

PFE-SELF-RAG: BALANCING SELF-RAG EVALUATION METRICS VIA PARETO EFFICIENCY

F. HOSSEINI  AND M. EFTEKHARI  ✉

Article type: Research Article

(Received: 29 July 2025, Received in revised form 22 November 2025)

(Accepted: 30 January 2026, Published Online: 30 January 2026)

ABSTRACT. Self-RAG enhances Retrieval-Augmented Generation (RAG) by enabling Large Language Models (LLMs) to dynamically retrieve external knowledge and self-evaluate outputs. However, the original Self-RAG heavily relies on a manually tuned weighted-sum mechanism for combining critique scores, rendering the system brittle and poorly adaptable to diverse query distributions. To address these limitations, Pareto Front Enhanced Self-RAG (PFE-SELF-RAG) is proposed as a tuning-free Multi-Objective Optimization(MOO) framework. It first applies Maximal Marginal Relevance (MMR) to enrich context diversity, then incorporates two evaluation strategies: Pareto Front-based selection and Geometric Mean (GM) Aggregation. The primary significance of this approach lies in eliminating fragile manual weight tuning. By mathematically modeling the trade-off between factual accuracy and relevance, PFE-SELF-RAG enables adaptive candidate selection, allowing the number and quality of outputs to vary dynamically. This represents the first formal application of Pareto optimization to candidate ranking in self-reflective RAG systems, establishing a principled alternative to heuristic aggregation. Evaluations on PopQA, ARC Challenge, PubHealth, and TriviaQA demonstrate substantial impact. The Full Pareto Set strategy consistently outperforms the Self-RAG baseline, achieving %58.6 on PopQA (%+3.7), %68.0 on ARC Challenge (%+1.6), %73.0 on PubHealth (%+0.6), and %71.3 on TriviaQA (%+4.3). These improvements underscore the practical impact of replacing brittle heuristics with principled optimization, establishing PFE-SELF-RAG as a robust and scalable standard for self-reflective RAG systems.

Keywords: Retrieval-Augmented Generation, Pareto Optimization, Geometric Mean, Self-RAG, LLM Evaluation

2020 MSC: Primary 68T50, 68T07, 68T37

1. Introduction

LLMs have fundamentally advanced Natural Language Processing (NLP), demonstrating exceptional performance in tasks such as text generation, reasoning, and dialogue [5, 25]. Despite these capabilities, LLMs frequently exhibit *hallucination*, generating fluent but factually incorrect information [17]. This

✉ m.eftekhari@uk.ac.ir, ORCID: 0000-0002-0381-8225

<https://doi.org/10.22103/jmnr.2026.25661.1841>

Publisher: Shahid Bahonar University of Kerman

How to cite: F. Hosseini, M. Eftekhari, *PFE-SELF-RAG: Balancing self-RAG evaluation metrics via Pareto efficiency*, J. Mahani Math. Res. 2026; 15(2): 179-208.



© the Author(s)

limitation is particularly critical in high-stakes domains—such as healthcare, law, and scientific research—where factual accuracy is paramount.

To address this challenge, RAG has emerged as a hybrid framework that enhances factual accuracy by grounding generated outputs in external knowledge [18]. Unlike models relying solely on internal parameters, RAG retrieves relevant documents from a knowledge base to inform the generation process. This approach enables the incorporation of up-to-date, domain-specific information without requiring model retraining, thereby improving reliability in tasks such as open-domain question answering and fact verification.

Several RAG implementations have been developed, integrating retrievers like BM25 [23] or DPR [16] with generative models such as BART or T5. Advanced RAG variants further incorporate multi-hop retrieval [26] and iterative refinement [12]. While these methods improve document relevance, most conventional RAG systems retrieve a fixed number of passages regardless of necessity. This rigidity can introduce irrelevant or redundant information, potentially degrading the quality of the generated output.

To overcome these limitations, Self-RAG [2] introduces a dynamic, reflective inference approach. Unlike traditional RAG, Self-RAG enables the model to determine when retrieval is necessary and to evaluate the quality of its outputs through a self-reflection mechanism. By employing reflection tokens, Self-RAG triggers retrieval on demand and critiques its own responses, selecting the most accurate and coherent output from multiple candidates. This self-assessment enhances factuality and adaptability while preserving generative flexibility.

Despite its innovations, the inference strategy of Self-RAG is constrained by fundamental limitations. A central weakness is the reliance on a static, manually-tuned weighted sum to aggregate critique scores—a process that is brittle and fails to adapt to the unique trade-offs of each query, often leading to suboptimal selections. Additionally, certain evaluation metrics used in the process lack semantic richness and fail to comprehensively capture key attributes such as coherence, relevance, and utility. Furthermore, specific metrics have been shown to be redundant, as their removal does not result in a measurable performance decline. Collectively, these issues necessitate a more adaptive, semantically informed, and tuning-free inference mechanism.

To address these shortcomings, this paper introduces PFE-SELF-RAG, a framework that fundamentally reframes the inference process by replacing the brittle, manually-tuned weighted sum with a principled MOO approach. Innovations are incorporated at two critical stages:

- **Post-retrieval stage:** MMR [6] is employed to curate a document set that is both highly relevant and informationally diverse, directly enhancing the quality of the context provided to the generator.
- **Response selection stage:** The response selection mechanism is overhauled by introducing two distinct evaluation strategies:

- (1) **Pareto Front–based selection**, which formally models the trade-offs between factual utility and relevance.
- (2) **GM aggregation**, which offers a robust, holistic score that penalizes candidates with any single weak critique.

This dual-strategy design not only eliminates the need for sensitive hyperparameter tuning but also allows the Pareto-based method to adaptively identify a set of optimal candidates, surpassing the single-answer restriction of prior work. Consequently, greater robustness and adaptability are achieved, supporting more accurate and contextually aligned responses across a broad range of knowledge-intensive tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on RAG and Self-RAG, highlighting existing retrieval strategies and prior efforts in MOO for text generation to situate our contribution within the broader landscape. Section 3 presents the preliminaries, including the Pareto front formulation for MOO, MMR for balancing relevance and diversity, and the Self-RAG framework.

Section 4 introduces the proposed evaluation metrics for the final response selection, focusing on Pareto Front-Based selection and GM. Section 5 details the PFE-Self-RAG method, describing the full inference pipeline, MMR-based document refinement, parallel candidate generation, and adaptive response selection.

Section 6 presents the experimental results, detailing datasets, baseline models, parameter choices, and evaluation methodology. Results are analyzed quantitatively through statistical tests (Quade and Li procedures) and qualitatively via score distribution visualizations. Finally, Section 7 concludes the paper by summarizing key findings, discussing limitations, and outlining directions for future work.

2. Related work

RAG [18] has become a widely adopted paradigm for improving the factual accuracy of LLMs without resorting to costly fine-tuning. By retrieving relevant external documents and conditioning the generation process on this evidence, RAG can improve factual correctness and adapt more effectively to domain-specific tasks [16]. Recent advances such as Self-RAG [2] extend this idea by introducing a dynamic retrieval mechanism, where the model determines when external knowledge is needed and uses self-reflection tokens to critique and select the best response from multiple candidates. While Self-RAG improves adaptability and factual grounding, its response selection relies on a brittle, manually tuned weighted sum over multiple critique dimensions—a static aggregation that often fails to capture nuanced trade-offs inherent in multi-dimensional evaluation.

To address such limitations, MOO offers a principled alternative by explicitly modeling trade-offs across competing objectives [9, 21]. In the context of

aligning LLMs with diverse human preferences, He and Maghsudi [13] proposed the Pareto Multi-Objective Alignment (PAMA) algorithm, demonstrating that Pareto-based alignment yields sets of equally optimal solutions without sacrificing one objective for another. Beyond NLP, the efficacy of MOO has been demonstrated in various complex domains. For instance, [22] utilized a Deep Reinforcement Learning (DRL) framework to optimize conflicting objectives in BIM-based green building design, while [3] enhanced the Multi-Objective Cuckoo Search (MOCS) with migration operators to address scheduling trade-offs in cloud-fog computing. These studies highlight a consensus that dealing with conflicting goals explicitly via Pareto-based methods yield superior robustness compared to static aggregation. Inspired by this scalability, our PFE-SELF-RAG framework leverages explicit Pareto front construction over small candidate sets to preserve optimal trade-offs between factual utility and relevance.

In the specific domain of RAG, recent research has increasingly focused on addressing inherent trade-offs—such as latency, cost, and accuracy—through adaptive strategies. For dynamic adaptation, Adaptive-RAG [14] introduces a classifier to select retrieval strategies based on query complexity. To enhance retrieval reliability, Corrective RAG (CRAG) [27] proposes a lightweight evaluator to trigger corrective actions when the retrieved context is insufficient. Optimization efforts have also targeted the pipeline configuration; very recently, Barker et al. [4] proposed a multi-objective hyperparameter optimization framework using Bayesian optimization to find a Pareto front of system configurations. They demonstrate that optimal configurations are task-dependent, making fixed setups suboptimal in many real-world cases.

Despite these significant advancements, a comparative analysis reveals a critical gap in the inference-time candidate selection phase. Existing methods primarily operate at different stages of the pipeline. For instance, approaches like Adaptive-RAG [14] and CRAG [27] prioritize optimizing the retrieval strategy or correcting inputs prior to generation, rather than explicitly modeling trade-offs in the final generated candidates. Furthermore, frameworks such as that of Barker et al. [4] utilize MOO for offline hyperparameter tuning, which, while effective for system configuration, does not allow for dynamic, per-query trade-off resolution during inference. Meanwhile, current self-reflective frameworks like the original Self-RAG [2] continue to rely on brittle heuristics to aggregate critique scores, forcing a fixed preference that often discards Pareto-optimal candidates. PFE-SELF-RAG addresses this specific limitation by introducing a tuning-free, Pareto-based inference mechanism. Unlike the aforementioned works, our approach dynamically models the continuous trade-off between *relevance* and *factual utility* for individual generated candidates at inference time, ensuring robustness without the need for costly hyperparameter search or external corrective pipelines.

3. Preliminaries

This section outlines the foundational concepts underpinning the proposed method, specifically MOO and the Self-RAG framework.

3.1. Multi-Objective Optimization (MOO). MOO addresses optimization problems involving multiple conflicting objectives, a scenario frequently encountered in real-world applications [9, 21]. In contrast to single-objective optimization, which yields a single optimal solution, MOO produces a set of Pareto-optimal solutions. Within this set, no single solution is superior across all objectives simultaneously. These solutions collectively form the Pareto front, representing the trade-offs among objectives. The general formulation of MOO is given by:

$$(1) \quad \begin{aligned} \text{optimize: } & \mathbf{Y} = \mathbf{F}(\mathbf{X}) = [f_1(\mathbf{X}), f_2(\mathbf{X}), \dots, f_k(\mathbf{X})] \\ \text{subject to: } & g_i(\mathbf{X}) \geq 0 \quad (i = 1, 2, \dots, q) \\ & h_i(\mathbf{X}) = 0 \quad (i = q + 1, \dots, m) \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^n$ denotes the decision vector, \mathbf{Y} represents the objective vector, and the problem is subject to m constraints within an n -dimensional space [9, 21].

The primary objective in MOO is to identify non-dominated solutions that constitute the Pareto front. A solution is defined as Pareto-optimal if no alternative solution exists that improves upon it in all objectives simultaneously [20]. Search strategies, such as beam search, approximate the Pareto front by maintaining a limited set of promising solutions, referred to as the optimal beam. In this process, solutions are evaluated for dominance, and non-dominated candidates are retained. Over successive iterations, the beam evolves to enhance both solution diversity and quality, aiming for convergence toward the true Pareto front [29]. This methodology efficiently captures optimal trade-offs inherent in multi-objective problems [10].

A prevalent method for solving MOO problems is the weighted sum approach, which transforms the multi-objective problem into a single-objective formulation by aggregating weighted objectives:

$$(2) \quad \text{maximize: } y = \sum_{i=1}^k w_i f_i(\mathbf{X}),$$

where w_i represents the weight assigned to the i -th objective, typically normalized to ensure appropriate scaling [7]. Varying weight configurations yields distinct solutions. However, this method exhibits significant limitations: it generates only a subset of Pareto-optimal solutions and often fails in non-convex problem spaces, potentially missing solutions located on the true Pareto front. To mitigate these issues, the aggregated function must be optimized across various weight combinations to incrementally expand the solution set; however, this process remains computationally intensive [9].

In the context of RAG, quality is rarely one-dimensional. Responses must be both *factually supported* and *contextually relevant*; however, these goals frequently conflict when retrieved evidence is incomplete or noisy. Aggregating these qualities via a weighted sum risks over-prioritizing one objective at the expense of the other. By framing response selection as a multi-objective problem, these trade-offs can be explicitly modeled. Rather than collapsing diverse quality signals into a single, brittle scalar, Pareto analysis preserves the optimal candidates across varying balances of relevance and support. This approach concretizes the abstract notion of the Pareto front within the RAG setting, identifying answers that are optimal regarding both factual utility and contextual relevance. Consequently, it provides a principled basis for final selection under uncertainty.

3.2. Maximum Marginal Relevance (MMR). MMR is a prominent technique in information retrieval and recommendation systems, employed to optimize the balance between relevance and diversity within a selected set of items, such as documents or sentences. The primary objective of MMR is to curate a subset of items that are highly relevant to a specific query while simultaneously minimizing redundancy among them. Formally, MMR is defined with respect to a query Q , a set of candidate items D , and a subset $S \subseteq D$ representing items already selected. For any candidate item $d \in D \setminus S$, the MMR score is computed as:

$$(3) \quad \text{MMR}(d) = \lambda \cdot \text{Sim}_1(d, Q) - (1 - \lambda) \cdot \max_{d' \in S} \text{Sim}_2(d, d'),$$

where $\text{Sim}_1(d, Q)$ quantifies the relevance of item d to the query Q , $\text{Sim}_2(d, d')$ measures the similarity between items d and d' , and $\lambda \in [0, 1]$ is a parameter governing the trade-off between relevance and diversity [6].

In the context of RAG systems, retrieving documents for fact-seeking queries—such as “What is the capital of Burkina Faso?”—often yields multiple passages containing redundant information (e.g., reiterating “The capital is Ouagadougou”). Including such redundant documents in the context window is inefficient and offers diminishing informational returns. MMR addresses this inefficiency by selecting documents based on the dual criteria formulated in Equation (3):

- **Relevance (Sim_1):** The degree to which a document aligns with the users’ query.
- **Diversity (Sim_2):** The dissimilarity of a document relative to those already selected.

In practice, once the most relevant document identifying “Ouagadougou” is selected, MMR penalizes subsequent candidates that merely repeat this fact. Priority is instead assigned to documents that, while relevant, provide complementary information, such as demographic or historical details. By modulating the λ parameter, MMR constructs a context that is informationally dense and minimizes repetition, thereby enhancing the input quality for the generator.

3.3. Self-Retrieval-Augmented Generation (Self-RAG). Self-RAG is a framework designed to enhance the reliability and adaptability of LLMs by integrating dynamic retrieval, self-assessment, and iterative refinement. The architecture comprises two primary components: a generator model and a critic model. The generator, a fine-tuned LLM, interleaves text generation with self-reflection tokens (e.g., *ISREL*, *ISSUP*, *ISUSE*). As these tokens correspond to probabilities estimated by the critic model, their values are inherently normalized within the $[0, 1]$ interval. These scores are acquired via training on synthetic feedback provided by the critic.

During inference, the generator processes the input query. If the query or context provides insufficient information, the model emits a Retrieve token, triggering the retrieval of relevant documents from an external corpus. For each retrieved document d , the generator produces multiple candidate continuations, each annotated with critique tokens. The quality of these candidates is evaluated using the segment score:

$$(4) \quad f(y_t, d, \text{Critique}) = p(y_t \mid x, d, y_{<t}) + S(\text{Critique}),$$

where $f(y_t, d, \text{Critique})$ denotes the composite quality score for a generated segment y_t , conditioned on document d and the self-reflection critique. The first term, $p(y_t \mid x, d, y_{<t})$, represents the intrinsic generation probability assigned to token y_t given the input query x , the retrieved document d , and the preceding output sequence $y_{<t}$. The second term, $S(\text{Critique})$, captures the aggregated assessment from the critic model across multiple quality dimensions, ensuring that both generation likelihood and qualitative evaluation jointly influence the ranking of candidate responses. The aggregation is defined as:

$$(5) \quad S(\text{Critique}) = \sum_{G \in \mathcal{G}} w_G s_t^G, \quad \text{for } \mathcal{G} = \{\text{ISREL}, \text{ISSUP}, \text{ISUSE}\}.$$

Here, $S(\text{Critique})$ represents the weighted sum of critique scores over the evaluation dimensions \mathcal{G} . The term w_G is a static weighting coefficient associated with the critique token G , and $s_t^G \in [0, 1]$ denotes the probability assigned by the critic at generation step t . This value quantifies the critic’s confidence that the generated content satisfies specific criteria, such as relevance (*ISREL*), evidential support (*ISSUP*), or utility (*ISUSE*).

4. Proposed Metrics

This section introduces two novel evaluation metrics for the final response selection stage, both derived from the self-reflection tokens generated by the critic model.

4.1. Pareto Front-Based Selection. Standard Self-RAG aggregation integrates critique tokens (*ISREL*, *ISSUP*, and *ISUSE*) via manually tuned weights. To eliminate dependency on heuristic tuning while aligning with the principles of MOO, these tokens are transformed into two aggregate metrics:

the F_1 Score and Relevance. Formal Pareto optimization is applied to candidate response selection within the Self-RAG framework, representing a novel contribution to the domain. Consequently, the limitations inherent in simple linear aggregations are transcended by explicitly modeling the trade-offs between distinct quality dimensions.

The *ISUSE* token reflects the functional utility of the response, effectively serving as a proxy for precision by assessing how well the retrieved documents contribute to the final output. The *ISSUP* token measures the degree of evidential support, serving as an analogue to recall by indicating whether sufficient information has been retrieved to address the query. While *ISUSE* and *ISSUP* do not correspond strictly to standard information retrieval precision and recall, they represent parallel concepts within the Self-RAG architecture. Consequently, the F_1 Score is introduced to balance factual grounding and task utility, derived as the harmonic mean of these two components:

$$(6) \quad F_1 = \frac{2 \times \text{ISUSE} \times \text{ISSUP}}{\text{ISUSE} + \text{ISSUP}}.$$

In this formulation, *ISUSE* and *ISSUP* are normalized critique scores in $[0, 1]$ representing the usefulness and evidential support of the generated output, respectively. The harmonic mean is selected for its inherent property of penalizing imbalances. Unlike the arithmetic mean, which might assign a high score to a response that is useful but unsupported, the F_1 score remains low if either component is deficient. This ensures that only candidates demonstrating both high utility and strong support are ranked favorably. Furthermore, given that *ISUSE* and *ISSUP* frequently exhibit similar behaviors and high correlation, merging them into a single composite score reduces redundancy in the optimization space.

The *ISREL* token is explicitly employed as an independent metric for the assessment of passage relevance. MOO, free from heuristic weight tuning, is realized through the construction of a Pareto front, utilizing F_1 (factual utility) and F_2 (where $F_2 = \text{ISREL}$) as the objective dimensions. Formally, the finite set of k candidate responses, $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$, is denoted as being generated from the refined document set \mathcal{R} produced by MMR filtering. For each candidate c_j , a factual-utility score $F_1(c_j) \in [0, 1]$ and a relevance score $F_2(c_j) \in [0, 1]$ are defined. The MOO problem is consequently formulated as:

$$(7) \quad \begin{aligned} & \text{maximize} && (F_1(c_j), F_2(c_j)) \\ & \text{subject to} && c_j \in \mathcal{C}, \\ & && \mathcal{C} \text{ generated using documents in } \mathcal{R}, \quad |\mathcal{R}| \leq 5, \\ & && F_1(c_j), F_2(c_j) \in [0, 1] \quad \forall j. \end{aligned}$$

A candidate $c_a \in \mathcal{C}$ is defined as *Pareto optimal* if there exists no other $c_b \in \mathcal{C}$ such that $F_1(c_b) \geq F_1(c_a)$ and $F_2(c_b) \geq F_2(c_a)$, with at least one inequality being strict. Given the constraint $k \leq 5$ in our setting, this discrete optimization is solved exactly by pairwise dominance checks, yielding the Pareto set $\mathcal{P} \subseteq \mathcal{C}$. A response candidate is considered non-dominated if no other candidate achieves strictly higher scores in both metrics. This methodology eliminates reliance on hand-crafted weight combinations, enabling a mathematically principled and adaptive selection process.

We investigate two response selection strategies based on this formulation. The first strategy returns all non-dominated responses on the Pareto front, prioritizing completeness by capturing the full spectrum of optimal trade-offs between factual utility and relevance. The second strategy selects a single response by identifying the non-dominated candidate that minimizes the Euclidean distance to the ideal point $[1, 1]$ in the (F_1, F_2) objective space. This method ensures a robust, deterministic selection that simultaneously maximizes both objectives.

Given the limited cardinality of the candidate response set per query ($k = 5$), the Pareto front is approximated utilizing exhaustive pairwise comparison. Dominance is subsequently determined by comparing each candidate against every other candidate in the set. For small k , this approach is deemed computationally negligible and guarantees the identification of the true Pareto front for the candidate set, thereby obviating the need for heuristic approximations such as beam search.

4.2. Geometric Mean Aggregation (GM Aggregation). In contrast to the Pareto-based metric, which aggregates *ISSUP* and *ISUSE* into a composite F_1 score, the GM Aggregation metric retains the distinct values of all three critique tokens—*ISUSE*, *ISSUP*, and *ISREL*—combining them to ensure no quality dimension is disregarded. This approach fosters a more rigorous and a comprehensive evaluation of responses. Rather than employing a weighted sum, this method utilizes the GM, allowing each critique factor to contribute multiplicatively while preserving the granularity of token-level signals.

Since the critique scores are generated by the critic model as probabilities, they are inherently normalized to the unit interval $[0, 1]$, ensuring direct compatibility without requiring auxiliary scaling. The joint product P of the three scores is calculated as:

$$(8) \quad P = \text{ISUSE} \times \text{ISSUP} \times \text{ISREL}.$$

The Geometric Mean (GM) score is subsequently derived by taking the cube root of this product:

$$(9) \quad GM_{Score} = \sqrt[3]{P}.$$

In this formulation, each factor represents a distinct evaluation dimension: usefulness (*ISUSE*), evidential support (*ISSUP*), and relevance (*ISREL*). Formally, the response selection process is modeled as an optimization problem:

$$\begin{aligned}
& \text{maximize} && \text{GMScore}(c_j) = (\text{ISUSE}(c_j) \cdot \text{ISSUP}(c_j) \cdot \text{ISREL}(c_j))^{1/3} \\
(10) \quad & \text{subject to} && c_j \in \mathcal{C}, \\
& && \text{ISUSE}(c_j), \text{ISSUP}(c_j), \text{ISREL}(c_j) \in [0, 1],
\end{aligned}$$

where \mathcal{C} denotes the finite set of candidate responses generated for the query. The decision variable is the specific candidate c_j selected from \mathcal{C} , with the objective of maximizing the GMScore. Consequently, the optimal selection c^* is defined as:

$$(11) \quad c^* = \arg \max_{c_j \in \mathcal{C}} \text{GMScore}(c_j).$$

This formalization explicitly captures the optimization objectives, decision variables, domains, and constraints governing the GM aggregation strategy. By utilizing the GM, the method ensures that all three dimensions contribute equiproportionally while effectively penalizing candidates that exhibit weakness in any single dimension. This property yields a stricter and more balanced assessment than additive methods, mitigating the influence of isolated extreme values and guiding the model toward responses that maintain consistently high quality across all evaluation criteria.

5. Pareto-Front Enhanced SELF-RAG (PFE-SELF-RAG)

This section details the PFE-SELF-RAG inference pipeline, a principled framework designed to enhance response quality through the systematic refinement of both retrieved context and final candidate selection. The process comprises three distinct stages: Document Refinement, Parallel Generation, and Response Selection via multi-objective evaluation.

The procedure initiates with post-retrieval re-ranking using MMR, applied immediately following the initial retrieval of ten documents. The objective of MMR is to curate a refined subset of five documents that are highly relevant to the query while maximizing mutual diversity. For instance, if two retrieved documents contain the target answer but one offers additional historical background, MMR prioritizes the latter to improve contextual breadth and minimize redundancy. To govern the trade-off between relevance and diversity, the parameter λ is held constant across experiments, selected to prioritize relevance while penalizing redundancy (see Section 6.2). This initial refinement ensures that the subsequent generation and evaluation stages operate on a document set characterized by broader and more informative coverage. Subsequently, the focus shifts to the final inference stage, where candidate responses are generated using this refined document set. In contrast to the traditional Self-RAG approach, which applies a single, static weighted sum to critique tokens, our framework evaluates candidates using one of two alternative multi-objective strategies. The first, Pareto Front-based selection, identifies responses that optimally balance the F_1 and $ISREL$ (F_2) metrics by constructing a set of

non-dominated solutions. Alternatively, the GM aggregation method combines all critique tokens into a holistic GM_Score , providing a robust evaluation metric. These strategies are distinct and are not applied sequentially. This framework enables a more adaptive inference process by integrating relevance, support, and utility without reliance on manual tuning.

PFE-SELF-RAG addresses the fundamental limitations of static scoring through a dual-level adaptive mechanism. First, the MMR filtering step tailors the document context on a per-query basis. Second, the replacement of a fixed weighted sum with principled MOO allows the cardinality of the final answer set to vary according to the query and its candidate score distribution.

By identifying the optimal trade-off between factual utility (F_1) and relevance (F_2) dynamically, the system transcends rigid single-output constraints. This capability allows the output set—as seen in the Full Pareto Set strategy—to adapt to the quality of available evidence, potentially yielding multiple optimal answers rather than enforcing a restricted single output. Prior to detailing the algorithm, the symbols used in the PFE-SELF-RAG framework are defined in Table 1.

TABLE 1. Notation Used in the PFE-SELF-RAG Algorithm.

Symbol	Description
Q	The input user query.
D	The initial set of n documents retrieved for q .
R	The refined set of k documents after MMR filtering.
C	The set of k candidate responses generated from R .
c	A single candidate response, where $c \in C$.
P	The Pareto set; subset of non-dominated candidates from C .
$F_1(c)$	Factual Utility score (harmonic mean of $ISUSE$ and $ISSUP$).
$F_2(c)$	Relevance score (from the $ISREL$ token).
$GMscore(c)$	Geometric Mean score of all three critique tokens.
n	Number of documents initially retrieved (set to 10).
k	Number of documents after MMR filtering (set to 5).
λ	MMR hyperparameter balancing relevance and diversity (set to 0.9).

Algorithm 1: PFE-SELF-RAG

```

1 ENHANCED SELF-RAG Input: User query  $Q$ 
2 Output: Optimal response
3
4 // 1. Document Retrieval & MMR Filtering
4  $D \leftarrow \text{Retrieve}(Q, 10)$ 
5  $R \leftarrow \emptyset$ 
6 while  $|R| < 5$  do
7   Select  $d \in D$  maximizing
7    $\{\lambda \cdot \text{Sim}_1(d, Q) - (1 - \lambda) \cdot \max_{d' \in R} \text{Sim}_2(d, d')\}$ 
8    $R \leftarrow R \cup \{d\}; D \leftarrow D \setminus \{d\}$ 
9 end
10
11 // 2. Parallel Generation
11  $C \leftarrow \emptyset$ 
12 foreach  $d \in R$  do
13    $[\text{Resp}, \text{SELF} - \text{TOKENS}] \leftarrow \text{Generate}(Q, d)$ 
14    $\text{Resp}.F_1 \leftarrow \text{CalculateF1}(\text{SELF} - \text{TOKENS})$ 
15    $\text{Resp}.F_2 \leftarrow \text{CalculateF2}(\text{SELF} - \text{TOKENS})$ 
16    $\text{Resp}.GM \leftarrow \text{CalculateGM}(\text{SELF} - \text{TOKENS})$ 
17    $C \leftarrow C \cup \{\text{Resp}\}$ 
18 end
19
20 // 3. Pareto Analysis
20  $P \leftarrow \{r \in C \mid \nexists r' \in C : (r'.F_1 > r.F_1 \wedge r'.F_2 \geq r.F_2) \vee (r'.F_2 >$ 
20    $r.F_2 \wedge r'.F_1 \geq r.F_1)\}$ 
21
22
23 // 4. Output Selection, three following cases can be
23   selected
23 return  $\begin{cases} \text{case 1 : All } r^* & \text{if } r^* \in P \\ \text{case 2 : The Pareto Optimal closest solution to the ideal point} & r^+ \in P \\ \text{Case 3: } \text{argmax}_{c \in C} c.GM \end{cases}$ 

```

First, an initial retrieval is done such that given a user query, ten relevant documents from the knowledge base are retrieved. Then the following steps are performed based on Algorithm 1.

- (1) **MMR Filtering:** Apply MMR to select 5 refined documents that:
 - Maximize similarity between document and query
 - Minimize similarity between selected documents
- (2) **Parallel Generation:** For each of the 5 refined documents:
 - Generate candidate responses with self-reflection tokens
 - Compute F_1 and F_2 (*ISREL*) scores

- (3) **Solution Evaluation:**
- Construct Pareto optimal front using $F1/F2$ scores
 - Identify all Pareto optimal solutions
 - Compute geometric mean of self-reflection tokens across all responses
- (4) **Response Selection:** The algorithm can either select one of these items as the final solution.
- The geometric mean solution
 - The set of all answers correspond to all Pareto-optimal set.
 - Pareto-optimal solution closest to Ideal Point

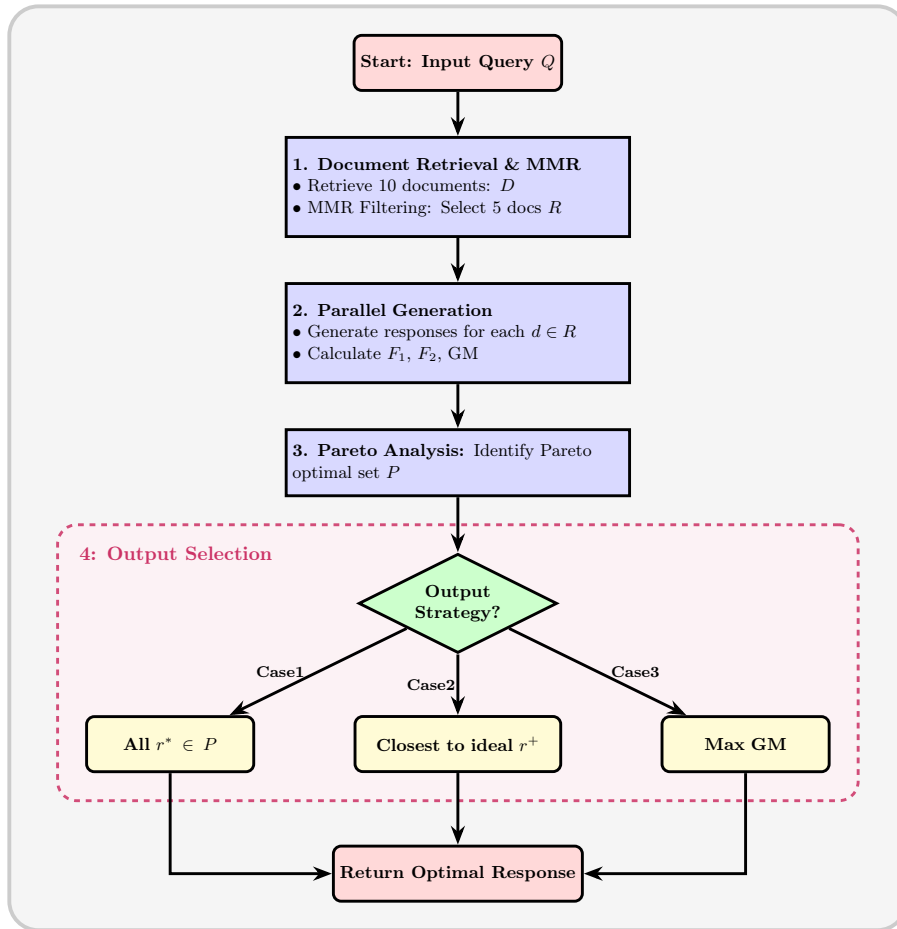


FIGURE 1. Flowchart representation of the PFE-SELF-RAG algorithm.

For further illustration, Figure 1 provides a more detailed depiction of Algorithm 1.

6. Experimental Results

The experimental evaluation of the proposed PFE-SELF-RAG framework is presented in this section. The structure commences with a detailed presentation of the datasets and task specifications, followed by a description of the baseline methods utilized for comparison. Subsequently, the evaluation methodology, based on binary matching, is introduced. A comprehensive analysis of the quantitative results is provided, complemented by a qualitative discussion supported by the visualization of score distributions.

6.1. Datasets and Tasks. To assess the generalizability and robustness of the proposed method, four diverse benchmarks were selected. These benchmarks represent distinct domains and task formats, specifically PopQA [19], TriviaQA [15], ARC-Challenge [8], and PubHealth [28]. This selection ensures that the proposed framework is validated across varied reasoning and generation patterns, rather than being limited to a single task type. All evaluations were conducted under a zero-shot paradigm. The specific characteristics of each dataset are as follows:

Fact Verification and Scientific Reasoning.

- **PubHealth** [28]: A domain-specific fact verification dataset consisting of expert-curated public health claims (e.g., “Vaccines cause autism”). The task requires classifying claims as *Supported* or *Not Supported* based on biomedical evidence, utilizing accuracy as the primary evaluation metric.
- **ARC-Challenge** [8]: A multiple-choice question-answering benchmark derived from grade-school science exams (e.g., “Which force causes tides?”) The test set, comprising 1,172 examples, is utilized. This set is specifically designed to assess complex reasoning capabilities over scientific concepts.

Short-Form Open-Domain Question Answering.

- **PopQA** [19]: The utilized dataset constitutes an open-domain benchmark comprising 14,000 factual queries. Evaluation of performance is focused on the long-tail subset of 1,399 rare-entity questions (e.g., “What is the capital of Burkina Faso?”) to specifically assess knowledge less likely to be memorized by the model.
- **TriviaQA** [15]: This dataset is designed for the evaluation of retrieval and reasoning capabilities over textual evidence. From the original set of 11,313 questions, a long-tail subset comprising 3,000 rare-entity questions was sampled. This selection was conducted to rigorously test

the model’s ability to effectively utilize retrieved documents for query answering.

6.2. Parameter Setup and Implementation. PFE-SELF-RAG is implemented strictly as an inference-time enhancement to the standard Self-RAG framework, necessitating no additional model training or fine-tuning. All experiments were conducted on the Kaggle platform utilizing dual NVIDIA T4 GPUs (15 GB VRAM each). The software environment leveraged Hugging Face Transformers for model orchestration, SentenceTransformers for embedding computation, and vLLM for efficient large language model serving. Consistent with the inference-only design, no PyTorch-based training loops were executed.

The `selfrag_llama2_7b` checkpoint was employed for the processes of both candidate generation and critique scoring. This model is responsible for the concurrent generation of response candidates and the required self-reflection tokens (*ISUSE*, *ISSUP*, *ISREL*). In the MMR filtering stage, query and document embeddings were computed using the `all-MiniLM-L6-v2` model from the SentenceTransformers library. Cosine similarity was utilized to quantify both query–document relevance and inter-document similarity within the MMR calculation.

A critical hyperparameter in PFE-SELF-RAG is the MMR trade-off coefficient, λ , which governs the balance between relevance and diversity during post-retrieval selection. Higher values of λ bias the ranking toward documents with high query relevance, whereas lower values prioritize diversity to minimize redundancy. The hyperparameter λ was empirically evaluated across the range [0.5, 0.95] utilizing validation instances derived from the target tasks. Results indicated that the value $\lambda = 0.9$ yielded optimal performance consistently across PopQA, TriviaQA, ARC-Challenge, and PubHealth. This configuration prioritizes strong topical alignment while retaining sufficient diversity to capture complementary evidence, ensuring the refined document set is both precise and informative.

The retrieval parameters—initial pool size ($n = 10$) and refined set size ($k = 5$)—were selected to optimize the trade-off between retrieval recall and computational efficiency. The time complexity of the framework is dominated by the parallel generation phase, which scales linearly with k (i.e., $\mathcal{O}(k \cdot T_{\text{gen}})$). While increasing k may marginally enhance the probability of including relevant evidence, it imposes a substantial penalty on inference latency. Preliminary evaluations confirmed that the $10 \rightarrow 5$ reduction provides a robust performance baseline, maintaining manageable inference costs without compromising output quality.

6.3. Baseline Comparisons. To provide a comprehensive evaluation, comparisons are made against baselines that incorporate external information through two distinct paradigms.

Test-Time Augmentation (Standard RAG). Standard RAG implementations [18] utilizing Llama2-7B [24] and Alpaca-7B [11] are employed. In these configurations, retrieved documents are prepended to the input query to facilitate retrieval-augmented inference without altering the underlying model weights. Additionally, these models are evaluated in their standard closed-book configurations (without retrieval) to establish a performance baseline that isolates the specific contribution of external knowledge.

Training-Time Retrieval Integration (Self-RAG). Self-RAG-7B [2], a model trained to seamlessly integrate retrieval into the generation process, is assessed. This framework employs special reflection tokens to dynamically trigger retrieval requirements and aggregates passages during inference via a weighted sum of learned relevance and utility metrics.

This mechanism enables the model to critically evaluate the quality of retrieved content and adaptively synthesize the final response.

6.4. Evaluation Methodology. The performance of the proposed framework and baselines is evaluated using a binary matching criterion. Under this protocol, a generated prediction is classified as correct if it corresponds to at least one of the provided ground-truth answers. This approach explicitly accounts for instances where multiple valid responses exist. The final accuracy is computed as the ratio of correctly matched predictions to the total number of evaluated samples.

TABLE 2. The average accuracy of different methods across PopQA, ARC, PubHealth, and TriviaQA datasets over 30 runs.

Category	Model	PopQA	ARC	PubHealth	TriviaQA
Without Retrieval	Alpaca 7B	23.6	45.0	49.8	54.5
	LLaMA2 7B	14.7	21.8	34.2	30.5
With Retrieval	Alpaca 7B + Retrieval	46.7	48.0	40.2	64.1
	LLaMA2 7B + Retrieval	38.2	48.0	30.0	42.5
	SELF-RAG 7B	54.9	66.4	72.4	67.0
	Pareto Front-Based (Closest to Ideal Point)	55.75	67.4	72.4	67.0
	GM	55.8	67.4	72.4	65.7
	Full Pareto Set	58.6	68.0	73.0	71.3

6.5. Results and Discussion. The predictive performance of Llama2-7B, Alpaca-7B, Self-RAG-7B, and the proposed PFE-SELF-RAG variants is evaluated across the PopQA, ARC-Challenge, PubHealth, and TriviaQA datasets. Assessments are conducted under both non-augmented (closed-book) and augmented (retrieval-based) configurations. Table 2 reports the mean accuracy for each method, calculated over 30 independent experimental runs. To facilitate comparative analysis, Figure 2 provides a visual representation of the quantitative data summarized in Table 2.

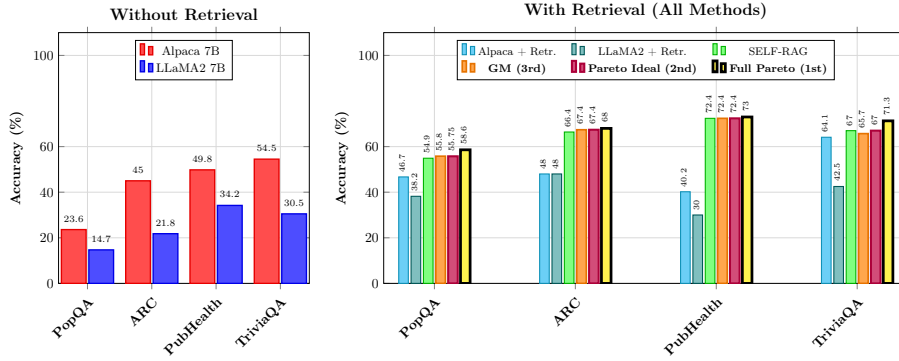


FIGURE 2. Performance comparison: (Left) Baseline methods without retrieval. (Right) All retrieval-augmented methods, with Full Pareto Set achieving the best results, followed by Pareto Ideal and GM.

In the non-retrieval setting, Alpaca-7B consistently outperformed Llama2-7B across all datasets. This disparity indicates that Alpaca-7B possesses superior baseline reasoning and generation capabilities when operating without access to external information.

Upon the integration of retrieval augmentation, all models exhibited substantial performance improvements. Notably, the accuracy of Alpaca-7B on PopQA nearly doubled, while Llama2-7B demonstrated significant gains, particularly on the ARC-Challenge. Among the retrieval-based methods, Self-RAG-7B established a robust benchmark, achieving scores of 54.9 on PopQA, 66.4 on ARC, 72.4 on PubHealth, and 67.0 on TriviaQA.

The proposed refinement strategies utilizing Pareto optimization yielded superior results. Both the Pareto Front-Based metric and the GM metric surpassed the Self-RAG baseline on PopQA, TriviaQA, and ARC-Challenge, while matching its performance on PubHealth.

Most notably, the strategy leveraging the Full Pareto Set (Table 2) achieved the highest aggregate scores. This result provides strong empirical validation for the proposed adaptive candidate selection mechanism. By retaining all non-dominated options, the model avoids the premature discarding of correct answers that represent distinct yet equally valid trade-offs. This adaptive behavior—wherein the cardinality of the final output set varies according to the query-specific score distribution rather than being fixed—constitutes a significant operational advantage over the rigid, single-output strategy of Self-RAG and is directly responsible for the superior performance observed.

While both the Pareto-based and GM methods consistently outperformed the Self-RAG baseline, the Full Pareto Set strategy exhibited the most significant and consistent gains, particularly on open-ended generation tasks such

as PopQA and TriviaQA, where multiple valid answer formulations frequently coexist. The GM approach proved to be a highly competitive alternative, often securing the second-highest rank and demonstrating robustness as a single-point evaluation metric.

The subsequent subsections present a non-parametric statistical test to validate the significance of these results, followed by illustrative figures demonstrating the qualitative superiority of the Full Pareto Set approach compared to the GM, Ideal Point, and Self-RAG variants.

6.5.1. Statistical Comparison of Models. To rigorously validate the performance disparities observed in Table 2, the Quade test [1] is employed. This non-parametric statistical method is designed for the comparison of multiple related groups (models) across distinct blocks (datasets).

The Quade test is particularly appropriate for this analysis as it not only considers the ordinal ranks of models within each dataset but also weights the blocks based on the magnitude of performance variation observed [1]. This weighting mechanism assigns greater significance to datasets where distinctions between model performances are most pronounced.

The null hypothesis (H_0) posits that no significant difference exists in the average accuracies of the compared models. Upon the detection of a significant difference (i.e., the rejection of H_0), a post-hoc analysis is conducted using the Li test. This procedure facilitates pairwise comparisons to identify specific models that significantly outperform others while controlling for statistical error. This two-step validation process formally determines whether the observed superiority of the proposed PFE methods is statistically significant.

The Quade test was conducted to assess the differences in average accuracy across all evaluated models. The analysis yielded a statistically significant difference in performance, producing an F -statistic of 12.40 and a p -value of approximately 3.34×10^{-6} . With 7 and 21 degrees of freedom ($F(7, 21) = 12.40, p < 0.001$), the null hypothesis that all models perform equally is confidently rejected.

The average rankings derived from the Quade test are presented in Table 3, illustrating a clear performance hierarchy. The proposed Full Pareto Set method achieved the premier rank (1.0), confirming its consistent superiority across the datasets. The alternative PFE variants, GM (2.6) and Pareto Closest (2.75), also secured high rankings, outperforming the baseline SELF-RAG 7B (3.65). In contrast, the non-retrieval baselines, LLaMA2 7B (7.8) and Alpaca 7B (6.5), received the lowest ranks.

TABLE 3. Average Rankings of the algorithms (Quade)

Algorithm	Ranking
Alpaca 7B	6.5
LLaMA2 7B	7.8
Alpaca 7B + Retrieval	5.4
LLaMA2 7B + Retrieval	6.3
SELF-RAG 7B	3.65
Pareto Closest	2.75
GM	2.6
Full Pareto Set	1

A post-hoc analysis, utilizing the Li test, was performed to identify the specific models significantly outperformed by the best performing method. The Full Pareto Set was designated as the control model for this comparison. As detailed in Table 4, the Li procedure establishes that models with a corresponding p -value of ≤ 0.018307 are determined to perform significantly worse than the control.

TABLE 4. Post Hoc comparison Table for $\alpha = 0.05$. Li’s procedure rejects those hypotheses that have an unadjusted $P_{Li} \leq 0.018307$.

Algorithm	$z = (R_0 - R_i)/SE$	p	P_{Li}
LLaMA2 7B	1.915683	0.055405	0.018307
Alpaca 7B	1.549449	0.121274	0.018307
LLaMA2 7B + Retrieval	1.493106	0.135409	0.018307
Alpaca 7B + Retrieval	1.23956	0.215138	0.018307
SELF-RAG 7B	0.746553	0.455333	0.018307
Pareto Closest	0.493007	0.622008	0.018307
GM	0.450749	0.652171	0.05

The post-hoc analysis indicates that, despite a distinct performance trend evident in the rankings, pairwise comparisons among the top-performing retrieval models do not universally meet the strict threshold for statistical significance. Nonetheless, the omnibus Quade test result provides robust statistical validation that the proposed PFE-based methods—specifically the Full Pareto Set approach—represent a significant improvement over both the SELF-RAG baseline and standard retrieval-augmented architectures. While the Full Pareto Set strategy achieves the highest aggregate ranking, the post-hoc analysis reveals no statistically significant divergence between it and the GM method. This performance proximity between the two leading strategies warrants a more granular graphical comparison to elucidate their respective advantages.

6.5.2. *Qualitative Comparison of PFE Methods.* Figure 3 illustrates the T-SNE visualization of document embeddings relative to the query embedding within a two-dimensional space. The process initiates with the retrieval of 10 documents using the standard Self-RAG protocol. Two distinct configurations are depicted:

- **Full Retrieval (filled blue circles):** The complete set of 10 retrieved documents is utilized as input to the model.
- **MMR Selection (red-circled):** A post-retrieval refinement step is applied using the MMR algorithm to select a subset of 5 documents. This selection prioritizes candidates that maximize both query relevance and mutual diversity.

The efficacy of MMR refinement is visually substantiated by the spatial distribution of the red-circled documents. Compared to the unrefined set, these selected documents exhibit greater proximity to the query centroid (yellow point) and increased dispersion, indicating enhanced relevance and reduced redundancy.

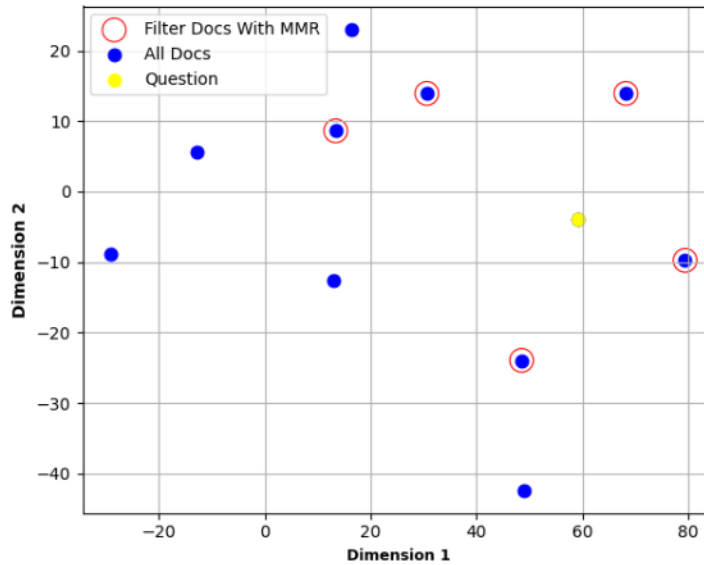


FIGURE 3. The visualization of documents and query in two dimensional space by t-SNE. The impact of refining documents using MMR can be seen in the red-circled documents, which are diverse and closer to the yellow point (query).

As illustrated in Figure 4 and Figure 5, a comparative analysis of the inference strategies is performed. Each of the ten retrieved documents is processed by the `selfrag_llama2_7b` model to generate self-reflection tokens, from which F_1 and F_2 scores are derived according to the procedures defined in Section 3. Candidate answers are projected into a two-dimensional objective space to facilitate visual comparison. The left panels depict candidates generated from the five documents selected via MMR refinement, evaluated using the proposed Pareto and GM strategies. The right panels display candidates derived from the full set of ten documents, evaluated using the baseline Self-RAG inference method. Selected answers are distinguished by unique markers. The following analysis discusses the specific divergences observed between these strategies.

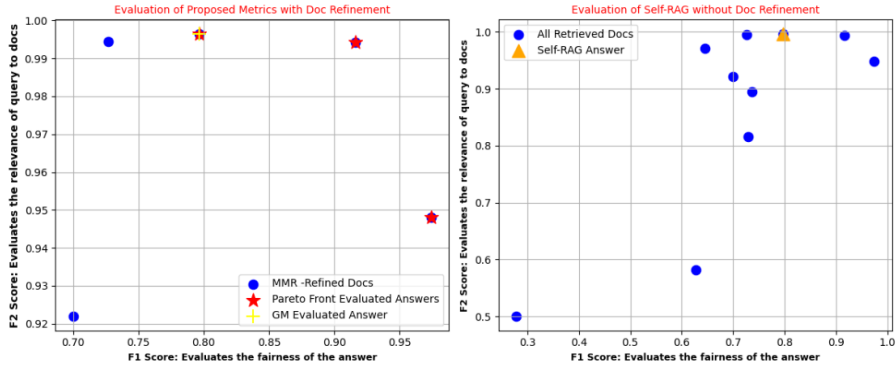
In Figure 4a, the Pareto Set strategy demonstrates a distinct advantage in completeness by identifying three non-dominated answers within the (F_1, F_2) score space. These selections encompass all correct responses present in the MMR-refined set. While one selection overlaps with the choice made by both GM and Self-RAG, the Pareto method uniquely recovers two additional correct answers overlooked by the baseline. This result illustrates the capacity of the Full Pareto Set strategy to retain distinct yet equally valid trade-offs between factual utility and relevance, thereby ensuring maximal coverage of optimal solutions in scenarios where multiple correct formulations exist.

Figure 4b highlights a more pronounced divergence: the baseline Self-RAG weighted-sum aggregation selects a plausible but incorrect answer, whereas both the GM and Pareto strategies identify the correct response. This failure stems from the compensatory nature of fixed weightings, where a high score in one dimension can mask severe deficits in others. Consequently, a candidate that is highly relevant but evidentially unsupported may be favored over a more balanced, factually correct candidate. Unlike the weighted-sum approach, the GM and Pareto methods enforce strict trade-offs—either multiplicatively or through dominance analysis—ensuring that low performance in any critical dimension significantly penalizes the candidate’s overall ranking.

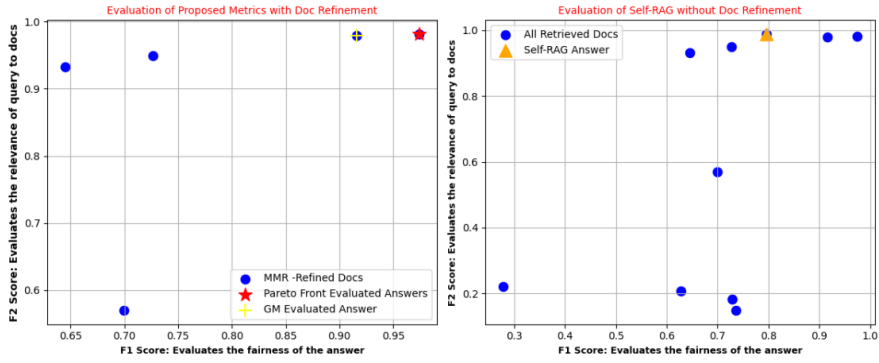
In Figure 5a, both the GM and Pareto strategies converge on the same correct answer, while Self-RAG selects a suboptimal candidate. This agreement between the two proposed strategies underscores their reliability in producing consistently high-quality outputs, even amidst varying score distributions. The independent convergence of these methods on the same optimal choice reinforces the robustness of multi-objective evaluation compared to static aggregation.

Finally, Figure 5b presents a detailed analysis of the Self-RAG baseline without document refinement. The selected answer exhibits high relevance ($F_2 \approx 0.90$), driven by a critique score of ($ISREL = 0.88$), but possesses low factual utility ($F_1 \approx 0.3$) due to poor evidential support ($ISSUP = 0.10$) and utility ($ISUSE = 0.40$). This selection exemplifies the fundamental limitation of weighted-sum aggregation: the high relevance score disproportionately influences the final metric, effectively compensating for the lack of support.

Visually, the selected answer lies far from the ideal point $(1, 1)$, which represents the optimal balance of accuracy and relevance. This case demonstrates the brittleness of manually tuned, static weights, which fail to adapt to query-specific distributions and allow the selection of fluent but factually groundless responses.



(A)



(B)

FIGURE 4. Visualizing the F_1 and F_2 scores (Part 1).

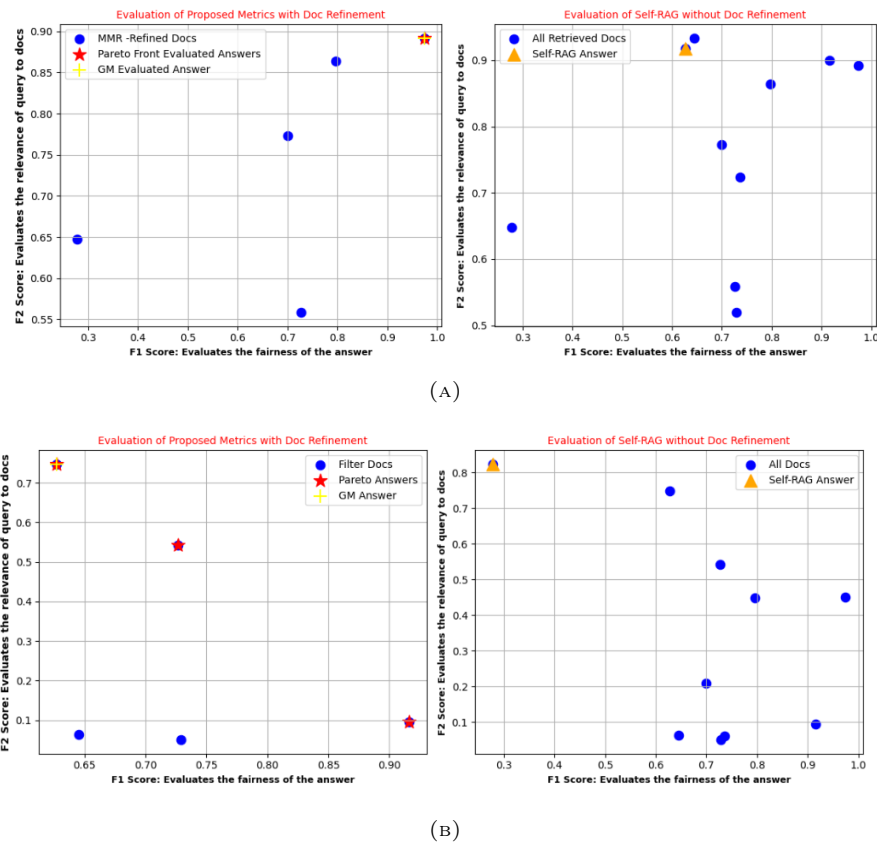


FIGURE 5. Visualizing the F_1 and F_2 scores (Part 2).

6.5.3. *Ablation study.* To isolate and quantify the specific contribution of the MMR document refinement stage, an ablation study was conducted on the three proposed selection strategies. In this experimental configuration, the MMR post-processing step was omitted, and the generator was applied directly to the top-10 documents retrieved by the initial search, devoid of diversity-aware re-ranking.

The empirical results of this ablation are summarized in Table 5. The exclusion of MMR resulted in substantial performance decrements for the Pareto Closest strategy, with reductions of 14.35 points on PopQA, 17.2 points on ARC-Challenge, and 14.3 points on TriviaQA. These significant declines underscore the strategy’s high dependence on a curated, diverse context to maintain accuracy. Conversely, the GM method exhibited greater resilience to unrefined contexts, registering minimal fluctuations: a decrease of 1.3 points on PopQA

, 0.4 points on ARC-Challenge and 0.7 points on TriviaQA. This stability suggests that GM aggregation offers robust performance even in the presence of less curated input.

The Full Pareto Set strategy, which achieved the highest accuracy in the full pipeline, experienced moderate performance reductions in the absence of MMR: -0.6 points on PopQA, -4.0 points on ARC-Challenge, and -6.4 points on TriviaQA. These results confirm that while the method remains functional without refinement, the benefits of contextual diversity are most pronounced in open-ended generation tasks. Across all evaluated strategies, performance on the PubHealth dataset remained invariant. This stability is likely attributable to the dataset’s constrained, fact-centric verification format, where the influence of informational diversity is negligible compared to direct evidential support.

Collectively, these findings corroborate the conclusion that the MMR mechanism serves as a critical determinant in maximizing performance, particularly for open-domain generation tasks such as PopQA and TriviaQA.

TABLE 5. Ablation study on the impact of MMR for the proposed methods. Accuracy (%) is reported with and without MMR. The Δ Change row shows the drop in accuracy when MMR is removed.

Method	Configuration	PopQA	ARC	PubHealth	TriviaQA
Pareto Closest	With MMR	55.75	67.4	72.4	67.0
	Without MMR	41.4	50.2	72.4	52.7
	Δ Change	-14.35	-17.2	0.0	-14.3
GM	With MMR	55.8	66.9	72.4	67.3
	Without MMR	54.5	66.5	72.4	66.6
	Δ Change	-1.3	-0.4	0.0	-0.7
Full Pareto Set	With MMR	58.6	67.8	73.0	71.3
	Without MMR	58.0	63.8	73.0	64.9
	Δ Change	-0.6	-4.0	0.0	-6.4

6.6. Computational Cost Analysis. Let n be the initial number of retrieved documents (10 in our case) and k be the number of documents selected by MMR for parallel generation (5 in our case). Let T_{ret} be the time for the initial retrieval call and T_{gen} be the time for a single **Generate** call.

- (1) **MMR Filtering (lines 6–9):** This loop runs k times. Inside the loop, it iterates over the remaining $n - |R|$ documents in D . The inner max operation iterates over the $|R|$ documents already in R . The complexity is

$$O(k \cdot n \cdot k) = O(n \cdot k^2).$$

Since n and k are small, fixed constants, this step is very fast in practice.

- (2) **Parallel Generation (lines 11–18):** This loop iterates k times. The dominant operation is `Generate(q, d)`. Although done in parallel, the total computational work is k times the work of one generation. The complexity is

$$O(k \cdot T_{\text{gen}}).$$

This is the primary computational bottleneck of the entire algorithm.

- (3) **Pareto Analysis (line 20):** This step involves pairwise comparisons among the k candidates in C . Each of the k candidates is compared against the other $k - 1$ candidates, resulting in a complexity of

$$O(k^2).$$

- (4) **Output Selection (line 23):** Finding the solution closest to the ideal point or the argmax of the GM score takes

$$O(k)$$

time.

Overall Time Complexity:

$$T_{\text{ret}} + O(n \cdot k^2) + O(k \cdot T_{\text{gen}}) + O(k^2).$$

Since T_{gen} is significantly larger than all other operations and n and k are small constants, the overall time complexity is dominated by the generation step:

$$O(k \cdot T_{\text{gen}}).$$

Space Complexity Analysis

Let L_{resp} be the maximum length of a generated response.

- (1) D and R store references to n and k documents, respectively. Assuming fixed-size embeddings or references, this requires

$$O(n)$$

space.

- (2) C stores k full response objects. The space required is

$$O(k \cdot L_{\text{resp}}).$$

- (3) P stores a subset of C , so its space is also bounded by

$$O(k \cdot L_{\text{resp}}).$$

Overall Space Complexity: The space is dominated by the need to hold the k generated candidate responses in memory:

$$O(k \cdot L_{\text{resp}}).$$

6.7. Discussion on Generalizability and Limitations. A distinguishing characteristic of the PFE-SELF-RAG framework is its robustness across varied operational conditions. The experimental validation presented herein utilized models at the 7B-parameter scale, a constraint imposed by available computational resources. While distinct performance gains were demonstrated at this scale, the evaluation of larger architectures (e.g., 13B and 70B parameters) remains a critical avenue for future research. Nevertheless, it is hypothesized that the fundamental advantages of the proposed framework are largely model-agnostic. Because PFE-SELF-RAG supplants heuristic, weighted-sum aggregation with a principled, tuning-free MOO mechanism, the observed improvements are expected to persist or potentially amplify when applied to larger LLMs capable of producing higher-quality base generations and more reliable critique tokens.

Furthermore, while the efficacy of any retrieval-augmented system is inherently contingent upon the quality of retrieved evidence, the proposed framework exhibits enhanced resilience to noisy corpora. The integration of MMR filtering minimizes the inclusion of redundant or marginally relevant documents. Concurrently, the subsequent Pareto and GM selection strategies direct the system toward responses that maximize evidential support, thereby mitigating susceptibility to errors stemming from low-quality retrieved passages.

7. Conclusion

This work presents PFE-SELF-RAG, an inference-time framework designed to enhance the performance of RAG by supplanting the brittle, manually tuned weighted sum mechanism of the original Self-RAG with a principled, MOO approach. By leveraging two distinct evaluation strategies—Pareto Front-Based selection and GM Aggregation—the framework robustly balances factual utility and relevance without necessitating hyperparameter tuning.

Experimental results across four diverse benchmarks demonstrate that PFE-SELF-RAG consistently surpasses the Self-RAG baseline in both accuracy and reliability. The Full Pareto Set strategy proved particularly effective, achieving peak accuracy on open-ended generation tasks, specifically PopQA (%58.6) and TriviaQA (%71.3). These findings validate the efficacy of adaptively retaining all optimal candidates. Furthermore, consistent improvements on the ARC-Challenge and PubHealth datasets underscore the generalizability of the method across domains requiring scientific reasoning and fact verification.

The analysis further highlights the potential misalignment between conventional scoring metrics and semantic correctness. By strictly targeting the post-retrieval selection phase, PFE-SELF-RAG provides a scalable, tuning-free mechanism that strengthens the factual grounding and relevance of generated outputs, thereby establishing a more reliable paradigm for RAG-based systems.

For future work, developing semantically aware scoring methods could enable more accurate alignment with the intended meaning of generated responses.

Further improvements may be achieved by integrating advanced post-retrieval refinement techniques, particularly reflective prompt strategies and the inclusion of additional reflection tokens beyond the baseline SELF-RAG setup, to enhance coherence and evaluation accuracy. Moreover, examining the scalability of PFE-SELF-RAG to larger language models (e.g., 13B and 70B parameters) and assessing its robustness across diverse domains constitute valuable directions for continued investigation.

8. Author Contributions

F. Hosseini: Conceptualization, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization.

M. Eftekhari: Conceptualization, validation, formal analysis, investigation, resources, data curation, writing—review and editing, visualization, supervision, project administration.

All authors have read and agreed to the published version of the manuscript.

9. Data Availability Statement

The datasets used in this study are publicly available. Their names and links are as follows:

ARC (AI2 Reasoning Challenge) is available at: <https://allenai.org/data/arc>

PubHealth dataset is available via Hugging Face: <https://huggingface.co/datasets/bigbio/pubhealth>

PopQA dataset is available via Hugging Face: <https://huggingface.co/datasets/akariasai/PopQA>

TriviaQA dataset is available via Hugging Face: https://huggingface.co/datasets/mandarjoshi/trivia_qa No new datasets were generated during the current study.

10. Acknowledgement

We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

11. Ethical considerations

The study does not involve any ethical concerns.

12. Conflict of interest

The authors declare no conflict of interest.

References

- [1] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., et al. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3), 307–318. <https://doi.org/10.1007/s00500-008-0323-y>
- [2] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511. <https://doi.org/10.48550/arXiv.2310.11511>
- [3] BahraniPour, F., Farshi, M., & Ebrahimi Mood, S. (2025). Enhanced multi-objective cuckoo search with migration operator for benchmark optimization and IoT task scheduling in cloud-fog computing. *The Journal of Supercomputing*, 81(8), 1024. <https://doi.org/10.1007/s11227-025-07531-0>
- [4] Barker, M., Bell, A., Thomas, E., Carr, J., Andrews, T., & Bhatt, U. (2025). Faster, cheaper, better: Multi-objective hyperparameter optimization for LLM and RAG systems. arXiv preprint arXiv:2502.18635. <https://doi.org/10.48550/arXiv.2502.18635>
- [5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Neelakantan, A. (2020). Language models are few-shot learners [Preprint]. arXiv preprint arXiv:2005.14165. <https://doi.org/10.48550/arXiv.2005.14165>
- [6] Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336. <https://doi.org/10.1145/290941.291025>
- [7] Chankong, V., & Haimes, Y. Y. (2008). *Multiobjective Decision Making: Theory and Methodology*. Dover Publications.
- [8] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafford, O. (2018). Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457. <https://doi.org/10.48550/arXiv.1803.05457>
- [9] Deb, K. (2011). Multi-objective optimization using evolutionary algorithms: An introduction. Technical Report No. 2011003, IIT Kanpur.
- [10] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197. <https://doi.org/10.1109/4235.996017>
- [11] Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., & Hashimoto, T. B. (2024). AlpacaFarm: A simulation framework for methods that learn from human feedback. arXiv preprint arXiv:2305.14387.

- <https://doi.org/10.48550/arXiv.2305.14387>
- [12] Guu, K., Lee, K., Tung, Z., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. arXiv preprint arXiv:2002.08909. <https://doi.org/10.48550/arXiv.2002.08909>
- [13] He, Q., & Maghsudi, S. (2025). Pareto multi-objective alignment for language models. arXiv preprint arXiv:2508.07768. <https://doi.org/10.48550/arXiv.2508.07768>
- [14] Jeong, S., Baek, H., Cho, S., Hwang, S. J., & Park, J. C. (2024). Adaptive-RAG: Synergy of retrieval and generation via adaptive strategies. arXiv preprint arXiv:2403.14403. <https://doi.org/10.48550/arXiv.2403.14403>
- [15] Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551. <https://doi.org/10.48550/arXiv.1705.03551>
- [16] Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906. <https://doi.org/10.48550/arXiv.2004.04906>
- [17] Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840. <https://doi.org/10.48550/arXiv.1910.12840>
- [18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401. <https://doi.org/10.48550/arXiv.2005.11401>
- [19] Mallen, A., Asai, A., Hajishirzi, H., Choi, E., & Khashabi, D. (2023). PopQA: An open-domain question answering benchmark for entity-centric long-tail queries. Proceedings of EMNLP, 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
- [20] Marler, R. T., & Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization, 26(6), 369–395. <https://doi.org/10.1007/s00158-003-0368-6>
- [21] Miettinen, K. (1999). Nonlinear Multiobjective Optimization. Springer. <https://doi.org/10.1007/978-1-4615-5563-6>
- [22] Pan, Y., Shen, Y., Qin, J., & Zhang, L. (2024). Deep reinforcement learning for multi-objective optimization in BIM-based green building design. Automation in Construction, 166, 105598. <https://doi.org/10.1016/j.autcon.2024.105598>
- [23] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. <https://doi.org/10.1561/1500000019>
- [24] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023).

- Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>
- [25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>
- [26] Xiong, W., Li, X. L., Iyer, S., Du, J., Lewis, P., Wang, W. Y., Mehdad, Y., Yih, W.-t., Riedel, S., Kiela, D., & Oguz, B. (2020). Answering complex open-domain questions with multi-hop dense retrieval. arXiv preprint arXiv:2009.12756. <https://doi.org/10.48550/arXiv.2009.12756>
- [27] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation. arXiv preprint arXiv:2401.15884. <https://doi.org/10.48550/arXiv.2401.15884>
- [28] Zhang, Y., Li, I., Swayamdipta, S., Choi, Y., & Smith, N. A. (2023). PubHealth: A dataset for fact verification in public health claims. Proceedings of the 61st Annual Meeting of the ACL, 12345–12360. <https://doi.org/10.18653/v1/2022.findings-naacl.1>
- [29] Zitzler, E., Deb, K., & Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation*, 8(2), 173–195. <https://doi.org/10.1162/106365600568202>

FATEMEH Hosseini

ORCID NUMBER: 0009-0000-9702-7581

DEPARTMENT OF COMPUTER ENGINEERING

SHAHID BAHONAR UNIVERSITY OF KERMAN

KERMAN, IRAN

Email address: fatemeh.hosseini.g108@gmail.com

MAHDI Eftekhari

ORCID NUMBER: 0000-0002-0381-8225

DEPARTMENT OF COMPUTER ENGINEERING

SHAHID BAHONAR UNIVERSITY OF KERMAN

KERMAN, IRAN

Email address: m.eftekhari@uk.ac.ir