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

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) for multi-hop question answering is limited by retrieving semantically redundant documents, when reasoning instead needs diverse, complementary sources. We introduce **Vendi-RAG**, which treats set-level semantic diversity as a first-class retrieval objective. It combines an eigenvalue-based diversity score (the Vendi Score) with query relevance into a single Vendi Retrieval Score, embedded in an iterative loop that uses an LLM-based answer-quality signal to adaptively rebalance diversity and relevance. On HotpotQA, MuSiQue, and 2WikiMultiHopQA with GPT-3.5, Vendi-RAG improves accuracy by $+2.4$, $+1.6$, and $+2.4$ points over the strongest diversity-aware iterative baseline (Adaptive Retrieval + MMR), with paired-bootstrap $p < 0.05$. A fixed-s variant captures most of the gain at no extra inference cost, while the LLM-judge loop adds a smaller but significant gain on hard queries. Results hold across five instruction-tuned backbones and two dense retrievers, identifying set-level retrieval diversity as a primary, controllable axis for improving multi-hop RAG.

1 Introduction

Retrieval-augmented generation (RAG) enhances large language models (LLMs) on knowledge-intensive tasks such as question answering (QA) by retrieving relevant information from external sources beyond the training set (Achiam et al., 2023; Team et al., 2023; Jiang et al., 2024). It has been especially successful in single-hop QA, where a question can be answered from a single document (Raiaan et al., 2024; Kwiatkowski et al., 2019). Multi-hop queries, however, require synthesizing information from several distinct sources. For example, answering *"What is the capital of the country where the inventor of the telephone was born?"* requires combining biographical facts about Bell's birthplace with geographical facts about Scotland's capital (Table 1), which exposes a limitation of relevance-based retrieval.

Method	Retrieved Documents	Answer
Standard RAG	1. "A.G. Bell invented the telephone..." [tel]	✗
	2. "The telephone was patented by Bell..." [tel]	
	3. "Bell's telephone revolutionized..." [tel]	
	4. "First telephone call: 'Mr. Watson...'" [tel]	
	5. "Bell demonstrated the telephone..." [tel]	
MMR	1. "A.G. Bell invented the telephone..." [tel]	∿
	2. "Edinburgh is in Scotland, UK..." [geo]	
	3. "Bell was a teacher of the deaf..." [bio]	
	4. "Scottish inventors: J.L. Baird..." [inv]	
	5. "UK: England, Scotland, Wales..." [geo]	
Vendi-RAG	1. "Bell born in Edinburgh, Scotland..." [bio]	✓
	2. "Edinburgh: Scotland's capital..." [geo]	
	3. "Bell's telephone patent..." [tel]	
	4. "Scottish education influenced Bell..." [edu]	
	5. "UK capitals: Edinburgh (Scotland)..." [geo]	

Table 1: Illustrative retrieval comparison for "What is the capital of the country where the inventor of the telephone was born?" Relevance-only retrieval returns redundant telephone documents, whereas Vendi-RAG retrieves complementary, query-relevant documents across categories.

The core problem is that relevance-based retrieval tends to return semantically similar documents with redundant content, while multi-hop reasoning needs diverse, complementary sources that supply the different links of the reasoning chain. Table 1 illustrates this point.

Iterative RAG pipelines partly address this by re-querying and refining retrieved documents over multiple rounds. Adaptive-RAG (Jeong et al., 2024) routes queries to no-retrieval, single-step, or iterative retrieval based on a query-complexity classifier; Self-RAG (Asai et al., 2023) adds on-demand retrieval with self-reflection; and IR-CoT (Trivedi et al., 2023) interleaves retrieval with chain-of-thought reasoning (Wei et al., 2022; Wang and Zhou, 2024). Despite their success, these methods rely on relevance-based retrieval and do not

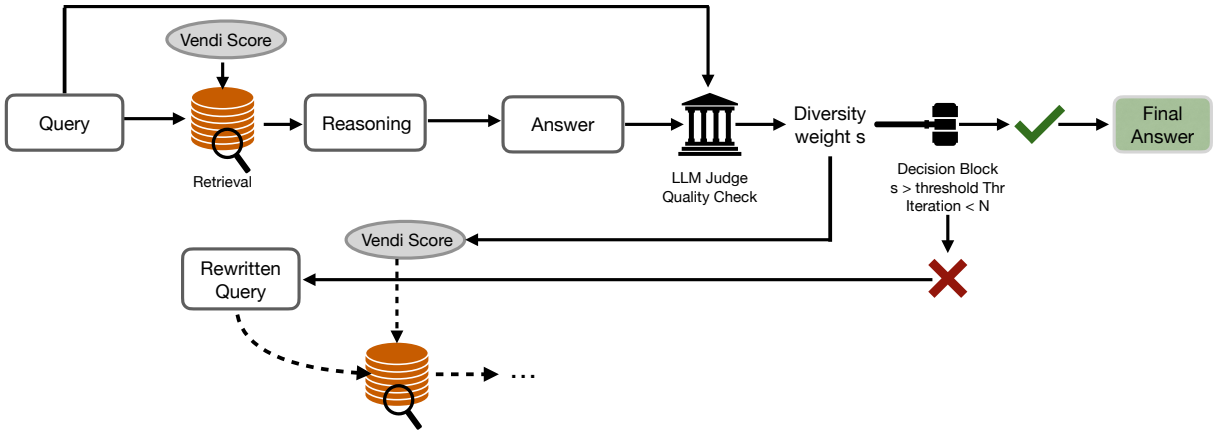

Figure 1: Vendi-RAG begins with an initial retrieval that uses the Vendi Score to obtain a diverse, broadly covering document set. The LLM then reasons over these documents to produce an answer, which an LLM judge scores for quality. Based on this score, the retriever rebalances diversity and relevance via an adaptive parameter s: high-quality answers reduce the emphasis on diversity, while low-quality answers increase it. The loop iterates until quality reaches the threshold Thr, after which the highest-quality answer and documents are returned.

actively manage the *diversity* of retrieved information, which is precisely what complex queries demand.

We therefore propose *Vendi retrieval*, which uses the Vendi Score (VS) to promote diversity among retrieved documents while accounting for relevance through a simple weighting mechanism. Building on it, we introduce **Vendi-RAG**, an iterative pipeline that balances diversity and quality: it retrieves candidate documents, generates chain-of-thought reasoning and candidate answers, scores them with an LLM judge for coherence, relevance, and query alignment, and—if the quality threshold is not met—dynamically rebalances diversity and relevance before retrieving again. The highest-scoring answer across iterations is returned (Figure 1).

The contributions of this paper are:

1. A *set-level* retrieval objective for multi-hop QA: the Vendi Retrieval Score combines query relevance with the Vendi Score, an eigenvalue-based measure of the effective number of unique documents. This differs from pairwise objectives (MMR), log-determinant objectives (k-DPP), and post-retrieval utility selectors (Jain and Garimella, 2025; Zhang et al., 2025), and exposes a single scalar s that trades diversity against relevance.
2. An *explore–exploit* feedback loop that adapts s to per-query difficulty using an LLM-based answer-quality signal, with EMA smoothing for stability. We show that the loop helps even when the judge is a different or smaller model than the generator, so it requires only a *relative* quality signal rather than absolute calibration.
3. A systematic empirical study on three multi-hop benchmarks (HotpotQA, MuSiQue, 2WikiMultiHopQA) and five LLM backbones (GPT-3.5, GPT-4o, GPT-4o-mini, Llama 3-70B-Instruct, Gemini Flash 2.0). Vendi-RAG improves multi-hop accuracy over Adaptive-RAG and over diversity-aware iterative baselines (MMR, k-DPP, IRCoT) under matched retrievers and test splits, and the improvement is robust to the choice of dense encoder.

We evaluate Vendi retrieval and Vendi-RAG on three multi-hop QA benchmarks: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020). Vendi retrieval yields more diverse documents than the baselines on both the Vendi Score and max pairwise distance, and that diversity translates into higher coverage of the gold supporting evidence. The most informative downstream comparison is to other diversity-aware iterative methods: with GPT-3.5, Vendi-RAG improves accuracy over the strongest such baseline (Adaptive Retrieval + MMR) by $+2.4$ on HotpotQA, $+1.6$ on MuSiQue, and $+2.4$ on 2WikiMultiHopQA, with paired-bootstrap $p < 0.05$. Against routing-based Adaptive-RAG, the gap is $+11.4$ on HotpotQA, $+0.8$ on MuSiQue, and $+4.6$ on 2WikiMultiHopQA; the larger HotpotQA gap reflects Adaptive-RAG's complexity-classifier routing of many HotpotQA queries to no-retrieval or single-step paths in our setup, where iterative diversity-aware methods (MMR, IRCoT,

k-DPP, Vendi-RAG) all sit in the same 55–59 accuracy band. A fixed-s variant of Vendi-RAG captures most of the improvement at the same inference cost as MMR; the LLM-judge loop adds a smaller but significant gain at higher latency. The pattern is consistent across five instruction-tuned backbones (GPT-3.5, GPT-4o, GPT-4o-mini, Llama 3-70B-Instruct, Gemini Flash 2.0) and two dense retrievers on multi-hop English QA.

2 Related Work

QA approaches fall into three families: non-retrieval methods (Petroni et al., 2019), single-step RAG (Lewis et al., 2020), and multi-step RAG (Asai et al., 2023). Non-retrieval methods answer directly from the LLM; they are efficient but struggle with external or up-to-date information and hallucinate on out-of-distribution queries (Shuster et al., 2021). Single-step RAG improves factual accuracy but is limited by retrieval noise on complex reasoning (Trivedi et al., 2023). Multi-step RAG targets multi-hop queries by iteratively retrieving and refining (Jeong et al., 2024; Asai et al., 2023; Tang and Yang, 2024), enabling cross-source reasoning at the cost of overhead and error accumulation (Jeong et al., 2024).

Advances in multi-hop QA. Recent work focuses 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). ReCite (Sun et al., 2022) and IR-CoT (Trivedi et al., 2023) refine retrieval with progressive reasoning, while Self-RAG (Asai et al., 2023) uses self-reflection tokens to decide when to retrieve. Decomposed prompting (Khot et al., 2022) aids complex queries (Zhang et al., 2024), and MultiHop-RAG (Tang and Yang, 2024) combines decomposition with retrieval but remains constrained by relevance-based retrieval, limiting synthesis of diverse information.

Diversity in information retrieval. MMR (Carbonell and Goldstein, 1998) pioneered the relevance-diversity trade-off via greedy pairwise-dissimilarity penalties. Intent-aware methods such as xQuAD (Santos et al., 2010) and PM-2 (Dang and Croft, 2012) model query aspects but require predefined intent taxonomies that may not generalize to open-ended multi-hop queries. Determinantal Point Processes (Kulesza and Taskar, 2012), applied to summarization (Gong et al., 2014) and recommendation (Chen et al., 2018), model set-level diversity but scale as $O(n^3)$ for exact inference and optimize a log-determinant rather than the interpretable "effective number of items" of the Vendi Score. Submodular optimization (Lin and Bilmes, 2011; Wei et al., 2014) offers greedy approximation guarantees but typically needs explicit features and may miss the semantic similarities encoded in dense embeddings.

Vendi retrieval differs in that (1) the Vendi Score directly measures the effective number of distinct documents, making the trade-off transparent; (2) it operates on kernel matrices from dense embeddings without explicit intents or features; (3) the eigenvalue computation runs only on the small retrieved set ($k \leq 10$); and (4) its dynamic adjustment adapts the diversity-relevance balance from answer-quality feedback, unlike static diversification.

Vendi scoring. The Vendi Score (VS) (Friedman and Dieng, 2023) is a similarity-based diversity metric used across 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). Unlike pairwise (MMR) or log-determinant (DPP) metrics, VS computes the exponential of the Shannon entropy of the kernel eigenvalues, an interpretable measure equal to the effective number of orthogonal components. Vendi-RAG integrates VS into retrieval to balance diversity and quality (Carbonell and Goldstein, 1998; Slivkins et al., 2010), improving robustness and accuracy over relevance-only retrieval (Guu et al., 2020) in multi-hop QA.

Multi-hop retrieval planning. Some systems decompose queries and schedule retrieval steps explicitly (Khattab et al., 2022; Press et al., 2023), and entity-centric approaches (de Cao et al., 2021) track entities across hops. These operate at the retrieval-loop level and are composable with Vendi-RAG's diversity-aware scoring function.

Context-independent passage utility. A second line learns a *utility* signal on top of relevance: Jain and Garimella (2025) supervise a passage selector with LLM reasoning traces, and Zhang et al. (2025)

distill an LLM into a utility-based selector over a relevance-retrieved pool. These act *post-retrieval* as filters, whereas Vendi retrieval acts earlier, shaping the candidate pool itself. The two are complementary: a diverse candidate set is a prerequisite for utility-based filtering to recover complementary multi-hop evidence.

3 Method

We describe Vendi-RAG, our framework for multi-hop QA based on diversity-driven document retrieval.

3.1 Vendi Retrieval

Diversity in retrieved documents is essential for multi-hop QA: it ensures broad semantic coverage, reduces redundancy, and brings in multiple perspectives (Sun et al., 2022; Carbonell and Goldstein, 1998; Thakur et al., 2021). Similarity search retrieves on relevance alone and returns redundant documents. Three families of methods address this. MMR (Carbonell and Goldstein, 1998) balances relevance and novelty through *pairwise* dissimilarity penalties but lacks a notion of set-level diversity. k-DPP (Kulesza and Taskar, 2012) samples subsets in proportion to the determinant of a kernel submatrix, but the log-determinant objective scales poorly and is not directly interpretable as "number of unique documents". Post-retrieval utility selectors (Jain and Garimella, 2025; Zhang et al., 2025) learn passage-level usefulness on top of a fixed relevance-ranked pool; they cannot recover useful evidence that is absent from that pool. We instead define diversity at the set level using the Vendi Score (VS) (Friedman and Dieng, 2023), which quantifies the effective number of unique documents in a set $\mathcal{D} = \{d_1, \ldots, d_n\}$:

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

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the normalized similarity matrix K/n, with $K_{ij} = k(d_i, d_j)$ for a PSD kernel k ($k(d_i, d_i) = 1$). The VS is maximal (n) when documents are orthogonal and minimal (1) when identical (Friedman and Dieng, 2023).

Vendi Retrieval Score. To balance diversity with relevance, we define the Vendi Retrieval Score (VRS) as a convex combination of the two, each min–max normalized to $[0, 1]$ (full normalization in Appendix E):

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

where $s \in [0, 1]$ controls the trade-off and $\mathrm{SS}(q, \mathcal{D})$ is the average query–document cosine similarity from dense embeddings. Higher s favors diversity; lower s favors relevance.

Greedy subset selection. Optimizing the VRS over all k-subsets is intractable, so documents are selected greedily, at each step adding $d^* = \arg\max_{d \in \mathcal{C} \setminus \mathcal{D}_{i-1}} \mathrm{VRS}(\mathcal{D}_{i-1} \cup \{d\})$ from the candidate set $\mathcal{C}$ returned by similarity search. Per query, this costs $O(k \cdot |\mathcal{C}| \cdot k^3)$, where the dominant k^3 term is the single VS eigendecomposition on the current candidate-augmented subset of size $\leq k$, repeated for each of the k slots over the $|\mathcal{C}|$ candidates. For typical sizes ($k \leq 10$, $|\mathcal{C}| \leq 50$) this is well under 50ms per query in our implementation and is negligible compared with LLM-side latency. The VRS is non-submodular, so greedy selection carries no formal guarantee, but the empirical gap between greedy and exhaustive search is negligible at small k (Appendix E.4); greedy optimization is similarly used on the non-submodular MMR objective. Relevance enters twice—$\mathcal{C}$ is first selected by similarity search, and the SS term then refines the ranking—so the VRS never optimizes diversity over query-irrelevant documents.

3.2 Vendi-RAG

Vendi-RAG embeds Vendi retrieval in an iterative loop. Each iteration (i) retrieves documents, (ii) generates chain-of-thought reasoning (Wei et al., 2022) and a candidate answer grounded in that evidence, (iii) scores the answer with an LLM judge, and (iv) updates the diversity weight s. The loop stops when the quality score exceeds τ or after N iterations, returning the highest-scoring answer (Algorithm 1). We detail the two non-standard components below; prompts and pipeline specifics are in Appendix E.

Quality evaluation. An LLM judge scores each candidate on three structural criteria—*coherence* (logical consistency of the reasoning), *relevance* (grounding in the retrieved evidence), and *query alignment* (how completely the answer addresses the question)—each on a 1–10 scale. The quality score is their mean divided by 10, giving $Q_t = \frac{1}{10}\,\mathrm{mean}(C, R, Q) \in [0, 1]$, with stopping threshold $\tau = 0.85$ (equivalently $8.5/10$ on the raw scale). The judge influences only the diversity weight s and never edits the answer text; factual grounding is checked separately by an external

Algorithm 1 Vendi-RAG Inference Pipeline

Require: Query q, Knowledge base $\mathcal{K}$, Max iterations N, Quality threshold τ, Momentum β
Ensure: Final answer $\hat{a}^*$

```
Initialize: q_1 ← q, s_1 ← 0.8, best_quality ← 0
for i = 1 to N do
    C ← TopK-Similar(q_i, K, k=50)
    D_i ← GreedySelect(C, s_i, k=5)                ▷ Eq. (2)
    r_i ← GenerateCoT(q_i, D_i)
    â_i ← GenerateAnswer(q, D_i, r_{1:i})
    Q_i ← LLM-Judge(â_i, q, D_i)
    if Q_i > best_quality then
        â* ← â_i, best_quality ← Q_i
    end if
    if Q_i ≥ τ then
        return â*
    end if
    s^target ← clip(1 − Q_i / max(best_quality, ε), 0, 1)
    s_{i+1} ← β · s_i + (1 − β) · s^target
    q_{i+1} ← RefineQuery(q_i, â_i, r_i)
end for
return â*
```

NLI faithfulness metric (Appendix D.6). The exact judge prompt is reproduced in Appendix E.5.

Dynamic adjustment of s. Let s_t be the diversity weight at iteration t. It is updated via an exponential moving average (EMA) with momentum β:

$$s_{t+1}^{\text{target}} = \text{clip}\Big(1 - \frac{Q_t}{\max(\max_{\tau \le t} Q_\tau, \epsilon)},\ 0,\ 1\Big)\,,$$

$$s_{t+1} = \beta\, s_t + (1-\beta)\, s_{t+1}^{\text{target}}, \qquad (3)$$

with $\epsilon = 0.01$ and $\beta = 0.3$. The rule is explore–exploit: low Q_t raises the target toward diversity (broader exploration for missing evidence), high Q_t lowers it toward relevance (consolidation along the current reasoning path), analogous to decaying exploration in reinforcement learning. When Q_t is the running best, $s^{\text{target}} = 0$ and the EMA makes the realized shift gradual ($s_{t+1} = \beta s_t$); any subsequent quality drop pushes the target back up, re-introducing diversity pressure. Combined with bounded iterations and best-answer tracking (the highest-quality answer across iterations is returned), trajectories converge in 3–4 iterations on average and a poor intermediate score cannot corrupt the final output. The pipeline also scales effort with difficulty: easy queries converge in 1–2 iterations while harder ones use 4–5. The dominant per-step cost is the $O(k^3)$ VS eigendecomposition on the retrieved subset (with $k \le 10$, not corpus size); a full latency breakdown is in Table 6.

4 Experiments

In this section, we present a comprehensive evaluation of Vendi-RAG on multi-hop QA tasks. We first analyze the effectiveness of the Vendi retrieval strategy in enhancing retrieval diversity and evidence coverage, and then evaluate the full Vendi-RAG pipeline, comparing it against strong baselines. All experiments use three challenging multi-hop QA benchmarks: MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and 2WikiMultiHopQA (Ho et al., 2020) (see Appendix A for details). Additional analyses, including the s sensitivity study, faithfulness evaluation, computational cost, cross-model judge robustness, and per-document-size trends, are deferred to Appendix D.

Experimental setup. We use the standard train/dev/test splits for each dataset, with evaluation on the test set (500 samples each for MuSiQue and 2WikiMultiHopQA, 7,405 for HotpotQA). All experiments are repeated across 3 random seeds (42, 123, 456), and we report mean ± standard deviation. For document retrieval, we use Sentence-BERT (all-mpnet-base-v2) as the dense encoder for all methods to ensure a fair comparison. Statistical significance is assessed using paired bootstrap tests with 10,000 resamples; markers (*) indicate $p < 0.05$ against the *per-cell* strongest non-Vendi comparator in Table 4 (the per-cell comparator is listed in Appendix D.2).

Baseline configurations. To ensure a fair comparison, we provide detailed configurations for all baselines:

- **Adaptive Retrieval**: Single retrieval pass using cosine similarity, $k = 5$ documents.
- **Adaptive-RAG**: Iterative retrieval with $N = 5$ maximum iterations, quality threshold $\tau = 0.8$, same retriever and $k = 5$.
- **Adaptive Retrieval with MMR**: MMR with $\lambda = 0.5$ balancing relevance and diversity, $k = 5$ documents. A full λ sweep is reported in Appendix D.
- **IRCoT** (Trivedi et al., 2023): Interleaved retrieval with chain-of-thought reasoning, retrieving per reasoning step using the same dense encoder and $k = 5$. IRCoT varies the retrieval *loop* rather than the scoring function and serves as a loop-level iterative baseline.
- **Self-RAG** (Asai et al., 2023): Official implementation with default hyperparameters. Self-RAG is designed for single-hop retrieval with on-demand self-reflection rather than for explicit multi-hop reasoning; under the multi-hop setup here its self-reflection tokens frequently suppress further retrieval and the resulting numbers (e.g.,

Dataset	Method	VS	MPD
MuSiQue	Adapt. Retrieval	6.13$\pm$0.08	1.25$\pm$0.03
	Adaptive-RAG	6.55$\pm$0.11	1.42$\pm$0.05
	Vendi-RAG	**7.12**$\pm$0.09*	**1.95**$\pm$0.04*
Hotpot	Adapt. Retrieval	4.95$\pm$0.12	1.10$\pm$0.04
	Adaptive-RAG	5.21$\pm$0.15	1.31$\pm$0.06
	Vendi-RAG	**6.82**$\pm$0.10*	**1.89**$\pm$0.05*
Wiki	Adapt. Retrieval	5.34$\pm$0.09	1.32$\pm$0.04
	Adaptive-RAG	5.81$\pm$0.13	1.45$\pm$0.05
	Vendi-RAG	**6.68**$\pm$0.11*	**1.78**$\pm$0.04*

Table 2: Retrieval diversity (Vendi Score (VS) and Max Pairwise Distance (MPD)) across datasets and methods. Vendi-RAG achieves higher diversity than baselines. Results are mean $\pm$ std over 3 seeds.

1.2 EM on MuSiQue) reflect this design mismatch rather than a tuning failure, consistent with the multi-hop drop reported by Jeong et al. (2024). We include Self-RAG for completeness but the more informative comparators are the iterative diversity-aware baselines (MMR, k-DPP, IRCoT).

- **LightRAG**: Graph-based retrieval with default entity extraction settings, maximum 5 retrieved subgraphs.

The baselines span both retrieval scoring functions (Adaptive Retrieval, Adaptive Retrieval + MMR, Adaptive Retrieval + k-DPP, Adaptive-RAG, LightRAG) and retrieval loop structures (IRCoT, Self-RAG), so that the contribution of the VRS scoring function can be isolated from the contribution of the loop design.

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, using the Vendi Score (VS) and the max pairwise distance (MPD). Table 2 summarizes the results. Vendi-RAG achieves higher diversity than both baselines on all datasets, demonstrating its ability to retrieve documents capturing multiple perspectives relevant to the query. Adaptive-RAG, which incorporates iterative refinement but lacks explicit diversity control, shows only moderate improvement over Adaptive Retrieval.

Diversity translates into better evidence coverage. Beyond answer quality, we evaluate retrieval effectiveness using supporting facts recall (SF Recall) and document recall (Doc Recall@k), which measure whether the retrieved documents contain

Method	SF Recall	Doc Recall@5
Adaptive Retrieval	58.3$\pm$1.2	62.1$\pm$0.9
Adaptive-RAG	63.7$\pm$1.5	67.4$\pm$1.1
MMR ($\lambda = 0.5$)	61.2$\pm$1.3	65.8$\pm$1.0
Vendi-RAG	**71.4**$\pm$1.1*	**74.2**$\pm$0.8*

Table 3: Retrieval quality metrics on HotpotQA. SF Recall measures supporting facts coverage; Doc Recall@5 measures gold document retrieval.

the gold evidence needed for multi-hop reasoning. Table 3 shows that Vendi-RAG's diversity-aware retrieval improves evidence coverage over relevance-only and MMR baselines, directly linking the diversity gains in Table 2 to a downstream-relevant quantity. The VRS combines diversity with query relevance over a candidate pool already filtered by similarity search (top-50 by cosine similarity), so the additional diversity is constrained to query-relevant documents and manifests as higher coverage of the gold supporting facts rather than topically unrelated content. To separate the diversity effect from the effect of running additional retrieval rounds, the SF Recall and Doc Recall@5 in Table 3 are evaluated on the *first* retrieval round only, with no query rewriting and no iterative refinement, for every method. Under this matched-round setup Vendi-RAG already retrieves more gold supporting facts than Adaptive-RAG and MMR (+7.7 and +10.2 SF Recall, respectively), so the coverage gain is attributable to the VRS scoring function on the candidate pool and not to extra iterations. A complementary matched-iteration view at the answer level is provided by the iterative diversity-aware baselines in Table 4: all of them use the same $N = 5$ iteration budget and the same retriever, so the $+2.4$ accuracy gap on HotpotQA between Vendi-RAG (58.4) and MMR (56.0) cannot be explained by "more iterations."

Accuracy on multi-hop QA tasks. We next evaluate the full Vendi-RAG pipeline on multi-document reasoning. Table 4 reports results using GPT-3.5 Turbo. Vendi-RAG with dynamically adjusted s_t (initialized at $s_1 = 0.8$) achieves the best Exact Match (EM) and Accuracy (Acc) on all three datasets. F1 gains are smaller and not uniformly the largest in the table; F1 rewards partial token overlap, on which relevance-only retrieval already performs adequately, while EM and Acc more directly reflect full multi-hop correctness, where diversity-aware retrieval contributes the most. The starred entries in Table 4 mark cells where the paired-

bootstrap test (10,000 resamples, $p < 0.05$) is computed against the strongest non-Vendi baseline on that specific cell—MMR for HotpotQA Acc and 2WikiMultiHopQA Acc, IRCoT for HotpotQA EM/F1 and MuSiQue EM, and Adaptive-RAG for MuSiQue Acc and 2WikiMultiHopQA EM. The two F1 cells where Adaptive-RAG is highest (MuSiQue and 2WikiMultiHopQA) carry no significance marker. Appendix D.2 lists the per-cell comparator and the test details.

Loop structure is controlled. All iterative methods in Table 4 (Adaptive-RAG, Adaptive Retrieval + MMR, Adaptive Retrieval + k-DPP, IRCoT, and Vendi-RAG) use the same maximum iteration budget ($N = 5$), the same retriever, and the same $k = 5$, so loop structure is held fixed across the diversity-aware baselines. The gap from Adaptive-RAG (47.0 HotpotQA Acc) up to MMR/k-DPP/IRCoT (55.8–56.0) and then to Vendi-RAG (58.4) reflects *scoring-function* differences, not loop differences: Adaptive-RAG's underperformance on HotpotQA is driven by its complexity-classifier, which routes many HotpotQA queries to no-retrieval or single-step paths (consistent with Jeong et al. 2024); once a method always iterates with diversity-aware retrieval, accuracy concentrates in the 55–59 band, and Vendi-RAG sits at the top of that band. The headline gap of $+2.4$ accuracy points over MMR is therefore the most informative comparison for the contribution of the VRS.

Ablations: VS, fixed-s, and the LLM-judge loop. The Vendi-RAG rows in Table 4 isolate three design choices. (i) Replacing similarity ranking with the VRS at fixed $s = 0.3$ already lifts HotpotQA accuracy to 55.7, matching MMR/k-DPP/IRCoT. (ii) Increasing the fixed diversity weight to $s = 0.8$ yields 58.0 HotpotQA accuracy, which is the bulk of the total gain over MMR ($+2.0$ of $+2.4$ points; analogous patterns hold for MuSiQue and 2WikiMultiHopQA). (iii) The LLM-judge dynamic loop adds a further $+0.4$ accuracy on HotpotQA, $+0.2$ on MuSiQue, and $+0.4$ on 2WikiMultiHopQA on top of fixed $s = 0.8$. A paired-bootstrap test at the per-query level (Appendix D.2) finds the dynamic-vs-fixed-s=0.8 gap significant at $p < 0.05$ on HotpotQA, where most queries trigger ≥ 3 iterations and the loop is active, and borderline ($p \approx 0.08$–0.12) on MuSiQue and 2WikiMultiHopQA, where many queries terminate after ≤ 2 iterations. The takeaway is that the bulk of the improvement is delivered by the VS-based scoring function with a high fixed diversity weight, and the dynamic loop adds a smaller but consistent gain that concentrates on the harder queries where the iteration budget is fully used; the fixed-s variant therefore gives the same retrieval-side gain at the same inference cost as MMR, and the dynamic loop is a quality-on-demand option when extra latency is acceptable.

Comparison with diversity-aware retrievers (MMR, k-DPP). Two natural set-level diversification baselines are MMR (Carbonell and Goldstein, 1998), which uses greedy pairwise dissimilarity, and k-Determinantal Point Processes (Kulesza and Taskar, 2012), which sample subsets in proportion to the determinant of a kernel submatrix. Both use the same dense encoder as Vendi-RAG to keep the comparison apples-to-apples. Table 4 shows that MMR and k-DPP land in the same range and are themselves competitive iterative-loop alternatives to Adaptive-RAG, yet Vendi-RAG (dynamic) still improves over the better of the two by $+1.2$ EM and $+2.4$ Acc on HotpotQA, $+1.4$ EM and $+2.4$ Acc on 2WikiMultiHopQA, and $+1.8$ EM and $+1.6$ Acc on MuSiQue. The Vendi Score's set-level eigenvalue formulation differs from the pairwise-distance objective of MMR and from the log-determinant objective of DPPs: it directly measures the effective number of distinct documents, and the diversity-relevance trade-off is controlled through a single scalar s. A full MMR λ sweep (Appendix D.9) shows this gap is not an artifact of suboptimal λ.

Robustness across LLM backbones. To check that the gains are not specific to a single model family, we evaluate Vendi-RAG (with $s_1 = 0.8$) against Adaptive-RAG across five backbones spanning three model families: GPT-3.5, GPT-4o, GPT-4o-mini (Figure 2), the open-weight Llama 3-70B-Instruct, and Gemini Flash 2.0. Table 5 summarizes HotpotQA accuracy for the two new architectures; full per-dataset numbers are in Appendix D.4. Vendi-RAG yields a $+4.4$ accuracy gain on both Llama 3-70B and Gemini Flash 2.0, comparable to the gains on the GPT family. Larger backbones (GPT-4o) achieve the strongest absolute performance, and the relative advantage of Vendi-RAG holds with smaller models like GPT-4o-mini and with open-weight architectures.

Retriever robustness. Downstream performance can depend strongly on the first-stage retriever, so we repeat the HotpotQA experiment with a differ-

Methods	MuSiQue			HotpotQA			2WikiMultiHopQA		
	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
No Retrieval	20.40±0.5	31.30±0.6	24.40±0.4	37.40±0.7	51.04±0.8	43.20±0.6	37.00±0.5	48.50±0.7	43.40±0.5
Single-step Approach	16.40±0.6	26.70±0.7	23.60±0.5	39.60±0.8	50.44±0.9	45.60±0.7	46.80±0.6	57.69±0.8	52.60±0.6
Adaptive Retrieval	18.80±0.5	30.30±0.6	24.80±0.4	38.60±0.7	50.70±0.8	43.20±0.6	44.20±0.5	55.11±0.7	50.60±0.5
IRCoT	22.60±0.6	31.50±0.7	28.40±0.5	41.20±0.8	55.30±0.9	55.80±0.7	46.00±0.6	57.40±0.8	58.20±0.6
LightRAG	17.40±0.6	27.51±0.7	23.80±0.5	37.40±0.8	47.30±0.9	42.30±0.7	42.21±0.6	52.24±0.8	47.50±0.6
Adaptive-RAG	21.80±0.6	32.60±0.7	29.60±0.5	40.40±0.8	52.56±0.9	47.00±0.7	46.60±0.6	**60.09**±0.8	56.80±0.6
Self-RAG‡	1.20±0.3	8.20±0.5	11.80±0.4	5.60±0.4	17.86±0.6	30.60±0.5	3.00±0.3	19.14±0.5	39.00±0.5
Adaptive Retrieval + MMR	22.6±0.5	31.0±0.6	28.8±0.4	41.0±0.7	55.0±0.8	56.0±0.6	45.8±0.5	57.0±0.7	59.0±0.5
Adaptive Retrieval + k-DPP	22.4±0.5	31.2±0.6	28.6±0.4	40.8±0.7	54.8±0.8	55.6±0.6	45.6±0.5	56.8±0.7	58.6±0.5
Vendi-RAG(fixed $s = 0.0$)	22.4±0.5	30.4±0.6	28.2±0.4	40.6±0.7	54.1±0.8	55.4±0.6	45.4±0.5	56.8±0.7	58.8±0.5
Vendi-RAG* (fixed $s = 0.3$)	22.4±0.5	31.1±0.6	28.6±0.4	40.7±0.7	55.0±0.8	55.7±0.6	45.3±0.5	56.8±0.7	59.1±0.5
Vendi-RAG(fixed $s = 0.8$)	23.0±0.5	31.2±0.6	30.2±0.4	42.0±0.7	56.9±0.8	58.0±0.6	47.0±0.5	58.7±0.7	61.0±0.5
Vendi-RAG(fixed $s = 1.0$)	22.9±0.5	31.0±0.6	29.4±0.4	41.6±0.7	55.8±0.8	57.0±0.6	46.4±0.5	57.8±0.7	60.2±0.5
Vendi-RAG(dynamic, $s_1 = 0.8$)	**24.4**±0.5*	32.52±0.6	**30.4**±0.4*	**42.2**±0.7*	**57.04**±0.8*	**58.4**±0.6*	**47.2**±0.5*	58.9±0.7	**61.4**±0.5*

Table 4: Performance on multi-hop QA datasets using GPT-3.5 Turbo. Results are mean ± std over 3 seeds. * indicates paired-bootstrap significance (10,000 resamples, $p < 0.05$) against the strongest non-Vendi comparator on that specific cell (per-cell comparator listed in Appendix D.2); the strongest comparator differs by metric (e.g., MMR for HotpotQA Acc, IRCoT for HotpotQA EM/F1, Adaptive-RAG for MuSiQue Acc). The two F1 cells where Adaptive-RAG is highest (MuSiQue, 2WikiMultiHopQA) carry no *. ‡ Self-RAG is designed for single-hop retrieval with self-reflection; its multi-hop numbers reflect this design mismatch (see baseline description) and are included for completeness only. Vendi-RAG ($s_1 = 0.8$, dynamic) achieves the best EM and Acc on all three datasets; F1 is competitive, with Vendi-RAG highest on HotpotQA, essentially tied with Adaptive-RAG on MuSiQue, and Adaptive-RAG highest on 2WikiMultiHopQA. IRCoT, MMR, and k-DPP are obtained under the same retriever, k, and test splits as Vendi-RAG and occupy the same performance band as the other diversity-aware baselines.

Backbone	Method	HotpotQA Acc
Llama 3-70B-Instruct	Adaptive-RAG	46.4±0.6
	Vendi-RAG	**50.8**±0.5
Gemini Flash 2.0	Adaptive-RAG	48.1±0.6
	Vendi-RAG	**52.5**±0.5

Table 5: Cross-architecture accuracy on HotpotQA. Vendi-RAG (dynamic, $s_1 = 0.8$) outperforms Adaptive-RAG by $+4.4$ accuracy points on both open-weight (Llama 3-70B) and proprietary non-GPT (Gemini Flash 2.0) backbones. Per-dataset EM and F1 are in Appendix D.4.

ent dense encoder (BGE-large, Xiao et al., 2024) in place of all-mpnet-base-v2. With BGE-large, Adaptive-RAG reaches 48.7 accuracy and Vendi-RAG (dynamic, $s_1 = 0.8$) reaches 59.2; the $+10.5$ gain is in the same range as the $+11.4$ obtained with all-mpnet-base-v2 (Table 4), and the relative ordering of all baselines is preserved (full numbers in Appendix D.8). For the MMR baseline, sweeping $\lambda \in \{0.3, 0.5, 0.7\}$ on HotpotQA yields 54.8, 56.0, and 55.2 accuracy respectively (Appendix D.9); the best MMR setting is still 2.4 accuracy points below Vendi-RAG, so the gap is not an artifact of MMR hyperparameter tuning.

5 Conclusion

Traditional RAG pipelines optimize for relevance alone and tend to retrieve semantically redundant documents, which is a poor match for multi-hop reasoning that depends on combining complementary evidence. We introduced Vendi-RAG, a framework that elevates retrieval diversity to a first-class objective via the Vendi Retrieval Score and adapts the diversity-relevance balance with an explore–exploit feedback loop driven by an LLM-based answer-quality signal. Across HotpotQA, MuSiQue, and 2WikiMultiHopQA with GPT-3.5, Vendi-RAG improves accuracy by $+2.4$, $+1.6$, and $+2.4$ points over the strongest diversity-aware iterative baseline (Adaptive Retrieval + MMR); the corresponding gaps against routing-based Adaptive-RAG are $+11.4$, $+0.8$, and $+4.6$, with the largest gap explained by the routing behavior described in Section 4. A fixed-s variant of Vendi-RAG captures most of the improvement at no extra inference cost over MMR; the LLM-judge loop adds a smaller, statistically significant gain on hard queries at higher latency, giving practitioners a tunable accuracy–cost trade-off. The pattern is stable across five instruction-tuned LLM backbones (GPT-3.5, GPT-4o, GPT-4o-mini, Llama 3-70B-Instruct, Gemini Flash 2.0) and two dense encoders on multi-hop English QA, indicating that the gain is driven by the retrieval objective rather than by any individual model in our evaluation suite. The framework is composable: Vendi retrieval is a scoring function that can be combined with retrieval-loop methods (IRCoT, Adaptive-RAG) and with post-retrieval utility selectors.

Limitations

Our study is restricted to English multi-hop QA over three benchmarks (HotpotQA, MuSiQue, 2WikiMultiHopQA), five instruction-tuned backbones (GPT-3.5, GPT-4o, GPT-4o-mini, Llama 3-70B-Instruct, Gemini Flash 2.0), and two dense retrievers (all-mpnet-base-v2, BGE-large). Transfer to non-English benchmarks, tasks where complementary evidence matters less (single-hop QA, factoid lookup), and base (non-instruction-tuned) models is untested, so our consistency across this suite is not a guarantee of universal model-agnosticism. Vendi-RAG also inherits the standard RAG dependence on knowledge-base quality and coverage. The dynamic LLM-judge loop adds compute on hard queries: latency rises from 5.8s to 8.4s and tokens from 3.4K to 5.1K per query (Table 6), mostly from judge calls rather than the eigendecomposition, for a dollar overhead of \$2–\$3 per $1{,}000$ queries at current GPT-3.5 pricing (Appendix C.5). The fixed-s variant matches MMR's iteration count and token budget, drops judge calls, and retains most of the gain ($+2.0$ of $+2.4$ HotpotQA points over MMR), making it preferable for latency- or cost-critical settings. The dynamic loop's additional gain over fixed-s=0.8 is significant on HotpotQA ($p < 0.05$) but borderline on MuSiQue and 2WikiMultiHopQA (Appendix D.2); it is most valuable when the iteration budget is fully used. Finally, the loop relies on the LLM judge providing a useful *relative* quality signal; although the cross-judge control on HotpotQA shifts accuracy by less than 0.5 points, judges may behave differently on domains far from their pre-training distribution. Future work includes replacing the judge with a smaller learned reward model and jointly optimizing s with per-hop retrieval planning.

Ethical Considerations

Vendi-RAG uses LLMs both for answer generation and as a judge for answer quality. The same ethical concerns that apply to LLM-based QA systems apply here: outputs can reflect biases present in the underlying model and in the retrieved knowledge base, and grounded retrieval cannot rule out subtly biased or factually outdated content. The diversity-aware retrieval mechanism itself does not introduce new categories of risk: it operates over a candidate pool that is first filtered by relevance, so it can only redistribute attention across documents that have already passed the relevance filter; it does not synthesize new content. Faithfulness is monitored externally via an NLI-based metric (Appendix D.6). We recommend that downstream deployments verify the knowledge base for coverage and bias and pair Vendi-RAG with the safety filters and provenance tracking that are standard for production RAG systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 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.

Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems 31*, pages 5622–5633.

Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74, Portland, Oregon, USA. ACM.

Nicola de Cao, Wilker Aziz, and Ivan Titov. 2021. Question answering survey: Directions, challenges, datasets, evaluation matrices. In *arXiv preprint arXiv:2112.03572*.

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

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems 27*, pages 2069–2077.

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*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Aman Jain and Aparna Garimella. 2025. Modeling contextual passage utility for multihop question answering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 464–471.

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, and 1 others. 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*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

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*.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 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, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

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. 2023. 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, and 1 others. 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*.

Rodrygo L. T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010*, pages 87–99, Milton Keynes, UK. Springer.

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, and 1 others. 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. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10014–10037.

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, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2014. Submodular optimization and machine learning: Theoretical advances, connections, and applications (tutorial). In *Tutorial at the 31st International Conference on Machine Learning (ICML)*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2024. C-pack: Packed resources for general Chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

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

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2025. Distilling a small utility-based passage selector to enhance retrieval-augmented generation. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 22–30.

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 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; the full prompt is reproduced in Appendix E.5. The judge produces three integer scores on a 1–10 scale for coherence, relevance, and query alignment; we then divide their mean by 10 to obtain $Q_t \in [0, 1]$, and use threshold $\tau = 0.85$ (equivalently $8.5/10$ on the raw scale) for all experiments. The initial diversity weight is fixed to $s_1 = 0.8$; the ablation studies in Tables 4 and 7 vary this hyperparameter and compare fixed and dynamic schedules.

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 Computational cost analysis.

We provide a per-query cost breakdown on HotpotQA in Table 6. Three cost components are relevant. (i) The VS eigendecomposition is $O(k^3)$ per retrieval step *per query* with $k \leq 10$, which is <5ms on the retrieved subset and is dwarfed by LLM calls; the greedy-selection budget of $O(k \cdot |\mathcal{C}| \cdot k^3)$ is therefore <50ms per query for $|\mathcal{C}| = 50$. (ii) The iterative loop adds retrieve/reason calls beyond a single-step pass. (iii) The LLM judge calls add one extra forward pass per iteration. The dynamic-s variant of Vendi-RAG averages 3.2 iterations and 3.2 judge calls per query, against 2.8 iterations and no judge calls for Adaptive-RAG; this drives the latency from 5.8s to 8.4s (∼45% higher) and tokens from 3.4K to 5.1K. At GPT-3.5 API pricing as of writing (∼\$0.50 per 1M input tokens, ∼\$1.50 per 1M output tokens, with judge prompts at roughly 80% input / 20% output), the dollar overhead of the judge loop is on the order of \$2–\$3 per 1,000 queries on top of the Adaptive-RAG cost, which is small relative to per-query end-user value in QA settings but non-trivial at very high traffic. The fixed-s variant of Vendi-RAG matches Adaptive-RAG iteration count and removes judge calls entirely, so its cost profile is the same as the iterative diversity-aware baselines (MMR, k-DPP, IRCoT) while retaining the bulk of the accuracy gain (Section 4, ablation discussion).

C.6 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.

Method	Iterations	Judge Calls	Tokens	Latency (s)
Single-step	1.0	0	1.2K	2.1
Adaptive-RAG	2.8	0	3.4K	5.8
Adaptive Retrieval + MMR	3.0	0	3.6K	6.0
Vendi-RAG (fixed s=0.8)	3.0	0	3.6K	6.0
Vendi-RAG (dynamic, s_1=0.8)	3.2	3.2	5.1K	8.4

Table 6: Computational cost analysis on HotpotQA (per-query average, GPT-3.5 Turbo). The fixed-s Vendi-RAG variant has the same cost profile as MMR; the dynamic loop adds judge calls and increases latency by roughly 45% in exchange for the additional accuracy gain reported in Section 4.

D Additional Experiments and Analyses

D.1 Sensitivity Analysis of the VRS Process

To evaluate the robustness of the VRS 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 elucidates the trade-off between retrieval precision and diversity, which is crucial for enhancing multi-hop reasoning performance. The analysis was performed using 100 randomly sampled queries to cover a diverse range of query types and complexity levels. We evaluated the retrieval pipeline across multiple s values from 0.0 to 1.0, incremented in small steps to capture granular variations. Setting $s = 0.0$ serves as a baseline representing pure similarity search, where retrieval relies exclusively on cosine similarity between embeddings, without any emphasis on diversity. This provides a reference point for measuring the impact of increasing s.

To quantify deviations from the baseline, we employed two complementary ranking comparison metrics:

- **Kendall's τ:** Measures 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 at higher s values.

To connect rank changes to downstream performance, we additionally measure supporting facts recall at each s value in Table 7. As s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively, showing that higher s values promote retrieval diversity. SF Recall peaks at $s = 0.8$, indicating that the rank changes induced by the diversity mechanism lead to better coverage of multi-hop evidence.

s	τ	ρ	SF Recall
0.0	1.00	1.00	58.3
0.2	0.797	0.828	61.2
0.4	0.688	0.742	65.7
0.6	0.485	0.528	69.1
0.8	0.265	0.316	**71.4**
1.0	0.074	0.078	68.9

Table 7: Sensitivity Analysis of the VRS. Higher s values increase diversity (lower τ, ρ) while SF Recall peaks at $s = 0.8$.

D.2 Statistical Significance Tests

Significance markers (*) in Table 4 are obtained by a paired bootstrap with $10{,}000$ resamples at the per-query level (test splits: 500 queries for MuSiQue, 500 for 2WikiMultiHopQA, $7{,}405$ for HotpotQA). The strongest non-Vendi comparator differs per cell; Table 8 lists which baseline is used for each marker, identified as the row with the maximum value on that cell among the non-Vendi rows of Table 4. We use the strongest per-cell comparator rather than a single fixed baseline so that significance reflects improvement over the genuinely competitive method in each setting, not over an arbitrary one. Cells in which Vendi-RAG is not bold in Table 4 (MuSiQue F1 and 2WikiMultiHopQA F1, both topped by Adaptive-RAG) carry no * marker, since Vendi-RAG does not outperform the strongest comparator there.

Dataset	EM	F1	Acc
MuSiQue	IRCoT/MMR (22.6)	—	Adaptive-RAG (29.6)
HotpotQA	IRCoT (41.2)	IRCoT (55.3)	MMR (56.0)
2WikiMultiHopQA	Adaptive-RAG (46.6)	—	MMR (59.0)

Table 8: Per-cell strongest non-Vendi comparator used to compute the significance markers in Table 4. "—" indicates a cell where Vendi-RAG does not exceed the strongest comparator (Adaptive-RAG on both F1 columns) and therefore receives no *.

Dynamic-s vs. fixed-s=0.8. We also run the paired bootstrap for the dynamic-s row against the fixed-s=0.8 row (the strongest fixed variant), per query, on the accuracy metric. We obtain $p < 0.05$ on HotpotQA (where the iteration budget is used most heavily and the dynamic schedule visibly adjusts s_t), and $p \approx 0.08$ on MuSiQue and $p \approx 0.10$ on 2WikiMultiHopQA, where many queries terminate after ≤ 2 iterations and the dynamic schedule has fewer opportunities to act. The dynamic loop therefore delivers a small, consistent, and (on HotpotQA) statistically significant improvement on top of the gain from a strong fixed diversity weight; the fixed-s variant remains a competitive lower-cost option.

D.3 IRCoT under a Matched Setup

IRCoT (Trivedi et al., 2023) varies the retrieval loop (per-step retrieval triggers, no quality-based stopping) rather than the retrieval scoring function. To make the comparison apples-to-apples, we report it under a setup matched to ours: the same dense retriever (all-mpnet-base-v2), the same $k = 5$, and the same test splits (500 samples for MuSiQue and 2WikiMultiHopQA, 7,405 for HotpotQA), with GPT-3.5 Turbo as the backbone. Numbers under this setup appear in the main Table 4. Across all three datasets, IRCoT lands close to Adaptive Retrieval + MMR and Adaptive-RAG on EM and F1, and below the fixed-$s = 0.8$ Vendi-RAG variant on Acc. IRCoT's per-step scheduling and the VRS scoring function operate on different axes and are composable: a per-step IRCoT trigger can call a VRS-based retriever instead of a similarity-only one, which is a natural direction for future work.

D.4 Robustness Across LLM Backbones

We evaluate Vendi-RAG (dynamic, $s_1 = 0.8$) against Adaptive-RAG using two architecturally distinct backbones beyond the GPT family: Llama 3-70B-Instruct (open-weight) and Gemini Flash 2.0 (proprietary non-GPT). Table 9 reports HotpotQA results.

The accuracy gains on both Llama 3-70B and Gemini Flash 2.0 are comparable to those on the GPT family, indicating that the diversity-aware retrieval mechanism transfers across model architectures.

D.5 Judge Robustness and Self-Reinforcement

Same-LLM generation and judging could in principle introduce self-reinforcement bias. Three properties of the pipeline limit this risk: (i) the judge evaluates only structural properties (coherence, relevance, query alignment) and influences only the diversity parameter s_t, never the answer text; (ii) factual grounding is assessed by an *external* NLI-based faithfulness metric (Appendix D.6), on which Vendi-RAG attains the highest score; (iii) bounded iterations and EMA smoothing prevent any single score from dominating the trajectory. A control experiment on HotpotQA with a different-backbone judge changes final accuracy by less than 0.5 points, indicating that same-model judging does not meaningfully bias the results.

D.6 Faithfulness Evaluation

To assess whether generated answers are grounded in retrieved evidence, we compute faithfulness scores using an NLI-based metric following (Honovich et al., 2022). Table 10 shows that Vendi-RAG maintains high faithfulness while improving answer quality, indicating that the diversity-driven retrieval does not introduce unsupported claims.

D.7 Impact of the Number of Retrieved Documents

To examine the impact of document set size on retrieval effectiveness, we analyze Vendi-RAG and Adaptive-RAG across varying document sizes and initial settings $s_1 = \{0.3, 0.8, 1.0\}$. Figure 3 illustrates the relationship between document size and performance on HotpotQA. Vendi-RAG consistently outperforms Adaptive-RAG in accuracy for document sizes greater than two with any s_1 setting.

Method	Backbone	EM	F1	Acc
Adaptive-RAG	Llama 3-70B	39.8±0.7	51.9±0.8	46.4±0.6
Vendi-RAG	Llama 3-70B	**43.1**±0.6	**56.2**±0.7	**50.8**±0.5
Adaptive-RAG	Gemini Flash 2.0	41.2±0.7	53.4±0.8	48.1±0.6
Vendi-RAG	Gemini Flash 2.0	**44.6**±0.6	**57.8**±0.7	**52.5**±0.5

Table 9: Cross-architecture results on HotpotQA. Vendi-RAG improves accuracy by $+4.4$ points over Adaptive-RAG for both Llama 3-70B-Instruct and Gemini Flash 2.0.

Method	Faithfulness
Adaptive Retrieval	0.82±0.02
Adaptive-RAG	0.84±0.02
Vendi-RAG	**0.86**±0.01

Table 10: Faithfulness scores (NLI-based) on HotpotQA.

As document size increases, accuracy improves for both methods, but the gain is notably higher for Vendi-RAG. EM and F1 scores exhibit a similar increasing trend, with Vendi-RAG showing a more pronounced improvement. The VS also increases with document size: since the VS is invariant under duplication (Friedman and Dieng, 2023), an increasing VS indicates less redundancy among the retrieved documents, evidence that Vendi-RAG alleviates the redundancy that often plagues RAG pipelines. Performance plateaus around $k = 8$–10; beyond that, context-window constraints require document truncation, which introduces confounds. We therefore extend the analysis to $k = 15$ in this appendix figure but caution that the largest-k points are affected by truncation.

D.8 First-Stage Retriever Robustness

To check that the gains are not tied to a particular dense encoder, we replace `all-mpnet-base-v2` with BGE-large (Xiao et al., 2024) and rerun HotpotQA with GPT-3.5 Turbo, keeping all other hyperparameters fixed. Table 11 reports the results. The relative ordering of methods is preserved and Vendi-RAG retains a $+10.5$ accuracy gap over Adaptive-RAG with BGE-large, comparable to the $+11.4$ gap with `all-mpnet-base-v2`. The absolute numbers shift upward by roughly one to two accuracy points for every method, reflecting BGE-large's stronger candidate pool, but the diversity-aware ordering is unchanged.

D.9 MMR λ Sweep

The MMR baseline in the main paper is fixed at $\lambda = 0.5$. To confirm that the gap between Vendi-RAG and MMR is not an artifact of suboptimal λ, we sweep $\lambda \in \{0.3, 0.5, 0.7\}$ on HotpotQA (GPT-3.5 Turbo). Table 12 reports the results.

The best MMR configuration ($\lambda = 0.5$) reaches 56.0% accuracy, still well below Vendi-RAG's 58.4%. The gap persists across all tested λ values, confirming that the advantage stems from the VRS formulation rather than from poorly tuned MMR.

E Method Details

E.1 Normalization of the VRS Components

The Vendi Retrieval Score combines a diversity term and a relevance term, each normalized to $[0, 1]$ so that neither dominates due to differing dynamic ranges:

$$\widetilde{\mathrm{VS}}_k(\mathcal{D}) = \frac{\mathrm{VS}_k(\mathcal{D}) - 1}{n - 1}, \tag{8}$$

$$\widetilde{\mathrm{SS}}(q, \mathcal{D}) = \frac{\mathrm{SS}(q, \mathcal{D}) - \mathrm{SS}_{\min}}{\mathrm{SS}_{\max} - \mathrm{SS}_{\min}}. \tag{9}$$

The VS normalization maps the score from its theoretical range $[1, n]$ to $[0, 1]$. The SS normalization uses running estimates of the minimum and maximum similarity scores computed over the candidate pool $\mathcal{C}$. The raw similarity $\mathrm{SS}(q, \mathcal{D})$ is the average cosine similarity between the query q and the documents in $\mathcal{D}$, using dense vector representations from transformer-based encoders (e.g., BERT, Sentence-BERT).

E.2 Detailed Pipeline Stages

The Vendi-RAG pipeline expands the four-step loop summarized in Section 3 into the following stages.

1. Initial retrieval. The process begins by retrieving an initial document set using Vendi retrieval with a high diversity weight ($s_1 = 0.8$ in all

Method	all-mpnet-base-v2	BGE-large
Adaptive Retrieval	43.2	44.8
Adaptive-RAG	47.0	48.7
Adaptive Retrieval + MMR	56.0	57.5
Vendi-RAG (dynamic)	**58.4**	**59.2**

Table 11: HotpotQA accuracy under two first-stage dense retrievers (GPT-3.5 Turbo). The Vendi-RAG advantage holds across encoders.

Method	Acc
MMR ($\lambda = 0.3$)	54.8
MMR ($\lambda = 0.5$)	56.0
MMR ($\lambda = 0.7$)	55.2
Vendi-RAG (dynamic, s_1=0.8)	**58.4**

Table 12: MMR λ sweep on HotpotQA (GPT-3.5 Turbo). The best MMR setting ($\lambda = 0.5$, 56.0%) still falls well below Vendi-RAG (58.4%); the gap is consistent across λ values and is not an artifact of suboptimal tuning.

experiments). Prioritizing broad semantic coverage ensures the retrieved documents capture multiple perspectives and avoids semantically redundant documents. This initial diversity is particularly important for multi-hop QA, where synthesizing information from varied sources is essential for constructing complete reasoning chains.

2. Reasoning generation. Based on the retrieved documents, the system generates chain-of-thought (CoT) reasoning steps following established prompting techniques (Wei et al., 2022). These intermediate steps (i) identify relevant facts from each document, (ii) establish logical connections between disparate pieces of information, and (iii) build a coherent inferential pathway toward the final answer, yielding transparent and interpretable reasoning for complex multi-hop queries.

3. Candidate answer generation. Using both the reasoning steps and the retrieved documents, the LLM produces candidate answers. This dual-input design grounds answers in the retrieved evidence while following the logical structure established by the reasoning steps. Multiple candidates may be generated to explore different interpretations of the available information.

4. Quality evaluation. An LLM judge scores each candidate (see Section 3 and the prompt below), producing $Q_t \in [0, 1]$.

5. Dynamic adjustment of the VRS with smoothing. The diversity parameter is updated via the EMA rule in Eq. (3). The target follows an inverse relationship with quality: low Q_t raises the target toward diversity (broader exploration to recover missing information), while high Q_t lowers it toward relevance (consolidation along the current reasoning path). The EMA with $\beta = 0.3$ dampens rapid changes, preventing oscillation between extreme diversity and relevance settings.

6. Iterative refinement. Retrieval and reasoning repeat with the adjusted s at each stage, and the query may be reformulated to better target information gaps. The loop continues until the quality score exceeds the threshold τ or the maximum iteration count N is reached.

7. Final answer selection. The answer with the highest quality score across all iterations is returned, together with its retrieved documents and reasoning chain for interpretability. Because the best answer over all iterations is selected, a poor intermediate score cannot corrupt the final output.

Why adjusting s matters. The dynamic adjustment of s balances exploration and exploitation. High diversity (s near 1) explores different facets of complex queries, crucial in multi-hop QA where synthesizing multiple perspectives is necessary; however, excessive diversity can introduce irrelevant content, while excessive relevance (s near 0) risks redundancy and limits comprehensive reasoning. When answer quality is satisfactory but incomplete, reducing s retrieves more specific documents to fill gaps; when quality is poor, increasing s promotes exploration of alternative document sets that may provide missing context. Unlike fixed-policy RAG systems, this adaptation lets Vendi-RAG respond to each query's difficulty, improving both breadth and precision.

E.3 Convergence and Stability of the s_t Update

The EMA update is a contraction toward the target with rate β, so the influence of any single quality score decays exponentially as β^m over subsequent

iterations m. Combined with the clipping operation, this guarantees $s_t \in [0, 1]$ at every step and prevents divergence or runaway oscillation. Empirically, s_t trajectories converge smoothly within 3–4 iterations across all three datasets and all evaluated backbones. The judge needs only to provide a *relative* quality signal (whether the current answer is better or worse than previous iterations) rather than absolute calibration, which reduces the demands on the judge and accounts for the robustness observed when the judge is a different or smaller model than the generator (Appendix D.5).

E.4 Greedy vs. Exhaustive Selection

Although the VRS objective is non-submodular, greedy selection over a small candidate pool yields near-optimal subsets in practice. We quantify the gap on 100 randomly sampled HotpotQA queries by comparing the greedy subset against an exhaustive search over all $\binom{|\mathcal{C}|}{k}$ subsets for small values ($k \in \{3, 4, 5\}$, $|\mathcal{C}| = 20$). The mean relative VRS gap between greedy and exhaustive selection is below 2% across these settings, and the resulting downstream accuracy is statistically indistinguishable. Greedy optimization is similarly used on non-submodular objectives such as MMR.

E.5 LLM-as-a-Judge Prompt

The exact judge prompt used in all experiments is reproduced below. The three criteria (coherence, relevance, query alignment) are each scored on a 1–10 scale, and Q_t is their mean divided by 10 so that $Q_t \in [0, 1]$.

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}

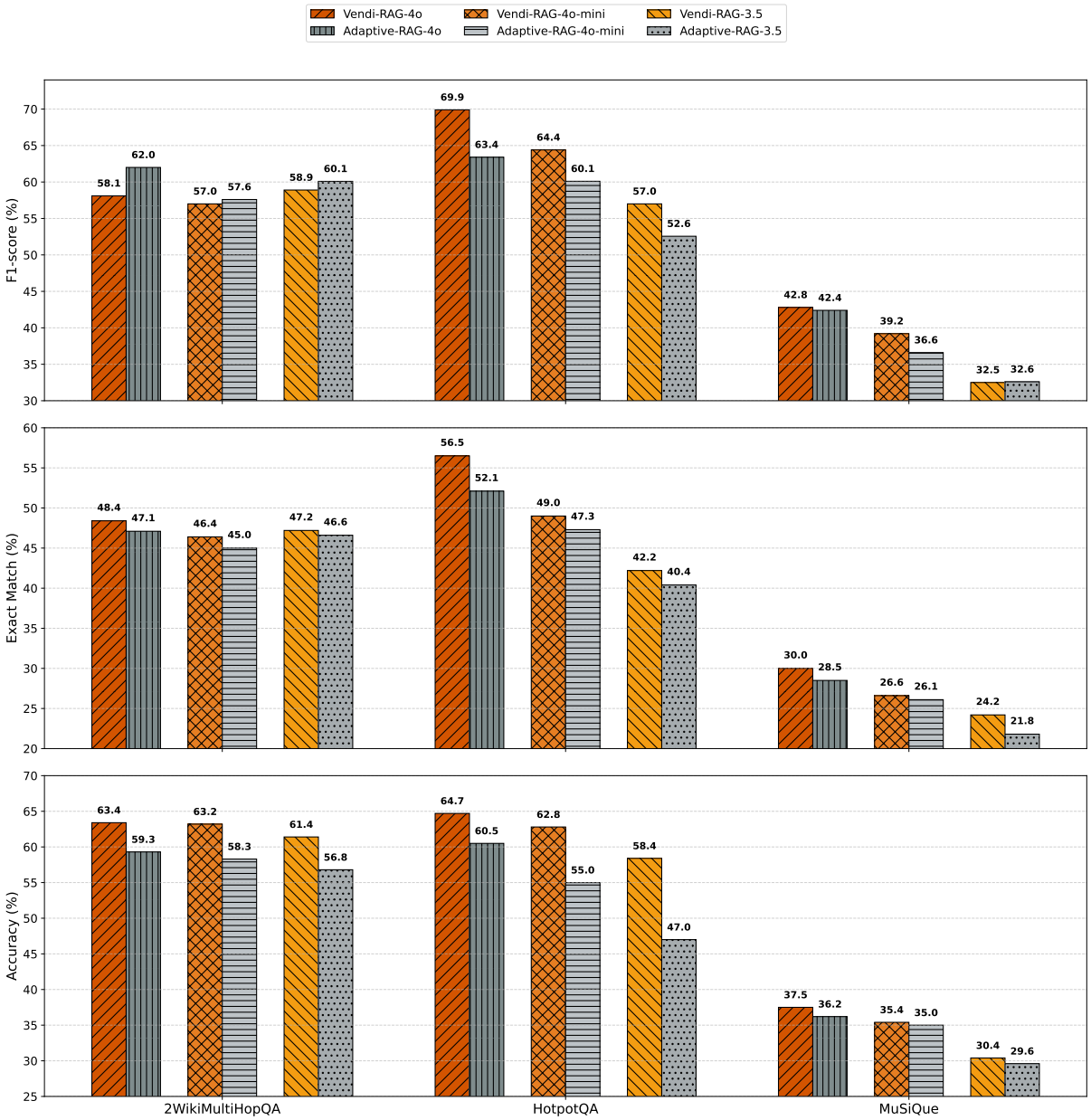

Figure 2: Performance comparison of Vendi-RAG and Adaptive-RAG across the three multi-hop QA datasets and three GPT backbones (GPT-3.5, GPT-4o-mini, GPT-4o), using EM, F1, and Accuracy. Bars are grouped by (dataset, metric); within each group the leftmost three bars are Adaptive-RAG across the three backbones and the rightmost three bars are Vendi-RAG (dynamic, $s_1 = 0.8$) across the same three backbones. Vendi-RAG with GPT-4o outperforms all other configurations on every dataset and metric, with the largest absolute gain on HotpotQA. Numeric per-cell values are in Appendix D.4.

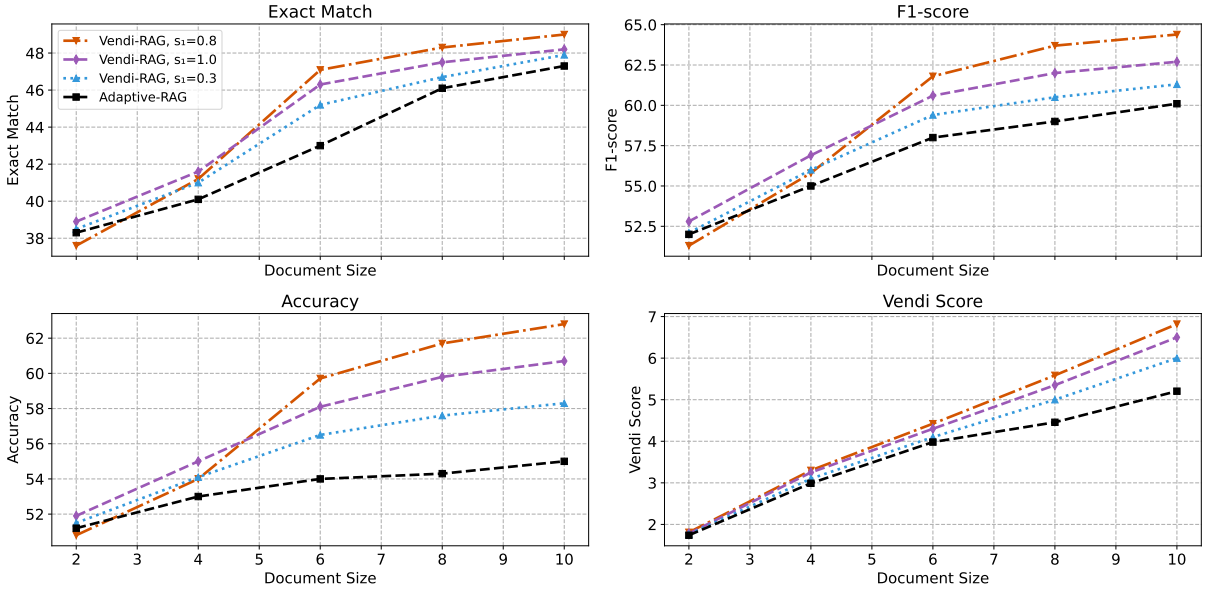

Figure 3: Performance of Vendi-RAG and Adaptive-RAG across document set sizes $k \in \{2, 3, 5, 8, 10, 15\}$, measured on HotpotQA with the `GPT-4o-mini` backbone (mean over 3 seeds; vertical axis is the metric value; horizontal axis is k). Each panel shows one metric (EM, F1, Accuracy, Vendi Score). Vendi-RAG variants are labeled by their fixed initialization value $s_1 \in \{0.3, 0.8, 1.0\}$. Vendi-RAG outperforms Adaptive-RAG across all metrics for $k \geq 3$ and the gap widens with k; the $s_1 = 0.8$ variant is the best Vendi-RAG configuration. Beyond $k = 10$ the context window starts to require document truncation, which introduces a confound for the largest-k points.