Enhancing Accuracy and Diversity in Retrieval-Augmented Generation through a Document-Structure-Aware Reranking Framework

Anonymous EMNLP submission

Abstract

Retrieval-Augmented Generation (RAG) systems often suffer from contextual redundancy and limited recognition of domain-specific entities in specialized domains, which degrades the quality and accuracy of responses generated by Large Language Models (LLMs). To address these challenges, we propose a document-structure-aware reranking framework that enhances both relevance and informational diversity, thereby improving the comprehensiveness and reliability of LLM outputs. Our approach consists of two key components: a multi-channel relevance scoring mechanism that combines thematic matching and entity-level signals, and a dynamic Maximal Marginal Relevance (MMR) algorithm based on thematic structure. This algorithm dynamically adjusts the trade-off parameter between relevance and diversity, effectively reducing semantic overlap among top-ranked passages. We conduct relevance evaluation on an internal benchmark dataset. Our method significantly outperforms existing baselines across multiple core metrics, with a 10.6% improvement in ranking accuracy over the internal baseline. Additionally, the framework further enhances the quality of model-generated responses by increasing the information density of the top-k document set.

1 Introduction

Large Language Models (LLMs) demonstrate impressive capabilities (Shao et al., 2024; Liang et al.), but their direct application often faces challenges such as outdated knowledge and factual hallucinations (Rawte et al., 2023). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) emerges as a leading approach to address these limitations by grounding LLM responses in up-to-date, external knowledge. This is particularly important in specialized domains—such as finance, law, and internal corporate knowledge bases—where high accuracy and access to lengthy, domain-specific documents are critical.

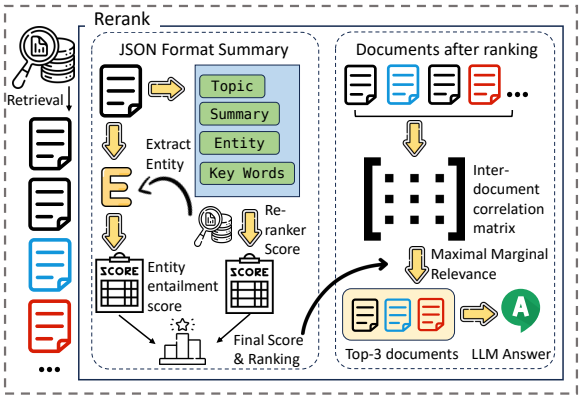

Figure 1: Our task primarily focuses on relevance ranking of long documents within a specific domain, considering user query needs from both accuracy and diversity perspectives.

The effectiveness of RAG systems largely depends on the quality and composition of context passages provided during retrieval (Fan et al., 2024; Park et al., 2025; Finardi et al., 2024) and reranking (Ampazis, 2024; Moreira et al., 2024). The retrieval stage is responsible for initially acquiring relevant documents, while the reranking stage refines the relevance of the retrieved documents. Standard RAG pipelines struggle with long, domain-specific documents due to redundancy, relevance decay, and the presence of specialized terminology or entities that generic models may fail to recognize, which impacts the comprehensiveness and accuracy of the output. Given the large size of our internal document library and the infrequency of changes in the retrieval process, our focus is on optimizing the reranking stage for better performance.

Current reranking methods in RAG systems remain limited in two key aspects: ineffective redundancy filtering in long documents and insufficient domain-specific entity recognition. While initial retrieval retrieves broad results, traditional rerankers primarily optimize for query-passage relevance, often returning top-k passages with high semantic

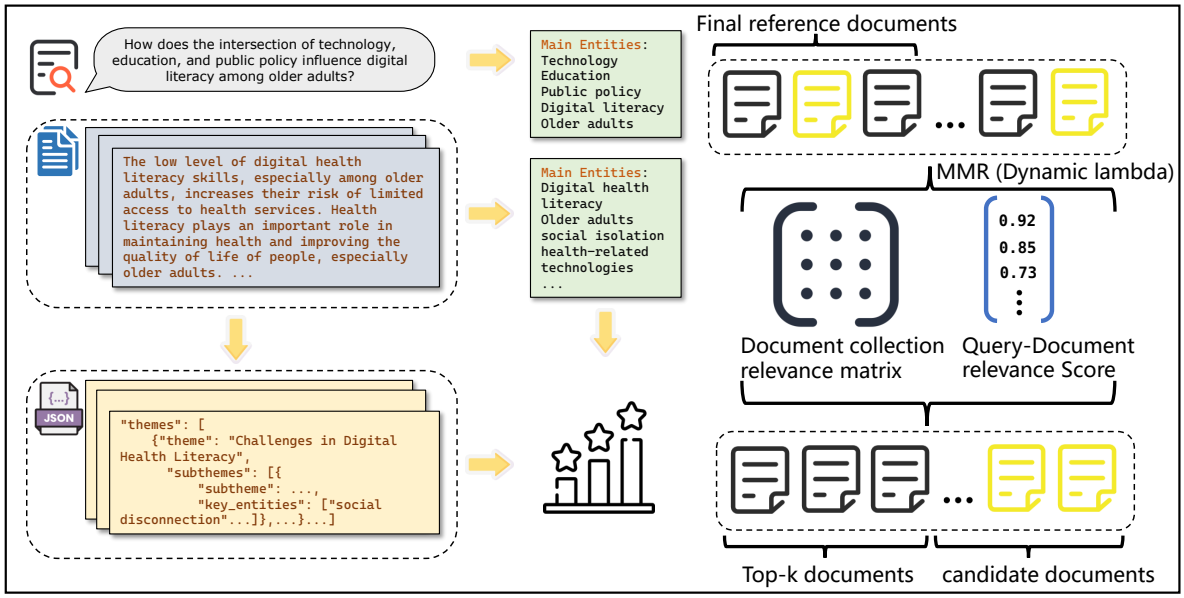

Figure 2: The workflow of our framework consists of two main components: accuracy and diversity. The accuracy part focuses on relevance scoring based on topic summarization and entity entailment. For diversity, we adopt MMR to dynamically adjust the top-k document set.

overlap. This redundancy restricts the LLM's capacity to integrate diverse perspectives, potentially biasing or fragmenting outputs. Additionally, many rerankers lack domain-aware entity sensitivity, increasing the risk of critical information omission.

To overcome these limitations, we propose a novel reranking framework designed to jointly enhance **accuracy** and **diversity**, especially for handling long documents in specialized domains. Our goal is not merely to diversify rankings but to curate a set of top-k passages that are highly relevant yet informationally complementary, thereby enriching the input for LLM generation.

We achieve this through a multi-channel relevance scoring mechanism that integrates thematic and entity-level signals, followed by a dynamic Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) based on semantic similarity among high-relevance candidates.

Our key contributions are as follows:

1. A multi-channel relevance calculation method combining thematic and entity-level signals, ensuring strong alignment with the core information need before introducing diversity considerations.

2. A theme-aware MMR algorithm with adaptive λ control that dynamically balances relevance and diversity based on the observed semantic overlap among top-ranked candidates.

3. Extensive experiments on specialized-domain datasets showing significant improvements over existing RAG ranking baselines in both retrieval accuracy and answer quality metrics.

2 Related Work

Retrieval-Augmented Generation enhances the accuracy and controllability of LLMs by integrating retrieval with generation techniques. A typical RAG pipeline consists of three stages: document indexing, initial retrieval (*e.g.*, using BM25 (Robertson et al., 2009) or DPR (Karpukhin et al., 2020)), and reranking. While these components perform well in general domains, they face notable challenges when applied to long documents in specialized fields—such as information redundancy, semantic repetition, and limited understanding of domain-specific terminology or entities.

Recent efforts focus on improving the reranking stage through advanced embedding models and fine-tuned rerankers, such as *bge-m3* and *bge-reranker-v2-m3* (Chen et al., 2024; Sturua et al., 2024; Zhang et al., 2024). These methods have shown improvements in query-passage relevance; however, they often fail to address semantic overlap among top-ranked passages and lack tailored mechanisms for modeling long documents in specialized domains.

To mitigate redundancy and promote diversity, several studies incorporate Maximal Marginal Rele-

vance (MMR), a strategy widely adopted in recommendation systems and document summarization. Traditional MMR implementations typically rely on bag-of-words or static vector representations, which are insufficient for capturing complex semantic structures in domain-specific content.

In summary, existing reranking methods struggle with semantic repetition, redundancy, and poor entity recognition in long, domain-specific texts. Our framework addresses these issues by combining thematic and entity-aware relevance scoring with dynamic λ adjustment, effectively balancing relevance and diversity for improved RAG performance in specialized domains.

3 Problem Setup

Given a user query q and a set of n retrieved long documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, the goal is to reorder these documents according to their relevance to the query. We aim to select and rank a subset $\mathcal{D}_k \subset \mathcal{D}$ of size k such that the selected documents are maximally relevant to q while maintaining diversity. Formally, this can be expressed as a combination of two objectives:

$$\mathcal{D}_k = \arg \max_{\substack{\mathcal{D}' \subset \mathcal{D}, \\ |\mathcal{D}'|=k}} \left(\lambda \cdot \mathcal{F}_{\text{acc}}(\mathcal{D}', q) + (1 - \lambda) \cdot \mathcal{F}_{\text{div}}(\mathcal{D}', q) \right), \tag{1}$$

where:

- $\mathcal{F}_{\text{acc}}$ denotes a scoring function focused on **accuracy**, defined by the semantic alignment between each document in $\mathcal{D}'$ and the query q;
- $\mathcal{F}_{\text{div}}$ represents a scoring function emphasizing **diversity**, encouraging broad coverage of the query topic while minimizing semantic overlap;
- $\lambda \in [0, 1]$ is a dynamic weighting parameter that balances the trade-off between accuracy and diversity.

When $\lambda = 1$, the optimization prioritizes **accuracy** alone, suitable for ranking-focused tasks:

$$\mathcal{D}_k = \arg \max_{\substack{\mathcal{D}' \subset \mathcal{D}, \\ |\mathcal{D}'|=k}} \mathcal{F}_{\text{acc}}(\mathcal{D}', q). \tag{2}$$

Conversely, when $\lambda = 0$, the objective shifts to maximizing **diversity**, ideal for LLM-based generation:

$$\mathcal{D}_k = \arg \max_{\substack{\mathcal{D}' \subset \mathcal{D}, \\ |\mathcal{D}'|=k}} \mathcal{F}_{\text{div}}(\mathcal{D}', q). \tag{3}$$

By adjusting λ, the system can dynamically adapt to downstream requirements: higher λ emphasizes precision for ranking, while lower λ enhances diversity for generative tasks.

4 Method

We propose a novel reranking framework for Retrieval-Augmented Generation (RAG) systems that jointly optimizes for both **accuracy** and **diversity** in the selection of top-k passages. The goal is to retrieve a set of documents that are not only highly relevant to the user query, but also collectively informative—capturing a broad range of complementary content to reduce redundancy and enrich the generative context.

Our framework consists of two main stages:

1. **Relevance Scoring via Multi-Channel Semantic Signals:** We compute fine-grained relevance scores for candidate passages by leveraging multiple semantic channels, including topic distributions and domain-specific entity recognition. These signals are derived from both general and specialized knowledge bases, allowing the system to more accurately capture the intent behind complex or technical queries. This step ensures that the initial document pool is both topically aligned and semantically precise.
2. **Topic-Structure-Aware Maximal Marginal Relevance (MMR):** To further refine the top-k selection, we employ an enhanced MMR algorithm that incorporates topic structure into the diversity-aware reranking process. Unlike standard MMR, our method dynamically adjusts the trade-off parameter λ based on pairwise semantic similarity among high-relevance candidates. This adaptive strategy promotes novel, non-redundant information while maintaining high overall relevance.

This two-stage approach is particularly effective for long and complex documents characterized by redundancy and dense terminology. By selecting passages that are both accurate and complementary, our framework improves the contextual input provided to Large Language Models (LLMs), thereby enhancing the *factuality*, *coverage*, and *informativeness* of the generated outputs.

4.1 Accuracy

In our reranking process, we prioritize *accuracy* to ensure that selected documents accurately reflect

the user's information needs while minimizing redundancy. To achieve this, we propose a hybrid scoring strategy that integrates hierarchical topic summarization and entity-level entailment, combining semantic understanding with explicit query grounding.

Formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ denote the set of retrieved long documents and q the user query. For each document $d_i \in \mathcal{D}$, we compute a final relevance score $R(d_i, q)$ as follows:

$$R(d_i, q) = \text{RRF}(s_{\text{topic}}(d_i, q), s_{\text{entity}}(d_i, q))$$

where $\text{RRF}(\cdot)$ denotes the Reciprocal Rank Fusion function, and the two components are defined below.

4.1.1 Topic-based Semantic Relevance

We leverage a LLM to perform hierarchical topic analysis on each document d_i, generating a structured summary $T_i = \{t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(m)}\}$ representing its main topics and subtopics. The topic-based semantic relevance score between the query q and the document is computed as:

$$s_{\text{topic}}(d_i, q) = \text{sim}(T_i, q)$$

Here, $\text{sim}(\cdot, \cdot)$ represents a semantic similarity function, typically calculated using cosine similarity between embeddings generated by a reranker or embedding model.

4.1.2 Entity-level Entailment

We extract the set of key entities $\mathcal{E}_q = \{e_1, e_2, \ldots, e_l\}$ from the query q using an LLM-based named entity recognition module. For each document d_i, we perform exact or fuzzy string matching to detect whether these entities appear in its full text, yielding a matched entity set $\mathcal{E}_{d_i} \subseteq \mathcal{E}_q$. The entity entailment score is computed as:

$$s_{\text{entity}}(d_i, q) = \frac{|\mathcal{E}_{d_i}|}{|\mathcal{E}_q|}$$

This ratio reflects the extent to which document d_i covers the core entities mentioned in the query, ensuring factual alignment and disambiguation in domain-specific contexts.

4.1.3 Final Relevance Fusion

To integrate the semantic and symbolic signals, we use Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to combine the independently ranked lists derived from s_{topic} and s_{entity}:

$$\text{RRF}(s_1, s_2) = \sum_{j=1}^{2} \frac{1}{k + \text{rank}_j}$$

where rank_j is the rank position of the document based on score s_j in the respective list, and k is a smoothing constant (set to 0). This fusion method ensures that documents performing well on either criterion are preserved in the top ranks while mitigating the dominance of a single feature type.

This relevance scoring strategy balances deep semantic representation with structured factual grounding, making it suitable for complex long-document reranking in retrieval-augmented systems.

4.2 Diversity

To enhance the diversity of reranked results, we introduce a dynamic adjustment mechanism based on the classical Maximal Marginal Relevance (MMR) algorithm. This approach aims to maximize the informational breadth of the selected document set while maintaining high relevance to the query. Unlike traditional MMR, which uses a fixed trade-off parameter λ, our method dynamically adjusts λ according to the characteristics of the initial ranking results.

Formally, let $\mathcal{D}_k = \{d_1, d_2, \ldots, d_k\}$ denote the top-k documents obtained from the initial ranking stage—such as through Reciprocal Rank Fusion (RRF). The MMR-based reranking score for each document d_i is defined as:

$$\text{MMR}(d_i, \mathcal{D}_k, q, \lambda) = \arg\max_{d_i \in \mathcal{D}_k} \big(\lambda \cdot \text{sim}(d_i, q) - (1 - \lambda) \cdot \max_{\substack{d_j \in \mathcal{D}_k \\ j \neq i}} \text{sim}(d_i, d_j)\big) \quad (4)$$

- $\text{sim}(d_i, q)$ denotes the similarity between document d_i and the query q, representing relevance.
- $\text{sim}(d_i, d_j)$ measures the similarity between documents d_i and d_j, indicating redundancy.
- $\lambda \in [0, 1]$ balances relevance and diversity in the selection.

To enable adaptive behavior, we analyze the initial top-k results by constructing a $k \times k$ relevance matrix S where each element $S_{ij} = \text{sim}(d_i, d_j)$ is computed using cosine similarity between embeddings generated from hierarchical topic summaries

extracted from each document. Based on this matrix, we calculate an average similarity score $\bar{S}$ across all pairs of documents within the top-k set:

$$\bar{S} = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i} S_{ij}$$

This metric reflects the overall redundancy among the top-k documents. A higher $\bar{S}$ indicates greater similarity and thus a stronger need for increased diversity. We then use this average similarity score to dynamically adjust λ:

$$\lambda_{\text{adjusted}} = f(\bar{S})$$

where $f(\cdot)$ is a mapping function that translates the average similarity into an appropriate λ value. In practice, we implement f using a monotonic transformation—such as a linear or sigmoid function—ensuring smooth transitions between relevance- and diversity-focused rankings.

By focusing on the average similarity across all document pairs rather than just the highest-ranked document, our method ensures a more balanced and comprehensive assessment of diversity needs. Specifically, the similarity $\text{sim}(d_i, d_j)$ is computed using embeddings derived from hierarchical topic summaries, ensuring that both semantic and structural aspects are considered.

This dynamic λ mechanism enables our reranking process to adapt to varying query characteristics and document distributions. As a result, it achieves a more balanced trade-off between accuracy and diversity, particularly beneficial when generating responses with large language models.

5 Experiment

In this section, we detail the experimental setup, models utilized, and datasets employed to evaluate the performance of our proposed reranking framework.

5.1 Experimental Setup

For the structural extraction task, we utilize the `qwen-turbo` (Bai et al., 2023). For reranking, the `bge-reranker-v2-m3` model is employed to reorder top candidate passages based on relevance and diversity considerations. The embedding model used in our experiments is `bge-m3`, which generates embeddings for queries and documents, facilitating efficient similarity computations. Lastly, for answer generation, we use the `qwen2.5-72b-instruct` (Qwen et al., 2025), a robust large language model capable of producing high-quality answers from retrieved documents.

5.2 Datasets

We employ two distinct datasets to evaluate both the accuracy and diversity of our approach.

5.2.1 Accuracy Dataset

To evaluate accuracy, we use an internal dataset sampled from Alibaba's online data. Queries are sourced from real user queries collected across various online platforms within the company. Documents in this dataset come from diverse sources, including official documentation, notification documents, and user-generated content such as solution posts, help requests, and experience-sharing articles. To form the initial candidate set, online retrieval is performed to recall the top-100 documents per query. Ground truth documents, representing the most relevant documents for each query, are selected based on user click behavior, ensuring a reliable signal for evaluating relevance.

5.2.2 Diversity Dataset

For diversity evaluation, we focus on complex analytical queries requiring information synthesis from multiple sources to generate detailed summaries. This dataset consists of 100 queries created using Qwen3 (Yang et al., 2025), typically open-ended and necessitating content integration from multiple sources. For each query, web retrieval retrieves the top-100 web pages, from which raw text is extracted as reference documents for reranking. This setup evaluates the system's ability to balance relevance with diversity by integrating information from various sources to produce comprehensive answers.

5.3 Evaluation Metrics

Our evaluation of the proposed reranking framework focuses on two key aspects: **accuracy** and **diversity**. Different strategies are applied depending on the downstream task objectives.

5.3.1 Accuracy Evaluation

To assess the accuracy of the reranking results, we apply standard ranking metrics:

- **Hit_Rate**: Measures the proportion of queries where at least one relevant document appears in the top-K results. Higher values indicate better promotion of relevant documents.

- **Mean Reciprocal Rank (MRR)**: Computes the average reciprocal rank of the first relevant document across all queries. MRR measures the model's effectiveness in ranking the most relevant document highly. Specifically, if the rank of the first relevant document is r, its reciprocal rank is $\frac{1}{r}$, and the final MRR averages these over all queries. MRR ranges from 0 to 1, with higher values indicating better performance (Voorhees et al., 1999).

- **Normalized Discounted Cumulative Gain (nDCG)**: Evaluates the quality of the ranked list by assigning higher weights to relevant documents appearing near the top. DCG accumulates relevance scores logarithmically based on position, and nDCG normalizes this by the ideal ranking. A higher nDCG indicates not only the presence but also the well-ranking of relevant documents (Kelly et al., 2009).

5.3.2 Diversity Evaluation

To evaluate the diversity of reranked results, we assess the quality of the final answers generated by an LLM when using reranked documents as input. We conduct pairwise comparisons of the output quality under different reranking strategies.

Specifically, for each query, answers are generated using the top-3 documents selected by different reranking methods. Pairwise comparisons between these answers are then conducted using a strong LLM judge, `Qwen3-235B-A22B` (Yang et al., 2025), a reasoning LLM. In each comparison, the LLM is tasked with determining which answer better responds to the query, considering factors like informativeness, coverage, clarity, and lack of redundancy.

By aggregating the pairwise wins, we derive a relative ranking of answer quality for each reranking strategy. This protocol directly assesses how document diversity influences the richness and usefulness of the final LLM-generated answer.

To evaluate the effectiveness of our reranking framework, we compare it against several widely adopted baseline methods, focusing on two core dimensions: **accuracy** and **diversity**. Below, we outline the baselines used for each dimension.

Accuracy Baselines For accuracy evaluation, we select representative ranking strategies including embedding-based retrieval models, cross-encoder rerankers, and a production hybrid ranking system:

- **Embedding-based Ranking**: We use the `bge-m3` model to encode both queries and documents into dense vector representations. Relevance scores are computed using cosine similarity between query and document embeddings.

- **Reranker-based Ranking**: We employ the `bge-reranker-v2-m3` model as a strong reranking baseline. This cross-encoder model directly evaluates the semantic interaction between query and document, offering improved performance on complex queries compared to embedding-based approaches.

- **Production Hybrid Ranking**: As an internal baseline, we consider a hybrid ranking strategy currently deployed in our production environment. It combines multiple scoring components, including dense and sparse retrieval models applied to both document titles and full content. The final score is obtained by equal-weighted summation:

$$\text{Score}_{\text{hybrid}} = \frac{1}{4}\big(s^{T}_{\text{dense}} + s^{T+B}_{\text{dense}} + s^{T}_{\text{sparse}} + s^{T+B}_{\text{sparse}}\big) \quad (5)$$

  where s^{T}_{dense} and s^{T+B}_{dense} denote the dense model scores using title and title+body respectively, and s^{T}_{sparse}, s^{T+B}_{sparse} are the corresponding sparse model scores.

Diversity Baselines To assess the contribution of diversity-aware reranking, we evaluate three configurations that vary in how they select the final Top-3 documents:

- **Top-3 from Initial Ranking**: We directly use the top-3 documents returned from the initial relevance ranking process (using Reciprocal Rank Fusion of topic and entity scores), without any further diversity adjustment.

- **MMR-based Reranking (Fixed λ)**: We apply the standard Maximal Marginal Relevance algorithm with a fixed λ parameter to rerank the Top-K candidates. This approach introduces a static trade-off between relevance and novelty in the final selection.

Dataset	Hit_Rate			MRR			nDCG		
	hit_rate@10	hit_rate@5	hit_rate@1	MRR@10	MRR@5	MRR@1	nDCG@10	nDCG@5	nDCG@1
Baseline									
Embedding	0.7280	0.6232	0.3929	0.4915	0.4775	0.3929	0.5477	0.5138	0.3929
Reranker	0.7876	0.6966	0.4726	0.5671	0.5549	0.4726	0.6197	0.5599	0.4726
Hybrid Ranking	0.7448	0.6513	0.4362	0.5279	0.5154	0.4362	0.5795	0.5493	0.4362
Our Method									
$Our_{embedding}$	0.8237	0.7326	0.4542	0.5760	0.5638	0.4542	0.6356	0.6062	0.4542
$Our_{reranker}$	**0.8489**	**0.7679**	**0.4982**	**0.6162**	**0.6052**	**0.4982**	**0.6724**	**0.6461**	**0.4982**

Table 1: Accuracy comparison of different ranking strategies based on the Top-100 documents, including embedding-based, reranker-based, and hybrid methods.

Comparison	Votes for A	Tie	Votes for B
Initial Ranking vs. MMR (Fixed λ)	65	56	79
Initial Ranking vs. MMR (Dynamic λ)	60	55	85

Table 2: LLM-as-Judge voting results for pairwise diversity comparisons. Each row shows the number of votes favoring method A, method B, and ties.

- **MMR with Dynamic λ**: This is our proposed method, where λ is dynamically adjusted based on the diversity demand of the initial Top-K set. The adjustment is guided by the average pairwise similarity among top documents and the quality of the highest-ranked item. When initial results are redundant, the algorithm increases diversity emphasis; when they are already diverse and relevant, it favors preserving top-ranked relevance.

We evaluate the LLM performance under each method by measuring the quality of answers generated based on the Top-3 ranked documents.

5.4 Results

The experimental results are shown in Table 1 and Table 2.

Experimental results show that, in terms of accuracy, our method outperforms the current online baseline across multiple retrieval metrics. Using the same embedding model, it achieves a 10.6% improvement in hit_rate@10 compared to Hybrid Ranking. In terms of diversity, applying the MMR algorithm generally yields better QA performance than directly using the top-ranked documents from the original ranking. Furthermore, performance improves even more when using a dynamic λ in MMR.

5.5 Ablation Study

Hybrid Ranking Components We extend the online baseline by incorporating topic-based score (s_{topic}) and entity-based score (s_{entity}) into the hybrid ranking formula. Based on this extension, we design three configurations to study the impact of these additional components: 1) adding only s_{topic}; 2) adding only s_{entity}; 3) adding both s_{topic} and s_{entity}.

In all cases, we apply uniform weighting across all used components. For the third setting (adding both), the final hybrid score is computed as:

$$\text{Score}_{\text{hybrid}} = \frac{1}{6}\Big(s^{T}_{\text{dense}} + s^{T+B}_{\text{dense}} + s^{T}_{\text{sparse}} + s^{T+B}_{\text{sparse}} + s_{\text{topic}} + s_{\text{entity}}\Big) \quad (6)$$

The experiment results are shown in Table 4. The experimental results show that directly incorporating our topic and entity scores as components into the online hybrid ranking algorithm can also yield significant improvements (a 6.4% increase in hit_rate@10).

Fusion Strategy To evaluate the impact of different fusion strategies on retrieval performance, we conduct a series of ablation experiments comparing two commonly used approaches: **Reciprocal Rank Fusion (RRF)** and **weighted sum fusion**. In particular, for RRF, we investigate the influence of its key hyperparameter k by testing multiple values—specifically $k = 0$, $k = 30$, and $k = 60$—to better understand how sensitive the method is to this parameter.

The experimental results are summarized in Table 5. From the data, it can be observed that RRF consistently outperforms weighted sum fusion across all tested settings, indicating its effectiveness in combining ranked lists from different sources. Among the different values of k, setting $k = 0$ yields the best overall performance, suggesting that under our experimental conditions, giving

Metrics	qwen-turbo-2025-04-28	qwen-turbo-2025-02-11	qwen2.5-1.5b-instruct
Overall Score Precision	94.96%	93.54%	87.65%
Overall Score Recall	91.04%	86.84%	77.78%
Topic Precision	95.02%	94.24%	88.43%
Topic Recall	94.44%	91.76%	84.08%
Keywords Precision	95.12%	93.08%	88.65%
Keywords Recall	90.26%	86.36%	78.57%
Summary Precision	95.04%	93.1%	86.61%
Summary Recall	91.26%	87.06%	77.92%
Entity Precision	96.02%	95.36%	88.69%
Entity Recall	89.8%	85.98%	74.1%
Attribute Precision	93.68%	91.88%	85.86%
Attribute Recall	89.3%	82.86%	74.29%

Table 3: Comparison of Accuracy and Recall Rates for Topic Structure Extraction by Different Models Against Manual Inspection (50 Cases)

Dataset	Hit_Rate			MRR			nDCG		
	hit_rate@10	hit_rate@5	hit_rate@1	MRR@10	MRR@5	MRR@1	nDCG@10	nDCG@5	nDCG@1
Hybrid Ranking	0.7448	0.6513	0.4362	0.5279	0.5154	0.4362	0.5795	0.5493	0.4362
Hybrid Ranking+topic	0.7507	0.6572	0.4404	0.5328	0.5202	0.4404	0.5847	0.5544	0.4404
Hybrid Ranking+entity	0.7900	0.7002	0.4818	0.5749	0.5628	0.4818	0.6261	0.5684	0.4818
Hybrid Ranking+topic+entity	**0.7928**	**0.7037**	**0.4841**	**0.5774**	**0.5655**	**0.4841**	**0.6288**	**0.5999**	**0.4841**

Table 4: Ablation results for hybrid ranking components. Incorporating topic and entity scores leads to noticeable improvements in ranking performance.

Dataset	Hit_Rate			MRR			nDCG		
	hit_rate@10	hit_rate@5	hit_rate@1	MRR@10	MRR@5	MRR@1	nDCG@10	nDCG@5	nDCG@1
$Our_{embedding}$ (RRF k=0)	**0.8237**	**0.7326**	**0.4542**	**0.5760**	**0.5638**	**0.4542**	**0.6356**	**0.6062**	**0.4542**
$Our_{embedding}$ (RRF k=30)	0.8051	0.6966	0.4482	0.5541	0.5396	0.4482	0.6138	0.5787	0.4482
$Our_{embedding}$ (RRF k=60)	0.7911	0.6881	0.4468	0.5497	0.5359	0.4468	0.6072	0.5738	0.4468
$Our_{embedding}$ (Weight 5:5)	0.7716	0.6762	0.4493	0.5458	0.5330	0.4493	0.5996	0.5687	0.4493
$Our_{embedding}$ (Weight 3:7)	0.7712	0.6759	0.4491	0.5456	0.5328	0.4491	0.5993	0.5684	0.4491
$Our_{embedding}$ (Weight 7:3)	0.7746	0.6794	0.4520	0.5487	0.5359	0.4520	0.6025	0.5716	0.4520

Table 5: Comparison of fusion strategies. RRF with $k = 0$ achieves the best performance, outperforming the weighted sum baseline.

higher weight to items appearing at the top of individual rankings significantly improves retrieval accuracy.

6 Conclusion

In this study, we address the challenges faced by Retrieval-Augmented Generation systems when handling long documents and specialized domain information, such as contextual redundancy and poor recognition of domain-specific entities. The proposed reranking framework aims to jointly enhance relevance and diversity through a multi-channel relevance scoring mechanism that incorporates thematic matching and entity-level signals. Additionally, a dynamic Maximal Marginal Relevance (MMR) algorithm based on thematic structure adjusts the trade-off between relevance and diversity.

Experimental results show that our approach outperforms existing baselines across multiple core metrics, particularly in accuracy and diversity. Ablation studies confirm the effectiveness of the method by evaluating different fusion strategies and model components. Overall, the reranking framework provides an effective solution for optimizing RAG systems, especially in handling complex and domain-specific content. Future work will focus on addressing current limitations and exploring broader applicability.

Limitations

Despite the significant improvements demonstrated by the proposed reranking framework across multiple core metrics, there are several limitations.

Firstly, the method relies on high-quality thematic analysis and entity recognition modules, which can be challenging to achieve in certain specialized domains, especially when sufficient training data is lacking. Secondly, while dynamically adjusting the λ parameter in the MMR algorithm effectively enhances diversity, its effectiveness heavily depends on the quality of the initial retrieval results; poor initial retrieval quality limits the improvement that reranking can achieve in the final output. Additionally, the current experiments primarily focus on internal benchmark datasets and specific types of query tasks, so the generalizability of this approach to broader tasks and datasets remains to be further validated.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

Datasets Source All data used in our publicly released evaluation set were collected exclusively from publicly accessible websites. We have carefully ensured that the dataset does not contain any personal or sensitive information, and there is no risk of privacy leakage.

AI assistants AI assistants (ChatGPT) were solely used to improve the grammatical structure of the text.

Acknowledgements

References

Nicholas Ampazis. 2024. Improving rag quality for large language models with topic-enhanced reranking. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 74–87. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024. The chronicles of rag: The retriever, the chunk and the generator. *arXiv preprint arXiv:2401.07883*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Gabriel de Souza P Moreira, Ronay Ak, Benedikt Schifferer, Mengyao Xu, Radek Osmulski, and Even Oldridge. 2024. Enhancing q&a text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag. *arXiv preprint arXiv:2409.07691*.

Heekyong Park, Martin Rees, Nils Kruger, Kenshiro Fuse, Victor M Castro, Vivian Gainer, Nich Wattanasin, Barbara Benoit, Kavishwar B Wagholikar, and Shawn Murphy. 2025. A comprehensive evaluation of llm phenotyping using retrieval-augmented generation (rag): Insights for rag optimization. *medRxiv*, pages 2025–04.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang,

Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.