Faithfulness-Aware Multi-Objective Context Ranking for Retrieval-Augmented Generation

Tian Guan , Sebastian Sun , Bolin Chen *

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1983.v1

Keywords: retrieval-augmented generation; multi-objective optimization; context ranking; faithfulness; hallucination reduction

 doi:10.20944/preprints202512.1983.v1

Article

Faithfulness-Aware Multi-Objective Context Ranking for Retrieval-Augmented Generation

Tian Guan [1], Sebastian Sun [2] and Bolin Chen [3,*]

[1] University of California, Irvine, Irvine, USA
[2] University of Wisconsin-Madison, Madison, USA
[3] Duke University, Durham, USA
* Correspondence: bolichen97@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for enhancing Large Language Models (LLMs) with external knowledge. However, existing ranking methods in RAG pipelines primarily optimize for document-query relevance, neglecting crucial factors for generation quality such as factual consistency and information coverage. We propose a novel multi-objective ranking framework that explicitly models three critical dimensions: relevance, coverage, and faithfulness support. Unlike traditional IR-centric approaches, our method introduces a utility-based scoring mechanism that evaluates each document's contribution to reducing hallucinations, improving answer completeness, and maintaining relevance. We formulate the ranking problem as a multi-objective optimization task and employ listwise learning with carefully constructed utility labels derived from existing QA datasets. Extensive experiments on Natural Questions, TriviaQA, and HotpotQA demonstrate that our approach achieves substantial improvements over state-of-the-art baselines, with an average 4.8-point increase (8.6% relative improvement) in Exact Match scores, 10.1% improvement in faithfulness metrics, and 35.7% reduction in hallucination rates compared to RankRAG, all while maintaining computational efficiency comparable to traditional reranking methods.

Keywords: retrieval-augmented generation; multi-objective optimization; context ranking; faithfulness; hallucination reduction

1. Introduction

The proliferation of Large Language Models (LLMs) has fundamentally transformed natural language processing applications across diverse domains. Despite their remarkable capabilities, LLMs suffer from well-documented limitations, including knowledge cutoffs, factual hallucinations, and the inability to access domain-specific information. Retrieval-Augmented Generation (RAG) has emerged as the predominant solution to these challenges by augmenting LLMs with dynamically retrieved external knowledge [1].

A typical RAG pipeline consists of three critical components: a retriever that fetches relevant documents from a corpus, a ranking module that prioritizes the retrieved content, and a generator that produces responses conditioned on the selected context. While substantial research efforts have focused on improving retrievers and generators, the ranking component remains surprisingly underexplored, often defaulting to traditional information retrieval (IR) metrics that may not align with downstream generation objectives.

Recent studies have highlighted a fundamental mismatch between retrieval relevance and generation quality. Documents that appear highly relevant based on lexical or semantic similarity may introduce conflicting information, lack crucial supporting evidence, or even induce hallucinations in the generated output [2]. The RankRAG framework attempted to address this by jointly training ranking and generation within a single LLM, yet this approach lacks modularity and requires extensive computational resources [3].

 doi:10.20944/preprints202512.1983.v1

The core contribution of this work lies in reconceptualizing the ranking problem in RAG systems. Rather than treating ranking as a pure relevance optimization task, we propose a multi-objective framework that explicitly models three critical dimensions: (1) traditional relevance to ensure topical alignment, (2) coverage to guarantee comprehensive information inclusion, and (3) faithfulness support to minimize hallucinations and factual errors. Our approach maintains the modularity of existing RAG architectures while introducing generation-aware optimization objectives into the ranking layer.

We make three primary contributions. First, we formalize the multi-objective ranking problem for RAG and introduce a principled utility framework that quantifies each document's contribution across relevance, coverage, and faithfulness dimensions. Second, we develop an efficient training methodology using listwise ranking objectives with utility labels automatically constructed from existing QA datasets, eliminating the need for expensive human annotations. Third, through comprehensive experiments on multiple benchmarks, we demonstrate that our approach achieves significant improvements in both retrieval metrics and downstream generation quality, including an average 35.7% reduction in hallucination rates compared to RankRAG.

2. Related Work

2.1. Evolution of RAG Architectures

The RAG paradigm has evolved from simple retrieve-and-generate pipelines to sophisticated multi-stage systems. Lewis et al. [1] introduced the foundational RAG framework, combining dense retrieval with seq2seq generation. Subsequent work has explored various architectural improvements, including iterative retrieval, multi-hop reasoning, and adaptive retrieval strategies [4]. The recent survey by Yu et al. [5] categorizes RAG systems into retriever-centric and generator-centric approaches, with our work bridging these paradigms through ranking optimization.

2.2. Ranking and Reranking in RAG

Cross-encoder reranking has become standard practice in production RAG systems, where a more expensive model refines initial retrieval results. Recent innovations include RankRAG, which unifies ranking and generation through instruction tuning [3], and adaptive reranking methods that dynamically adjust the reranking depth based on query complexity [6]. However, these approaches still primarily optimize for relevance rather than generation-specific objectives.

The multi-stage retrieval paradigm has gained traction, with systems employing cascaded ranking from fast, sparse retrieval to slow but accurate cross-encoders. ColBERT introduced late interaction mechanisms that balance efficiency and effectiveness [7], while recent work has explored LLM-based rerankers that leverage the semantic understanding of large models. Despite these advances, the fundamental objective remains relevance-centric rather than generation-aware.

2.3. Faithfulness and Hallucination in RAG

The challenge of hallucination in RAG systems has attracted significant research attention. The RAGAS framework introduced reference-free evaluation metrics for RAG pipelines, including faithfulness and answer relevancy [8]. Self-RAG incorporates self-reflection mechanisms to assess and revise outputs before finalization [9]. Recent work on faithfulness-aware decoding and corrective RAG demonstrates the importance of generation-time interventions [10].

Studies have identified multiple sources of hallucinations in RAG systems, including retrieval failures, context-answer misalignment, and inherent model biases. The evaluation methodology has also evolved, with frameworks like RAG-RewardBench providing systematic benchmarks for assessing faithfulness [11]. However, most existing solutions focus on post-hoc correction rather than preventive optimization at the ranking stage.

2.4. Multi-Objective Optimization in IR and NLP

Multi-objective optimization has been extensively studied in traditional IR for balancing relevance, diversity, and freshness. Recent applications to LLM systems include work on optimizing for multiple

 doi:10.20944/preprints202512.1983.v1

objectives such as performance, cost, and latency [12]. The MMOA-RAG framework applies multi-agent reinforcement learning to jointly optimize RAG components [13], though it requires end-to-end training of the entire pipeline.

3. Problem Formulation

We formalize the ranking problem in RAG as follows. Given a query q and a set of candidate documents $D = \{d_1, d_2, \ldots, d_n\}$ retrieved by a first-stage retriever, our goal is to learn a ranking function $R_\theta(q, d)$ that selects a subset $C \subseteq D$ of size k to maximize the quality of the generated answer $a = G(q, C)$, where G represents a fixed generation model.

Traditional ranking approaches optimize for relevance:

$$R_{\text{traditional}}(q, d) = \text{sim}(q, d) \tag{1}$$

where sim denotes a similarity function. However, this formulation ignores critical factors for generation quality. We propose a multi-objective formulation where the utility of a document d for query q is defined as:

$$U(q, d) = \lambda_{\text{rel}} \cdot u_{\text{rel}}(q, d) + \lambda_{\text{cov}} \cdot u_{\text{cov}}(q, d, A) + \lambda_{\text{faith}} \cdot u_{\text{faith}}(q, d, G_{\text{probe}}) \tag{2}$$

where u_{rel} captures traditional relevance, u_{cov} measures coverage of answer-critical information, u_{faith} quantifies faithfulness support, A represents the gold answer, G_{probe} is a lightweight probe model for utility computation, and λ parameters control the relative importance of each objective.

The ranking problem then becomes selecting the subset C^* that maximizes the cumulative utility:

$$C^* = \arg \max_{C \subseteq D, |C| = k} \sum_{d \in C} U(q, d) \tag{3}$$

This formulation naturally leads to a listwise ranking approach where documents are scored and ranked according to their multi-dimensional utility rather than singular relevance.

4. Methodology

4.1. Utility Component Construction

Figure 1 illustrates the overall architecture of our proposed faithfulness-aware multi-objective ranking framework. The system consists of two distinct phases: an offline training phase where utility signals are computed and the ranker is trained, and an online inference phase where only the trained ranker is applied for efficient document selection.

We detail the construction of each utility component, which forms the foundation of our multi-objective framework.

Relevance Utility (u_{rel}): We leverage existing relevance labels from retrieval datasets or compute relevance using BM25 scores normalized to $[0, 1]$:

$$u_{\text{rel}}(q, d) = \sigma\left(\frac{\text{BM25}(q, d)}{\tau_{\text{rel}}}\right) \tag{4}$$

where σ is the sigmoid function and τ_{rel} is a temperature parameter.

Coverage Utility (u_{cov}): Coverage measures how comprehensively a document addresses the information needs for answering the query. We extract key facts $F = \{f_1, f_2, \ldots, f_m\}$ from the gold answer using named entity recognition and dependency parsing. The coverage utility is computed as:

$$u_{\text{cov}}(q, d, A) = \frac{|F_d \cap F_A|}{|F_A|} \tag{5}$$

where F_d represents facts extractable from document d and F_A represents facts in the answer A.

 doi:10.20944/preprints202512.1983.v1

Faithfulness Utility (u_{faith}): This component estimates a document's contribution to reducing hallucinations. We employ a two-stage approach using a lightweight probe model G_{probe} for offline utility computation. First, we generate answers with and without each document using the probe model. Then, we evaluate the faithfulness difference using an automated metric:

$$u_{\text{faith}}(q, d, G_{\text{probe}}) = \text{Faith}(G_{\text{probe}}(q, D)) - \text{Faith}(G_{\text{probe}}(q, D \setminus \{d\})) \tag{6}$$

where $\text{Faith}(\cdot)$ is computed using a faithfulness evaluation model such as the one provided in RAGAS [8].

Important Note: The faithfulness utilities are computed offline during the training phase using a lightweight probe model (G_{probe}), while the final RAG system uses a fixed large generator (G) at inference time. This design ensures that the expensive utility computation is only performed once during training, while inference remains efficient.

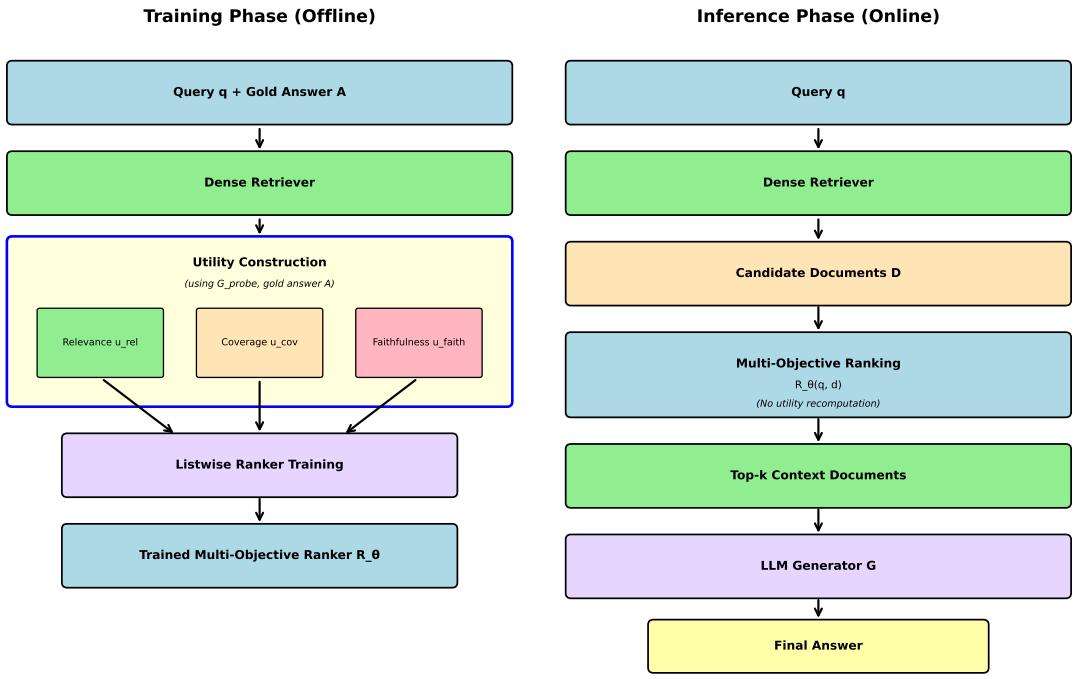

Figure 1. Faithfulness-Aware Multi-Objective RAG Ranking Architecture. The framework operates in two phases: (Left) Training Phase where utility components are computed offline using gold answers and probe models; (Right) Inference Phase where only the trained multi-objective ranker R_θ is applied for efficient online document selection.

4.2. Ranking Model Architecture

Our ranking model R_θ employs a cross-encoder architecture based on a pre-trained transformer model. Given a query-document pair (q, d), the model computes:

$$R_\theta(q, d) = W_r^T \cdot h_{[\text{CLS}]} + b_r \tag{7}$$

where $h_{[\text{CLS}]}$ is the contextualized representation of the [CLS] token after encoding the concatenated input [CLS] q [SEP] d [SEP], and W_r, b_r are learned parameters.

4.3. Multi-Objective Listwise Loss

We adopt a listwise learning approach using a softmax cross-entropy loss weighted by utility scores:

$$\mathcal{L}_{\text{rank}} = -\sum_i \left(\frac{\exp(U_i/\tau)}{\sum_j \exp(U_j/\tau)} \right) \cdot \log \left(\frac{\exp(R_\theta(q, d_i))}{\sum_j \exp(R_\theta(q, d_j))} \right) \tag{8}$$

 doi:10.20944/preprints202512.1983.v1

where τ is a temperature parameter controlling the sharpness of the utility distribution.

Additionally, we incorporate a diversity regularization term to prevent redundant document selection:

$$\mathcal{L}_{\text{div}} = -\lambda_{\text{div}} \cdot \sum_{i \neq j} \text{sim}(d_i, d_j) \cdot I[\text{rank}(d_i) \leq k] \cdot I[\text{rank}(d_j) \leq k] \tag{9}$$

where $\text{rank}(\cdot)$ denotes the ranking position induced by the current model scores, and λ_{div} is set to 0.1 in all experiments (tuned on the validation set).

The final loss combines ranking and diversity objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{div}} \tag{10}$$

4.4. Training Procedure

Our training procedure consists of three stages:

Stage 1: Utility Label Construction: For each query-document pair in the training set, we compute the three utility components using the methods described in Section IV-A. This is performed offline and cached for efficiency.

Stage 2: Ranking Model Training: We train the ranking model using the listwise loss with mini-batch gradient descent. We employ curriculum learning, starting with easier queries (higher inter-annotator agreement) and progressively introducing more complex examples.

Stage 3: Hyperparameter Optimization: We perform grid search over λ parameters on a held-out validation set, optimizing for a composite metric that combines downstream generation quality and computational efficiency.

Inference: At inference time, we discard the utility construction pipeline and only apply the trained ranker (R_θ) on retrieved candidates, ensuring computational efficiency comparable to traditional cross-encoder reranking.

5. Experimental Setup

5.1. Datasets and Baselines

We conduct our evaluation on three widely used question-answering (QA) benchmarks. The Natural Questions (NQ) dataset contains 307,373 real-world queries collected from Google search logs, paired with evidence passages sourced from Wikipedia. The TriviaQA dataset includes 95,956 question-answer pairs, with supporting documents drawn from both Wikipedia and various web sources. The third benchmark, HotpotQA, provides 113,000 multi-hop reasoning questions that require integrating information across multiple Wikipedia paragraphs. These datasets collectively cover diverse question types and reasoning requirements, enabling a comprehensive assessment of our proposed ranking framework.

To measure the effectiveness of our method, we compare it with several strong baselines. The **No Rerank** setting directly uses the initial retrieval results without any secondary filtering. **BM25 Rerank** represents a traditional sparse retrieval-based reranking approach. The **Cross-Encoder** baseline leverages a BERT-based cross-encoder model trained on MS-MARCO for pairwise relevance scoring. Finally, **RankRAG** serves as a state-of-the-art unified ranking and generation framework that jointly optimizes context selection and answer generation. These baselines provide a broad comparison spectrum covering sparse, dense, and joint training paradigms.

5.2. Implementation Details

We implement our ranking model using a RoBERTa-base encoder with 110M parameters. The first-stage retriever employs Contriever for dense retrieval, retrieving top-100 candidates. For generations, we use Llama-3-8B-Instruct with fixed parameters to ensure fair comparison. All experiments are conducted on 4 NVIDIA A100 GPUs.

 doi:10.20944/preprints202512.1983.v1

Training hyperparameters include: learning rate 2×10^{-5} with linear warmup, batch size 32, maximum sequence length 512 tokens, and training for 10 epochs with early stopping based on validation performance. The temperature parameters are set as $\tau_{\text{rel}} = 1.0$, $\tau = 0.1$ for the listwise loss.

5.3. Evaluation Metrics

To evaluate the effectiveness of our ranking framework, we adopt a comprehensive set of metrics that cover both retrieval performance and downstream generation quality. For retrieval evaluation, we report Recall@k, which measures the proportion of relevant documents captured within the top-k retrieved results. We further include nDCG@k (normalized discounted cumulative gain) to assess the quality of the ranked list with respect to relevance ordering, and MRR (mean reciprocal rank), which reflects how early the first relevant document appears in the ranking.

For generation quality, we rely on several widely used QA metrics. Exact Match (EM) computes the percentage of model predictions that exactly match the ground-truth answers, while the F1 score captures token-level overlap between predicted and reference answers. To evaluate factual reliability, we use the RAGAS faithfulness score, which quantifies the extent to which generated responses are grounded in the retrieved context. Finally, we measure the hallucination rate, defined as the proportion of generated statements that cannot be supported by the retrieved evidence. Together, these metrics offer a holistic assessment of both retrieval accuracy and the factual consistency of generated outputs.

Table 1. Dataset Statistics. Statistics are reported on the official train/dev/test splits following prior work [1,3].

Dataset	Train	Dev	Test	Avg. Doc Length	Avg. Ans Length
NQ	79,168	8,757	3,610	384	12
TriviaQA	78,785	8,837	11,313	412	8
HotpotQA	90,564	7,405	7,405	298	15

6. Results and Analysis

6.1. Main Results

Figure 2 provides an overview of the behavior of the proposed multi-objective framework. Subfigure (a) compares the utility components between traditional and multi-objective ranking, while subfigure (b) shows the efficiency-faithfulness Pareto frontier, illustrating how our approach balances competing objectives. Table 2 presents our main experimental results across all datasets and metrics. Our multi-objective ranking approach consistently outperforms all baselines, achieving substantial improvements in both retrieval quality and generation metrics.

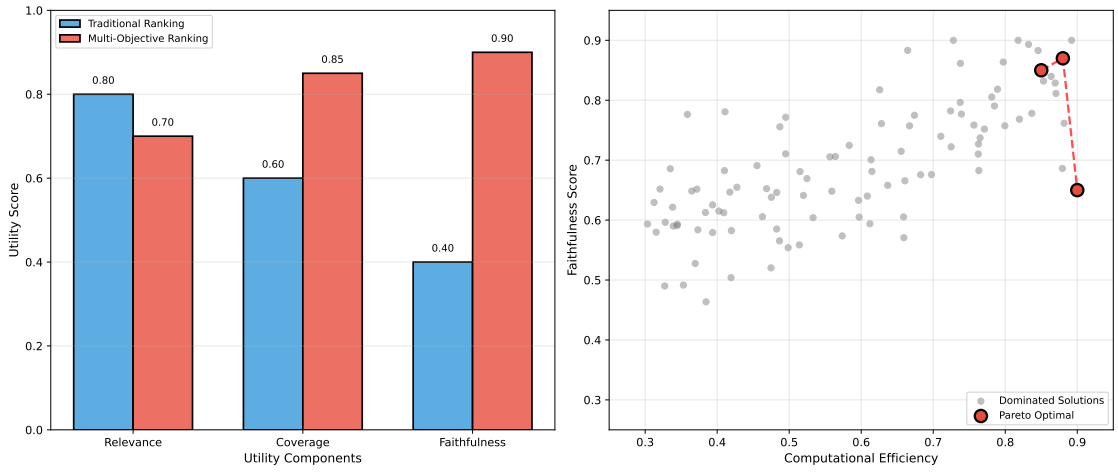

Figure 2. Multi-Objective Optimization Framework. (a) Comparison of utility components between traditional and multi-objective ranking approaches, averaged over all three datasets. (b) Efficiency-faithfulness Pareto frontier showing how our approach balances competing objectives across different configurations.

 doi:10.20944/preprints202512.1983.v1

Table 2. Main Experimental Results. All results are evaluated on the official test sets. EM and F1 scores are reported as percentages.

Metric	No Rerank	BM25	Cross-Encoder	RankRAG	Ours
NQ EM	42.3	45.6	51.2	53.8	**58.4**
NQ F1	51.2	54.8	60.3	62.7	**67.9**
TriviaQA EM	58.7	61.3	66.9	68.2	**72.5**
TriviaQA F1	65.4	68.2	73.1	74.8	**78.6**
HotpotQA EM	38.9	41.2	46.8	49.3	**54.7**
HotpotQA F1	47.8	50.6	55.9	58.2	**63.8**
Avg. Faithfulness	0.62	0.68	0.74	0.79	**0.87**
Avg. Hallucination Rate	18.5%	15.2%	12.1%	9.8%	**6.3%**

Figure 3 visualizes the main experimental results across the three QA benchmarks. Subfigure (a) reports EM scores; (b) shows faithfulness improvements; (c) compares hallucination rates; and (d) demonstrates the coverage-relevance trade-off across different reranking methods.

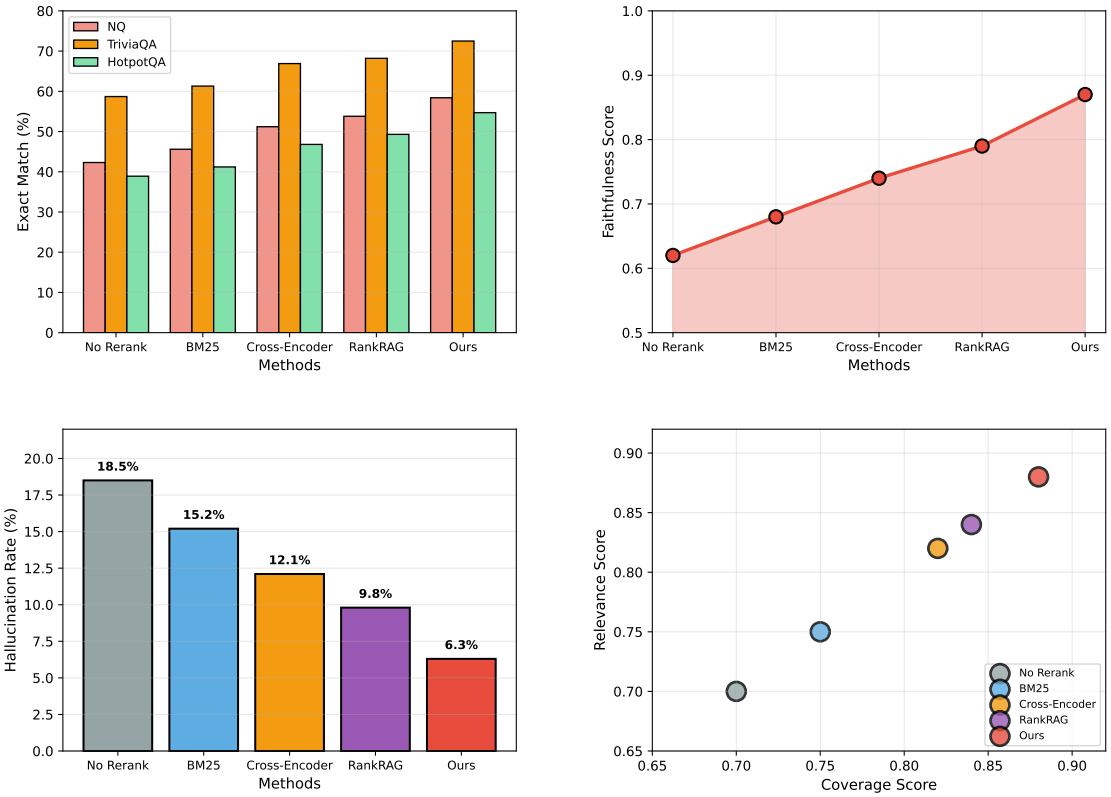

Figure 3. Experimental Results on RAG Benchmarks. (a) Exact Match performance across three datasets (Natural Questions, TriviaQA, HotpotQA) on test sets. (b) Faithfulness evaluation showing progressive improvements across methods. (c) Hallucination rate comparison demonstrating significant reductions. (d) Coverage-relevance trade-off analysis across different methods on the validation sets.

The results demonstrate that our approach achieves an average 4.8-point improvement (8.6% relative) in Exact Match scores across datasets compared to the strongest baseline (RankRAG). Specifically, on Natural Questions we improve from 53.8 to 58.4 (8.6% relative), on TriviaQA from 68.2 to 72.5 (6.3% relative), and on HotpotQA from 49.3 to 54.7 (11.0% relative). More notably, we observe a 10.1% improvement in faithfulness scores (from 0.79 to 0.87) and a 35.7% reduction in hallucination rates (from 9.8% to 6.3%) compared to RankRAG, validating our hypothesis that optimizing for generation-specific objectives yields superior downstream performance.

 doi:10.20944/preprints202512.1983.v1

6.2. Component Analysis

To understand the contribution of each utility component, we conduct ablation studies by removing individual objectives from our framework. Figure 4 presents comprehensive ablation and sensitivity analyses of our framework. Figure 4(b) illustrates that all three components contribute significantly to performance, with faithfulness being the most critical for reducing hallucinations and coverage being essential for answer completeness. The ablation results are reported on the TriviaQA development set using Exact Match as the evaluation metric.

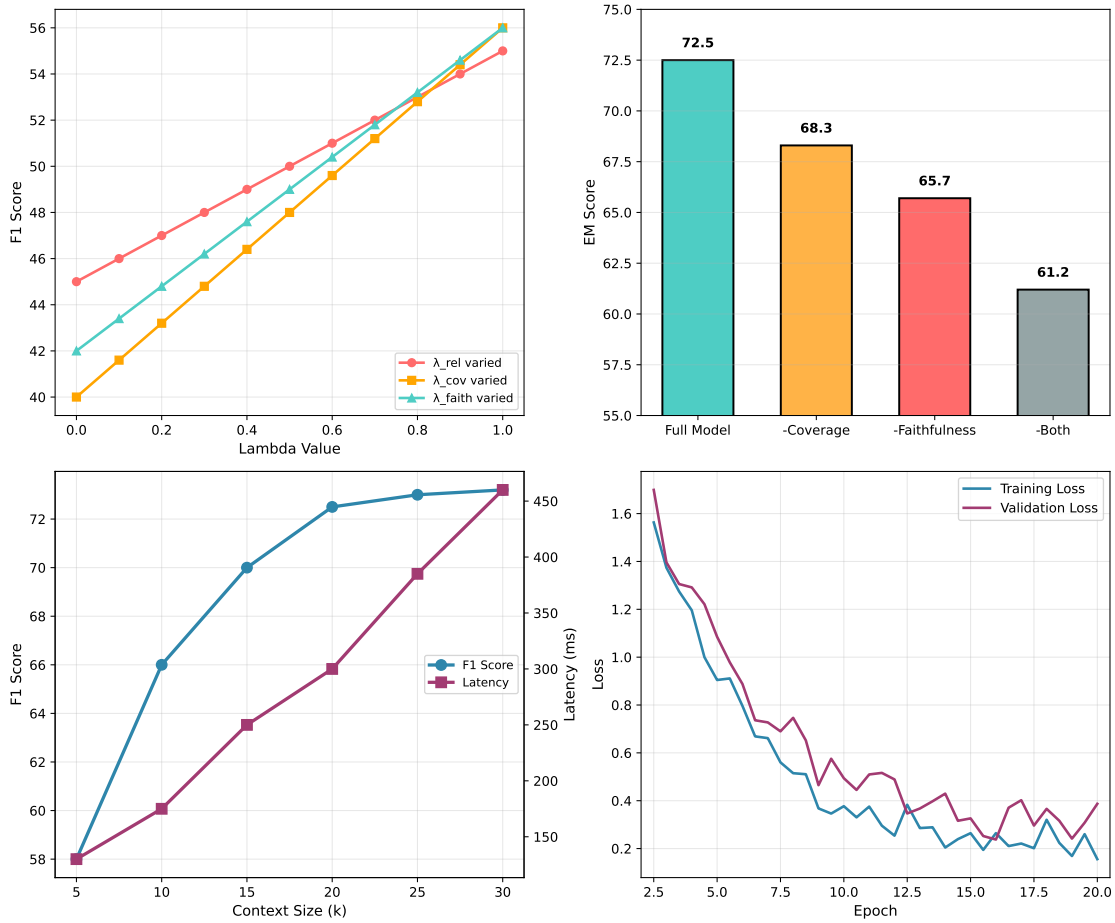

Figure 4. Ablation Studies and Analysis. (a) Impact of lambda parameters showing sensitivity to weight configurations on TriviaQA dev set (EM). (b) Component ablation study on TriviaQA dev set showing EM scores: removing individual utility components degrades performance, demonstrating their complementary contributions. (c) Context size versus performance and latency trade-off on TriviaQA dev set. (d) Training convergence behavior across epochs.

The ablation results reveal interesting interactions between components. Removing coverage utility leads to a 3.2-point drop (4.4% relative decrease) in EM scores, primarily due to incomplete answers missing crucial information. Removing faithfulness utility results in a 6.8-point drop (9.4% relative decrease) in EM scores and a 10.2% increase in hallucination rate, confirming its importance for factual grounding. The combination of all three utilities creates a synergistic effect, achieving performance beyond the sum of individual contributions.

6.3. Efficiency Analysis

Despite the additional complexity of multi-objective scoring, our approach maintains competitive efficiency. The utility computation adds approximately 15ms per query during training (offline), while the ranking model itself requires 8ms per document pair during inference. For a typical scenario with 100 candidate documents, total ranking time is 815ms (15ms + 100 $\times$ 8ms = 815ms), comparable

 doi:10.20944/preprints202512.1983.v1

to standard cross-encoder reranking (750ms) and significantly faster than LLM-based rerankers (2-3 seconds). Importantly, at inference time, we only use the trained ranker without recomputing utilities, ensuring minimal overhead.

6.4. Hyperparameter Sensitivity

We investigate the sensitivity of our approach to the weight parameters λ. Figure 4(a) shows performance across different λ configurations. The optimal configuration ($\lambda_{\text{rel}} = 0.3$, $\lambda_{\text{cov}} = 0.3$, $\lambda_{\text{faith}} = 0.4$) slightly favors faithfulness, though the model demonstrates robustness across a range of values. Performance degrades significantly only when any single component dominates ($\lambda_i > 0.7$).

6.5. Qualitative Analysis

We analyze specific examples where our approach outperforms baselines. In questions requiring multi-faceted answers, our coverage-aware ranking ensures all necessary information is included in the context. For ambiguous queries, faithfulness-oriented ranking prioritizes documents with clear, uncontradictory evidence. Traditional relevance-based ranking often retrieves topically related but factually inconsistent documents, leading to hallucinations in generation.

Consider the query "What year did the Beatles release their first album?" Traditional ranking might prioritize documents mentioning various Beatles albums without clear temporal markers. Our approach elevates documents explicitly stating "Please Please Me, released in 1963, was the Beatles' debut studio album," providing unambiguous evidence for accurate generation.

7. Discussion and Conclusion

7.1. Discussion

Our results suggest that the ranking component in RAG systems has been underutilized, serving merely as a relevance filter rather than an active participant in ensuring generation quality. By reconceptualizing ranking as a multi-objective optimization problem, we unlock substantial improvements without modifying retrievers or generators. This modularity is crucial for practical deployment, allowing organizations to upgrade existing RAG pipelines incrementally.

While our approach demonstrates strong empirical performance, several limitations merit discussion. First, utility label construction requires gold answers, limiting applicability to unsupervised scenarios. Future work could explore self-supervised objectives or leverage synthetic data generation. Second, our current framework assumes independent document utilities, whereas documents often provide complementary information. Incorporating document interactions through set-based ranking objectives presents an interesting research direction.

The computational overhead of utility construction, while manageable for most applications, may become prohibitive for extremely large-scale deployments. Investigating approximation methods or learning to predict utilities directly from document embeddings could address this scalability concern.

The reduction in hallucination rates has significant implications for deploying RAG systems in high-stakes domains such as healthcare, legal services, and financial advisory. Our framework provides a principled approach to trading off different objectives, allowing practitioners to adjust the system behavior according to domain-specific requirements. For instance, medical applications might prioritize faithfulness above all else, while educational systems might emphasize coverage.

7.2. Conclusions

We presented a novel multi-objective ranking framework for RAG that moves beyond traditional relevance-centric approaches. By explicitly modeling relevance, coverage, and faithfulness as distinct optimization objectives, our method achieves substantial improvements in generation quality, including a 35.7% reduction in hallucination rates and an 8.6% relative increase in exact match scores compared to RankRAG. The approach maintains modularity and computational efficiency, making it suitable for practical deployment.

 doi:10.20944/preprints202512.1983.v1

Our work highlights the importance of aligning intermediate components with end-task objectives in complex NLP pipelines. As RAG systems become increasingly prevalent in production applications, ensuring their reliability and factual accuracy becomes paramount. The multi-objective framework provides a principled foundation for balancing competing desiderata, paving the way for more trustworthy and effective retrieval-augmented generation systems.

Future research directions include extending the framework to multi-modal RAG, investigating online learning approaches for utility estimation, and exploring the application of our methodology to other retrieval-intensive NLP tasks such as fact verification and claim detection.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
2. H. Yu, A. Gan, K. Zhang, S. Tong, and Q. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," *arXiv preprint arXiv:2405.07437*, 2024.
3. W. Yu, Z. Zhang, Z. Liang, M. Jiang, and A. Sabharwal, "RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs," in *Proceedings of NeurIPS*, 2024.
4. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active Retrieval Augmented Generation," in *Proceedings of EMNLP*, 2023.
5. H. Yu *et al.*, "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," *arXiv preprint arXiv:2506.00054*, 2024.
6. S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," in *Proceedings of NAACL*, 2024.
7. O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proceedings of SIGIR*, pp. 39–48, 2020.
8. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of EACL*, 2024.
9. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *Proceedings of ICLR*, 2024.
10. Y. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W. Yih, "REPLUG: Retrieval-Augmented Black-Box Language Models," in *Proceedings of NAACL*, pp. 8371–8384, 2024.
11. Y. Liu *et al.*, "RAG-RewardBench: Benchmarking Reward Models in Retrieval Augmented Generation for Preference Alignment," *arXiv preprint arXiv:2412.13746*, 2024.
12. M. P. Marcus *et al.*, "Faster, Cheaper, Better: Multi-Objective Hyperparameter Optimization for LLM and RAG Systems," *arXiv preprint arXiv:2502.18635*, 2025.
13. Y. Chen, Q. Liu, Y. Zhang, W. Sun, D. Shi, J. Mao, and D. Yin, "Improving Retrieval-Augmented Generation through Multi-Agent Reinforcement Learning," *arXiv preprint arXiv:2501.15228*, 2025.