Vendi-RAG consistently outperforms other methods in response accuracy across all datasets, showcasing the efficacy of balancing retrieval diversity with quality. Additionally, Vendi-RAG achieves competitive performance in Exact Match (EM) and F1 score. These findings highlight Vendi-RAG's capability to enhance answer correctness for complex reasoning tasks through improved document retrieval. Additionally, we conducted a comprehensive ablation study comparing various retrieval strategies, including Vendi-RAG and MMR, as well as a dynamic adjustment of s_t against fixed s_t in the table. The results demonstrate that Vendi-RAG with dynamically adjusted s_t (with an initial setting of $s_1 = 0.8$) with the LLM-Judge consistently outperforms all baselines, including MMR, across the datasets in terms of EM, F1, and Acc.

Ablation Study on Retrieval Strategy and s_t Scheduling. To better understand the effectiveness of our retrieval strategy, we conducted a comprehensive ablation study, examining various configurations of the Vendi-RAG pipeline. First, we evaluated fixed values of s_t across all steps, testing settings such as $s_t = \{0.0, 0.3, 0.8, 1.0\}$. Among the fixed schedules, setting $s_t = 0.8$ consistently achieved the best overall performance in terms of Exact Match (EM), F1 score, and Accuracy across all datasets, highlighting the importance of dynamically balancing retrieval diversity with quality. We further compared Vendi-RAG against traditional retrieval baselines, including Adaptive Retrieval with MMR, and observed that our method out-

Methods	MuSiQue			HotpotQA			2WikiMultiHopQA		
	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
No Retrieval	20.40	31.30	24.40	37.40	51.04	43.20	37.00	48.50	43.40
Single-step Approach	16.40	26.70	23.60	39.60	50.44	45.60	46.80	57.69	52.60
Adaptive Retrieval	18.80	30.30	24.80	38.60	50.70	43.20	44.20	55.11	50.60
LightRAG	17.40	27.51	23.80	37.40	47.30	42.30	42.21	52.24	47.50
Adaptive-RAG	21.80	32.60	29.60	40.40	52.56	47.00	46.60	**60.09**	56.80
Self-RAG	1.20	8.20	11.80	5.60	17.86	30.60	3.00	19.14	39.00
Adaptive Retrieval with MMR	22.6	31.0	28.8	41.0	55.0	56.0	45.8	57.0	59.0
Graph-RAG	22.9	32.5	30.1	42.4	57.2	57.8	47.5	59.1	60.9
FiD-Reranker	23.2	**33.1**	29.8	41.8	56.7	56.9	46.6	59.3	60.1
Vendi-RAG(fixed-$s_{1:N} = 0.0$)	22.4	30.4	28.2	40.6	54.1	55.4	45.4	56.8	58.8
Vendi-RAG*(fixed-$s_{1:N} = 0.3$)	22.4	31.1	28.6	40.7	55.0	55.7	45.3	56.8	59.1
Vendi-RAG(fixed-$s_{1:N} = 0.8$)	22.6	30.4	29.1	41.2	55.7	57.2	46.4	57.4	60.2
Vendi-RAG(fixed-$s_{1:N} = 0.8$)	23.0	31.2	30.2	42.0	56.9	58.0	47.0	58.7	61.0
Vendi-RAG(fixed-$s_{1:N} = 1.0$)	22.9	31.0	29.4	41.6	55.8	57.0	46.4	57.8	60.2
Vendi-RAG($s_1 = 0.8$)	**24.4**	32.52	**30.4**	**42.2**	**57.04**	**58.4**	**47.2**	58.9	**61.4**

Table 3: Performance on multi-hop QA datasets using GPT-3.5 Turbo is evaluated across three metrics: exact match (EM), F1 score (F1), and traditional accuracy (Acc). Vendi-RAG with $s_1 = 0.8$ outperforms all baselines across the three datasets in terms of EM and Acc, while achieving comparable F1 scores to Adaptive-RAG. Here, Vendi-RAG* refers to the variant of Vendi-RAG that excludes the LLM-Judge component.

performed the MMR retrieval method across all metrics. Additionally, we tested a variant without dynamic scheduling (fixed s_t) and a variant without the LLM-Judge module (Vendi-RAG*). The results show that dynamically adjusting s_t during retrieval using the LLM-Judge significantly boosts performance compared to fixed schedules and simpler retrieval strategies. These findings emphasize the critical role of adaptive retrieval and document assessment in enabling Vendi-RAG to effectively handle complex multi-hop reasoning tasks.

Impact of the number of retrieved documents on performance. To further examine the impact of document size on retrieval effectiveness, we analyze the performance of Vendi-RAG and Adaptive-RAG across varying document sizes and initial settings of $s_1 = \{0.3, 0.8, 1.0\}$. Figure 2 illustrates the relationship between document size and performance on the HotPotQA dataset. Vendi-RAG consistently outperforms Adaptive-RAG in accuracy for document sizes greater than two with any s_1 setting. As document size increases, accuracy improves for both methods, but the gain is notably higher for Vendi-RAG. Similar to accuracy, EM and F1 scores exhibit an increasing trend as document size grows. Vendi-RAG shows a more pronounced improvement, underscoring its capacity to retrieve more informative and relevant documents, thereby enhancing answer quality. The VS also increases with document size. This is evidence that Vendi-RAG alleviates redundancy since the VS is known to be invariant under duplication (Friedman and Dieng, 2023). An increasing VS indicates less redundancy in the retrieved documents. By leveraging the VS in its retrieval process, Vendi-RAG avoids the redundancy issue that often plagues RAG pipelines. These results indicate that increasing document size enhances both retrieval diversity and answer correctness. Vendi-RAG is achieving superior gains in all metrics. However, we are computationally bottlenecked primarily by the LLM's context window limitation and processing time. As the number of retrieved documents increases, we must either truncate documents to fit within the model's maximum context length or process documents in multiple batches, both of which have significant computational overhead.

Performance for different LLM Backbones and retrieval strategies. To evaluate the impact of different LLM backbones and retrieval strategies on the performance of the Vendi-RAG (with $s_1 = 0.8$) framework, we conducted experiments using three LLMs: GPT-4o, GPT-4o-mini, and GPT-3.5, across all the multi-hop QA datasets described above. The results, shown in Figure 3, highlight the effectiveness of Vendi-RAG compared to Adaptive-RAG, the best baseline, across all datasets and LLM-backbones, except for F1-score on the 2WikiMultiHopQA dataset. In general, larger LLM backbones, such as GPT-4o, achieve superior perfor-

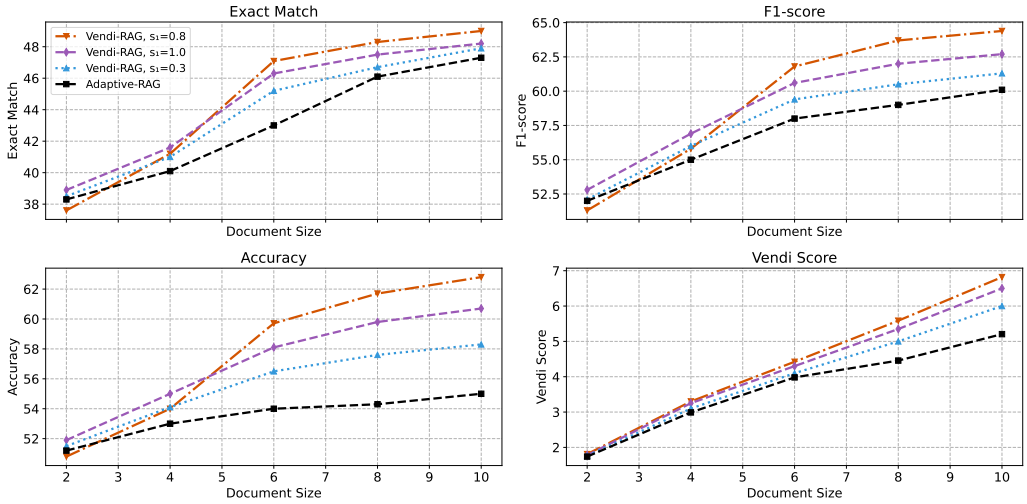

Figure 2: Performance comparison of Vendi-RAG and Adaptive-RAG across different document sizes in terms of Exact Match, F1-score, Accuracy, and Vendi Score on HotPotQA. The backbone LLM used is `GPT-4o-mini`. Vendi-RAG consistently outperforms Adaptive-RAG across all metrics. In particular, performance improves as the number of retrieved documents increases. Different variants of Vendi-RAG are plotted based on the fixed initialization value s_1 for the diversity-relevance parameter s_t, with $s_1 = 0.8$ achieving the best overall results.

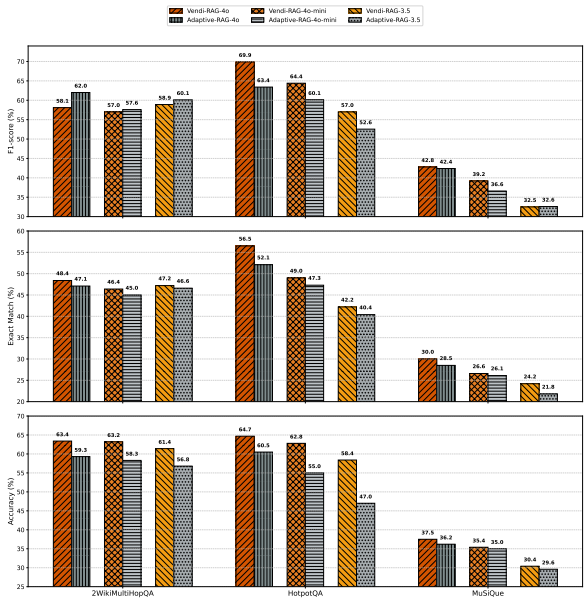

Figure 3: Performance comparison of Vendi-RAG and Adaptive-RAG variants across the three datasets using three evaluation metrics: F1-score, Exact Match, and Accuracy. Results show that Vendi-RAG-4o consistently outperforms other variants across all datasets and metrics, with a particularly strong performance on HotpotQA.

mance across all datasets, particularly for tasks requiring complex reasoning and synthesis across multiple documents. However, even with smaller models like GPT-4o-mini, the Vendi-RAG model maintains competitive performance, demonstrating its effectiveness.

VSR vs MMR. VS offers key advantages over traditional diversity metrics like MMR. While MMR relies on pairwise comparisons to balance relevance and novelty, it lacks a global view of semantic diversity across the retrieved set. In contrast, VS is a principled, global metric based on the eigenvalues of the normalized kernel matrix, directly measuring the effective number of distinct documents. It reaches its maximum when documents are entirely unique and its minimum when they are identical, providing an intuitive, mathematically grounded measure of diversity. This global perspective makes VS particularly effective for multi-hop QA, where broad semantic coverage is critical. Moreover, VS integrates naturally with Vendi-RAG's dynamic retrieval adjustment, enabling fine-grained control over the diversity-relevance trade-off via a single parameter, and addressing the challenge of balancing coverage and precision in complex reasoning tasks.

5 Conclusion

While RAG has proven effective in enhancing LLM performance for domain-specific QA tasks, traditional models often struggle with redundancy, particularly in multi-hop reasoning tasks. To address this shortcoming, we introduce Vendi-RAG, a novel framework that jointly optimizes retrieval diversity and answer quality through an iterative refinement process. Vendi-RAG leverages the Vendi Score and an LLM judge to promote semantic diversity while maintaining relevance during retrieval.

6 Limitations

Vendi-RAG introduces computational overhead due to LLM-based quality scoring, which may limit scalability in real-time applications. Additionally, like all RAG approaches, its performance depends on the quality and completeness of external knowledge sources, making it susceptible to biases or gaps in the retrieved information.

7 Ethics Statement

The deployment of LLMs, including their use in Vendi-RAG, necessitates careful ethical consideration. Since the model relies on external knowledge sources, concerns arise regarding the credibility and accuracy of retrieved content. Ensuring the reliability and factual integrity of information is crucial to mitigating risks related to bias and misinformation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. 2024. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pages 213–229. Springer.

Sebastian Berns, Simon Colton, and Christian Guckelsberger. 2023. Towards Mode Balancing of Generative Models via Diversity Weights. *arXiv preprint*. ArXiv:2304.11961 [cs.LG].

Utkarsh Bhardwaj, Vinayak Mishra, Suman Mondal, and Manoj Warrier. 2025. A robust machine learned interatomic potential for nb: Collision cascade simulations with accurate defect configurations. *arXiv preprint arXiv:2502.03126*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Dan Friedman and Adji Bousso Dieng. 2023. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. 2024. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models. *arXiv preprint arXiv:2411.02817*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jingyu Liu, Jiaen Lin, and Yong Liu. 2024a. How much can rag help the reasoning of llm? *arXiv preprint arXiv:2410.02338*.

Tsung-Wei Liu, Quan Nguyen, Adji Bousso Dieng, and Diego A Gómez-Gualdrón. 2024b. Diversity-driven, efficient exploration of a mof design space to optimize mof properties. *Chemical Science*, 15(45):18903–18919.

Mohsen Mousavi and Nasser Khalili. 2025. Vsi: An interpretable bayesian feature ranking method based on vendi score. *Available at SSRN 4924208*.

Quan Nguyen and Adji Bousso Dieng. 2024. Quality-Weighted Vendi Scores And Their Application To Diverse Experimental Design. In *International Conference on Machine Learning*.

Amey P Pasarkar, Gianluca M Bencomo, Simon Olsson, and Adji Bousso Dieng. 2023. Vendi sampling for molecular simulations: Diversity as a force for faster convergence and better exploration. *The Journal of chemical physics*, 159(14).

Amey P Pasarkar and Adji Bousso Dieng. 2023. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*.

Amey P. Pasarkar and Adji Bousso Dieng. 2025. The vendiscope: An algorithmic microscope for data collections. *arXiv preprint arXiv:2502.04593*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Mohammad Reza Rezaei and Adji Bousso Dieng. 2025. The $alpha$-alternator: Dynamic adaptation to varying noise levels in sequences using the vendi score for improved robustness and performance. *arXiv preprint arXiv:2502.04593*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. 2010. Learning optimally diverse rankings over large document collections. In *Proc. of the 27th International Conference on Machine Learning (ICML 2010)*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chunliang Yang, Rosalind Potts, and David R Shanks. 2018. Enhancing learning and retrieval of new information: a review of the forward testing effect. *NPJ science of learning*, 3(1):8.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Auto-rag: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*.

Zhihao Zhang, Alan Zhu, Lijie Yang, Yihua Xu, Lanting Li, Phitchaya Mangpo Phothilimthana, and Zhihao Jia. 2024. Accelerating retrieval-augmented language model serving with speculation. *arXiv preprint arXiv:2401.14021*.

A Datasets

MuSiQue evaluates a model's ability to synthesize facts spread across across multiple document sources. It includes questions spanning diverse domains such as history, science, and culture, requiring logical reasoning and synthesis of interdependent information. Given its emphasis on multi-step comprehension, this dataset challenges models to accurately identify and integrate relevant information to generate correct answers to queries. This is the most challenging dataset among the three.

HotpotQA assesses reasoning and fact verification across various domains, including geography, entertainment, and history. Its questions necessitate reasoning over two or more interconnected documents linked via hyperlinks. Additionally, the dataset includes "comparison" questions that require juxtaposing information from multiple sources, making it a challenging benchmark for evaluating both retrieval quality and reasoning ability.

2WikiMultiHopQA leverages Wikipedia's complex structure to pose complex reasoning challenges. Questions are derived from real-world knowledge graphs and require navigating reasoning paths across multiple documents. Topics span science, politics, and sports, emphasizing logical relationships such as cause-effect dependencies, making it an essential tool for evaluating structured knowledge reasoning.

B Evaluation Metrics

To compare model performance across different datasets, we employ the following key evaluation metrics:

- **Exact Match (EM)**: This metric calculates the percentage of predictions that exactly match the ground truth answers. It is defined as:

$$\text{EM} = \frac{\text{Number of exact matches}}{\text{Total number of queries}} \times 100 \quad (4)$$

  EM is a strict metric that grants credit only when the predicted answer matches the ground truth exactly in both content and format. It is particularly useful for assessing a model's precision in generating accurate responses.

- **F1 Score (F1)**: The F1 score captures the harmonic mean of precision and recall at the token level, providing a balanced measure of correctness. It is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

  where precision is the fraction of retrieved tokens that are relevant, and recall is the fraction of relevant tokens that are retrieved. The F1 score is particularly relevant for multi-hop QA tasks, where partial correctness (e.g., retrieving some but not all supporting evidence) is informative.

- **Accuracy (Acc)**: Accuracy measures the proportion of correct predictions over all evaluated queries. It is defined as:

$$\text{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of queries}} \times 100 \quad (6)$$

  Unlike EM, which requires exact matches, accuracy provides a broader assessment by capturing overall correctness, including responses that convey the intended meaning.

- **Max Pairwise Distance (MPD)**: This metric evaluates the maximum Euclidean distance between pairs of retrieved data points, measuring diversity. It is defined as:

$$\text{MPD} = \max_{i,j} \|x_i - x_j\|_2, \quad i \neq j \quad (7)$$

  where x_i and x_j represent document embeddings in the feature space. Higher values indicate greater diversity among retrieved documents.

Each of these metrics offers a unique perspective on model performance. EM is a stringent measure of precision, F1 balances precision and recall, and accuracy provides an overall correctness measure. Meanwhile, MPD and diversity-based metrics assess the variety and independence of retrieved documents, critical for multi-hop QA tasks requiring integration of diverse information.

C Implementation Details

The Vendi-RAG framework employs dense vector representations derived from transformer-based embeddings to compute the similarity score $\text{SS}(q, \mathcal{D})$ between a query q and a set of documents $\mathcal{D}$. This high-dimensional semantic comparison allows the

model to effectively capture contextual relationships and retrieve relevant documents across diverse domains. To evaluate answer quality, we use a consistent LLM backbone with a specifically designed prompt that positions the LLM as an expert judge. This judge-based evaluation framework assesses the generated answers according to this prompt:

LLM as a Judge Prompt

You are an expert LLM-based judge tasked with evaluating the quality of answers in a Retrieval-Augmented Generation (RAG) system. Your evaluation will consider the following aspects:

1. **Coherence**: Assess whether the provided answer is logically consistent and flows smoothly, without conflicting statements or gaps in reasoning.
2. **Relevance**: Evaluate how well the answer addresses the query based on the information from the retrieved documents.
3. **Query Alignment**: Determine how closely the answer aligns with the specific query asked, ensuring that the response is focused and appropriate.

Your evaluation will be quantified based on the following scoring system:

- Coherence Score (C): [1 - 10], where 10 is perfectly coherent.
- Relevance Score (R): [1 - 10], where 10 is highly relevant to the query.
- Query Alignment Score (Q): [1 - 10], where 10 is perfectly aligned.

Provide a quality score Q_t as the average of these individual scores:

$$Q_t = \text{mean}(C, R, Q)$$

Query: {query}

This judge assesses coherence, relevance, and alignment with the query to produce a quality score Q_t. The quality threshold (τ) is set to 0.85 for all experiments, ensuring a consistent standard of answer evaluation. While we initially set $s_1 = 0.8$ across experiments to prioritize diversity, our ablation study explores this hyperparameter through testing different fixed values and employing dynamic adjustment. This analysis highlights the importance of dynamic adjustment for improved performance across diverse datasets.

C.1 Dataset Processing and Chunking

Preparing datasets for question-answering requires transforming data into a searchable vector database to enable efficient retrieval. This workflow includes document chunking and semantic embedding to optimize performance.

The dataset, provided in JSON format with context paragraphs and metadata, is processed by splitting each document into smaller chunks. Each chunk has a maximum size of 512 tokens, with a 50-token overlap to preserve context across chunk boundaries and facilitate multi-hop reasoning in long documents.

C.2 Embedding Model and Vector Database

We use the **SentenceTransformer** model, specifically `all-mpnet-base-v2`, to generate dense vector representations for documents and queries. These embeddings are stored locally to avoid redundant downloads and improve reusability. The **Chroma** vector database efficiently stores and retrieves these vectorized documents along with metadata, such as document titles and chunk IDs.

C.3 Batch Processing and Database Population

To efficiently populate the vector database, document chunks are processed in batches of up to 10,000. This approach optimizes memory usage while ensuring completeness in the ingestion process. The total number of processed chunks is logged for verification.

C.4 Query Answering Workflow

For queries such as *"Who is the father-in-law of Queen Hyojeong?"*, relevant chunks are retrieved using **Chroma**'s similarity-based search mechanism. The system ranks the top 10 chunks based on their semantic similarity to the query, leveraging embeddings generated by `all-mpnet-base-v2` to ensure precise and relevant results.

C.5 Key Configuration Details

The system is configured with the following parameters:

- **Embedding Model:** `all-mpnet-base-v2`, optimized for semantic similarity tasks.
- **Vector Database:** Chroma, persisted to disk for efficient reuse.

- **Chunk Size:** 512 tokens per chunk, with a 50-token overlap.
- **Batch Size:** Up to 10,000 chunks per batch.