RFG Framework: Retrieval-Feedback-Grounded Multi-Query Expansion

Ronaldinho Vega Centeno Olivera[1][a], Allan M. de Souza[1,2][b] and Julio Cesar dos Reis[1,2][c]
[1]*Institute of Computing, University of Campinas (UNICAMP), Campinas, Brazil*
[2]*Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC), Campinas, Brazil*

Keywords: Information Retrieval, Query Expansion, Retrieval-Augmented Generation, Pseudo-Relevance Feedback.

Abstract: Information Retrieval (IR) systems face challenges such as query ambiguity and lexical mismatch, which limit the effectiveness of dense retrieval models, whose generalization capability to new domains or tasks is often limited. This study proposes a novel query expansion framework, named RFG, which integrates the capabilities of Large Language Models (LLMs) into an architecture that combines Retrieval-Augmented Generation (RAG) with Pseudo-Relevance Feedback (PRF). Our solution is based on using an initial document retrieval as a grounding context for the LLMs, a process that mitigates the generation of unsubstantiated information ("hallucinations") by guiding the creation of a diverse set of pseudo-queries. Following an evaluation across a broad spectrum of retrieval models, including unsupervised and supervised dense models, our experimental results demonstrate that RFG consistently outperforms baseline methods, such as *HyDE* and *Query2doc*. In contrast to previous findings that suggest a negative correlation between retriever performance and query expansion benefits, this study originally reveals that our approach not only benefits models with lower initial effectiveness but also improves the results of more robust retrievers. This positions the generation of multiple, contextualized queries as a versatile and highly effective expansion strategy.

1 INTRODUCTION

Information Retrieval (IR) systems face the fundamental challenge of bridging the lexical and semantic gap between user queries, which are often short and ambiguous, and the documents within an extensive corpus (Zhu et al., 2024). This "vocabulary mismatch" limits the effectiveness of both traditional sparse retrieval models, like BM25 (Robertson and Walker, 1994), and modern dense retrievers based on embeddings (Wang et al., 2023). Although dense retrievers have proven effective when abundant labeled training data is available, their effectiveness often degrades in zero-shot scenarios or when facing domain shifts, where the distribution of new queries and documents differs from the training data (Zhu et al., 2024).

Query expansion has been a widely studied technique for decades. Classical methods, such as Pseudo-Relevance Feedback (PRF) (Rocchio, 1971), aim to refine the query by incorporating terms from the top-ranked documents in an initial retrieval (Mackie et al., 2023). The positive outcome of these approaches heavily depends on the quality of the initial results; if they are not relevant, PRF can introduce noise and divert the query from its original topic (Rashid et al., 2024).

The advent of Large Language Models (LLMs) has led to new expansion strategies that leverage their vast world knowledge and generative capabilities. Approaches like *Query2doc* (Wang et al., 2023) generate "pseudo-documents" to be concatenated with the original query. HyDE (Gao et al., 2023), on the other hand, generates "hypothetical documents" to find similar real documents in the embedding space. While innovative, these techniques risk generating plausible, but incorrect, or disconnected content ("hallucinations") (Zhang et al., 2023; Lewis et al., 2020), as they often operate without being directly grounded in the target corpus (Gao et al., 2023).

Recent research has questioned the universal utility of query expansion. A comprehensive analysis by Weller *et al.* (Weller et al., 2024) found a strong negative correlation between a retrieval model's performance and the benefits gained from expansion. These techniques tend to improve weaker models, but harm

[a] https://orcid.org/0009-0001-4756-9726
[b] https://orcid.org/0000-0002-5518-8392
[c] https://orcid.org/0000-0002-9545-2098

Vega Centeno Olivera, R., de Souza, A. M. and dos Reis, J. C.
RFG Framework: Retrieval-Feedback-Grounded Multi-Query Expansion.
DOI: 10.5220/0013836900004000
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 509-516
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

stronger, more robust ones (Weller et al., 2024). This observation suggests that the addition of information, while potentially beneficial for recall, may introduce noise that degrades the precision of more advanced models. Concurrently, it has been argued that generating a single rewritten query may be insufficient. In this sense, creating multiple diverse queries is crucial for improving the coverage of relevant documents (Rackauckas, 2024).

This study proposes a novel query expansion framework named RFG (Retrieval-Feedback-Grounded) that addresses the aforementioned limitations. Our solution uniquely integrates the principles of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and PRF (Rocchio, 1971). Unlike other methods, RFG originally uses the documents from an initial retrieval as a grounding context for an LLM, rather than for direct term extraction. This grounding process guides the model to generate a diverse set of pseudo-queries that are semantically relevant and faithful to the corpus content, thereby significantly mitigating the risk of hallucinations (Lewis et al., 2020) by constraining the generation to the vocabulary and concepts present in the retrieved documents. This set of generated queries is then used to expand the original search intent.

Our experimental evaluation, conducted across a broad spectrum of retrieval models, including unsupervised dense and state-of-the-art supervised dense retrievers, demonstrates that RFG consistently outperforms baseline methods like *HyDE* (Gao et al., 2023) and *Query2doc* (Wang et al., 2023). Our findings challenge the notion that query expansion is only useful for low-performing models. We demonstrate that RFG enhances the effectiveness of weaker retrievers and improves that of the most robust and advanced models, positioning our designed approach as a versatile and highly effective expansion method.

This investigation provides the following contributions:

- The design and full implementation of the RFG, a novel query expansion framework that uses feedback from an initial retrieval to ground the generation of multiple diverse queries via an LLM, combining simultaneously the strengths of RAG and PRF.
- A comprehensive empirical evaluation demonstrating the superiority of our method over baseline query expansion architectures across various retrieval models.
- A key contribution to the ongoing debate about the utility of query expansion, showing evidence that a well-grounded expansion approach can benefit both low-performing and robust, state-of-the-art retrieval models.

2 RELATED WORK

Query expansion is a fundamental technique in IR designed to address the persistent problem of lexical mismatch between user queries and the documents in a corpus (Weller et al., 2024). Historically, one of the most influential approaches has been PRF (Rocchio, 1971), which assumes that the top-ranked documents from an initial retrieval are relevant and uses their terms to expand the original query. The main limitation of these methods refers to their high dependency on the quality of the initial retrieval, which can introduce noise and degrade quality if the first results are not pertinent.

The advent of LLMs has enabled new strategies to emerge that leverage their vast text generation capabilities to overcome these limitations. Approaches like *HyDE* (Hypothetical Document Embeddings) (Gao et al., 2023) use an LLM to generate multiple "hypothetical documents" that answer the query. It then averages the embeddings of these documents to create a single vector used to find semantically similar real documents in the corpus.

Query2doc (Wang et al., 2023) approach generates a single "pseudo-document" that is concatenated with the original query. Unlike zero-shot approaches, *Query2doc* relies on few-shot prompting (specifically, with four examples), which implies it requires access to a training dataset with query-document pairs to function. A common thread and key limitation of these methods is that the generation of new content relies exclusively on the internal and parametric knowledge of the LLM, making them susceptible to generating "hallucinations", suggesting information that, although plausible, may be factually incorrect or irrelevant to the target corpus and query resolution.

Other investigations have explored different angles. Generative Relevance Feedback (GRF) (Mackie et al., 2023), for example, generates long-form texts independently of the initial retrieval to avoid dependence on its quality. Another notable strategy is RAG-Fusion (Rackauckas, 2024), which rewrites the original query into multiple versions to retrieve a broader set of documents. It then utilizes the Reciprocal Rank Fusion (RRF) algorithm (Cormack et al., 2009) to rerank the combined results. Despite the demonstrated success of these techniques, their benefit is not universal.

A comprehensive study by Weller *et al.* (Weller et al., 2024) revealed a strong negative correlation

between a retrieval model's base results and the gains obtained from expansion, indicating that existing techniques benefit weaker models, but harm more robust ones. In this sense, high-performing models are negatively affected because the generated text introduces noise that dilutes the original relevance signal.

Our present investigation is situated at the intersection of these research lines. Our proposed RFG framework explores LLMs for query expansion. However, unlike *HyDE*, *RAG-Fusion*, or *Query2doc*, our approach grounds the LLM's generation using pseudo-relevance feedback to mitigate hallucinations. Unlike classic PRF, our solution does not use the retrieved documents for term extraction, but rather as a context to guide the generation of multiple diverse queries. Finally, our investigation directly addresses the question posed by Weller *et al.* (Weller et al., 2024), by examining whether our grounded expansion strategy approach can benefit both weak models and more robust ones.

3 THE RFG FRAMEWORK

Figure 1 presents the design of our solution for generating multiple high-quality queries that are grounded in relevant information from the corpus, aiming to enhance retrieval diversity. The process is organized into three main stages: 1) Retrieval and Grounded Query Generation, 2) Document Retrieval and Aggregation, and 3) Reranking and Final Context Selection.

The RFG workflow begins with an initial retrieval based on the original user query to obtain a set of context documents. These retrieved documents serve to "ground" an LLM, which then generates a set of new and diverse queries. Each of these new queries is used to perform a second round of document retrieval. Finally, all retrieved documents are aggregated and reranked to select a high-quality subset, which serves as the final context for answer generation. We detail each stage of the RFG framework along with its mathematical formulation.

Stage 1: Retrieval and Grounded Query Generation. Given an original user query q and a corpus of documents $C = \{d_1, d_2, \ldots, d_m\}$, the first step is to perform an initial retrieval to obtain a context set. This process can be formalized as:

$$D_c = \text{Retrieve}(q, C, N) \tag{1}$$

where Retrieve is a standard retrieval function that returns the top N most relevant documents from C for the query q. This set of contextual documents, $D_c = \{d_{c_1}, d_{c_2}, \ldots, d_{c_N}\}$, is used to ground the LLM's query generation expansion.

The LLM generates a set of M diverse and grounded queries, $Q' = \{q'_1, q'_2, \ldots, q'_M\}$. Each new query q'_j is the result of a generation function that takes the original query q, the context document set D_c, and a specific *prompt* P_j as input, which can be varied to induce different rewriting styles:

$$q'_j = \text{LLM}(q, D_c, P_j) \quad \forall j \in \{1, \ldots, M\} \tag{2}$$

This grounding mechanism ensures that the new queries do not solely depend on the parametric knowledge of the LLM. They are informed by relevant content from the corpus.

Stage 2: Document Retrieval and Aggregation. Once the set of diverse queries Q' is generated, each new query q'_j is used to perform an independent retrieval operation over the same corpus C, obtaining the top N most relevant documents for each, as follows.

$$L_j = \text{Retrieve}(q'_j, C, N) \tag{3}$$

where L_j is the list of documents retrieved for query q'_j. Subsequently, all the retrieved document lists are aggregated to form a single set of candidate documents, D_{cand}. This process is defined as the union of all retrieved lists:

$$D_{\text{cand}} = \bigcup_{j=1}^{M} L_j \tag{4}$$

Stage 3: Rank Fusion and Final Context Selection. The set of retrieved lists, $L_1, \ldots, L_M$, contains relevant information from the perspectives of multiple queries. To effectively combine these lists and obtain a single, unified ranking, we apply the **Reciprocal Rank Fusion (RRF)** algorithm (Cormack et al., 2009). This method computes a fusion score s_d for each unique document d present in the candidate set D_{cand}:

$$s_d = \sum_{j=1}^{M} \frac{1}{k + \text{rank}_j(d)} \tag{5}$$

where $\text{rank}_j(d)$ is the rank of document d in the list L_j (if a document does not appear in a list, its contribution for that list is zero), and k is a constant used to mitigate the impact of high ranks. All unique documents from the candidate set are then sorted in descending order according to their RRF score s_d, producing a final fused list, L_{fused}. Finally, the top-K subset of documents from this list is selected to form the final context that is provided to the downstream LLM for answer generation, as follows:

$$D_{\text{final}} = \text{Top-K}(L_{\text{fused}}) \tag{6}$$

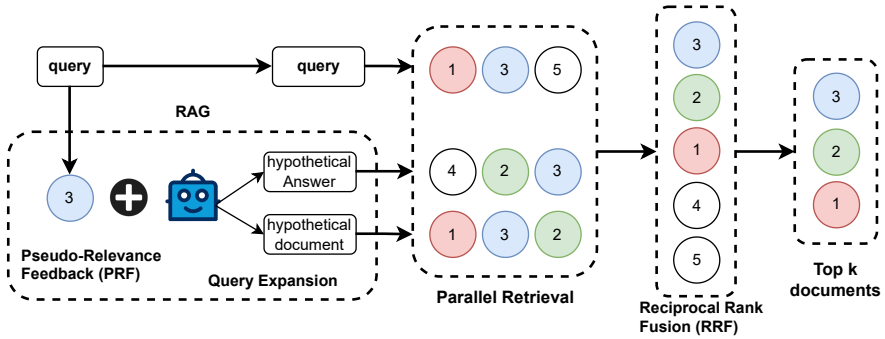

Figure 1: Complete flow of the RFG Framework.

This selected subset, D_{final}, represents the most relevant and diverse information, consolidated from multiple retrievals through a robust fusion process.

4 EVALUATION METHODOLOGY

This section outlines the experimental methodology used to evaluate the effectiveness of our RFG framework. We describe the evaluation benchmark and datasets, the retrieval models serving as baselines, the implementation details of our approach, and the evaluation metrics.

4.1 Evaluation Benchmark and Datasets

To ensure a robust and comprehensive evaluation across diverse retrieval challenges, we selected four datasets from the well-established **BEIR (Benchmarking IR)** benchmark (Thakur et al., 2021). This benchmark is specifically designed to assess the zero-shot generalization capabilities of retrieval models. The chosen datasets are:

- **ArguAna:** (Wachsmuth et al., 2018) This dataset is classified as a "Long Query Shift" task (Weller et al., 2024). It is chosen to test whether models, typically trained on short queries, can generalize to new length formats. The queries in *ArguAna* are notably long, with an average length of 197.1 words, effectively making them document-sized and posing a significant generalization challenge (Weller et al., 2024).
- **NFCorpus:** (Boteva et al., 2016) This is a "Domain Shift" dataset (Weller et al., 2024). It is used to evaluate a model's ability to generalize from its common training domain (e.g., general web documents from MS MARCO) to a new, specialized domain, such as medicine. This type of shift is made difficult by the specialized vocabulary inherent to the medical field (Weller et al., 2024).
- **FiQA-2018:** (Maia et al., 2018) Similar to NFCorpus, FiQA-2018 is used to evaluate "Domain Shift", testing model generalization from the training domain to the financial sector. This presents a challenge due to its specialized vocabulary (Weller et al., 2024).
- **SciDocs:** (Cohan et al., 2020) This is a "Domain Shift" dataset. Similar to NFCorpus and FiQA. It is used to evaluate the ability of models to generalize from a common training domain (web text) to a highly specialized domain, such as scientific literature. This shift is challenging due to the technical vocabulary and unique structure of scientific articles, testing the robustness of the retrieval model.

4.2 Baseline Models

To contextualize the results of our solution, we compared it against a broad spectrum of retrieval models, including:

Dense Retrievers. We evaluate both unsupervised and supervised dense models.

- **Unsupervised:** We include **Contriever** (Izacard et al., 2022), a dense retrieval model trained in an unsupervised manner.
- **Supervised:** We evaluate several high-performance models trained on benchmark corpora, such as **Contriever-ft** (fine-tuned on MS MARCO) (Izacard et al., 2022), **DPR** (fine-tuned on Natural Questions) (Karpukhin et al., 2020), and two of the most powerful recent embedding models, which consistently rank as top performers on the MTEB (Massive Text Embedding Benchmark) leaderboard (Muennighoff et al.,

2023)[1]: **BGE-large** (Xiao et al., 2024) and **GTE-large** (Li et al., 2023).

Furthermore, our query expansion approach was compared against two LLM-based baseline methods:

- **HyDE:** Evaluated on all the aforementioned embedding models and datasets.
- **Query2doc:** Evaluated on the embedding models for the NFCorpus and FiQA datasets, as these are the only ones that have the necessary training sets for its *few-shot prompting* approach.

4.3 Implementation Details

Query Generator. For the generation of pseudo-queries within our RFG framework, the **Mistral-small 3.1** model (Mistral AI, 2025) was used as the generative LLM. The first step is to retrieve the top $N = 5$ documents to serve as a grounding context. Next, the LLM generates a set of $M = 2$ diverse and grounded queries, $Q' = \{q'_1, q'_2\}$. Each query is created using a distinct strategy:

1. **Generated Document (q'_1):** A query that emulates the format and structure of a relevant document that could answer q. This strategy is similar to the "hypothetical document" concept in *HyDE* (Gao et al., 2023).
2. **Generated Answer (q'_2):** A direct answer to the original query q, containing keywords and information likely to be found in the correct documents.

Baseline Implementation. For *HyDE*, 4 pseudo-documents were generated for each original query, and their embeddings were averaged as suggested by the method. For *Query2doc*, 4 examples (*few-shots*) from the corresponding training sets were used to guide the generation of the pseudo-document.

Retrieval and Reranking. For managing and searching through the embedding vectors efficiently, we used **Qdrant** (Qdrant, 2024) as our vector store. Information retrieval for all dense models was performed using **cosine similarity** as the similarity measure. Our approach (RFG) performs three parallel retrieval operations, using the **Original Query**, the **Generated Document**, and the **Generated Answer** to each produce a ranked list of documents. These three resulting lists are then fused into a single, improved ranking using **Reciprocal Rank Fusion (RRF)** as the sole reranking strategy.

[1]Cf. the MTEB leaderboard at: https://huggingface.co/spaces/mteb/leaderboard

Infrastructure and Reproducibility. All experiments were conducted on a single **NVIDIA Quadro RTX 6000** GPU with 24 GB of VRAM. To ensure the reproducibility of our findings, the source code and configurations used in this research are made publicly available in a GitHub repository.[2]

4.4 Evaluation Metrics

The effectiveness of the different retrieval and query expansion methods was evaluated using standard metrics in the IR community. The primary metric for ranking is **nDCG@10 (Normalized Discounted Cumulative Gain at 10)**, which measures the quality of the ranking within the top 10 results.

4.5 Ablation Studies

To evaluate our framework, we performed two key analyses. First, an **ablation study** isolated the contribution of each query component (Generated Document and Generated Answer). Additionally, a **cost-benefit analysis** determined the practical viability by calculating the operational cost per 1,000 queries across the FiQA-2018 and NFCorpus datasets, using Mistral-small for generation and official token pricing[3].

5 EXPERIMENTAL RESULTS

We present the empirical results of our evaluation by analyzing the effectiveness of our proposed framework, RFG, in comparison with the baselines.

5.1 Overall Outcome Comparison

Table 1 presents the main results of our evaluation, comparing the effectiveness of the different query expansion methods on the four selected datasets. The reported metric is **nDCG@10**. The evaluated methods are: 1) **No Expansion**, which serves as our baseline; 2) **Query2doc**; 3) **HyDE**; and 4) **RFG**, our proposed method. Each method was evaluated on the full set of embedding models. To facilitate visual analysis, the best results for each model and dataset are marked in **bold** in Table 1. The background colors of the cells indicate the performance change concerning the "No

[2]The code for this research is available at: https://github.com/DinhoVCO/RFG

[3]Mistral AI pricing information, accessed July 24, 2025, available at: https://mistral.ai/pricing#api-pricing

Table 1: Consolidated results (nDCG@10, scaled by 100) comparing all expansion methods and ablations. The best result per column is in **bold** and the second-best is underlined. Cells with a green background indicate an improvement over the baseline (Original Query), while those with a red background indicate a performance degradation. Symbols: ♢ Original Query, □ Generated Document, ♣ Generated Answer.

Method	NFCorpus					FiQA-2018					SciDocs					ArguAna				
	DPR	Ctr	Ctr-ft	GTE-L	BGE-L	DPR	Ctr	Ctr-ft	GTE-L	BGE-L	DPR	Ctr	Ctr-ft	GTE-L	BGE-L	DPR	Ctr	Ctr-ft	GTE-L	BGE-L
Original Query ♢	14.8	26.0	30.7	37.9	36.1	5.9	11.9	27.0	44.5	44.3	**4.4**	11.6	15.3	22.7	22.5	**21.5**	46.8	**50.4**	**56.2**	**70.3**
State-of-the-art Methods																				
Query2Doc	12.3	21.1	24.9	28.6	30.8	5.4	20.5	28.0	43.8	43.6	-	-	-	-	-	-	-	-	-	-
HyDE	17.1	27.9	32.1	39.3	38.1	6.2	**24.5**	30.8	45.0	45.9	4.1	**14.3**	15.9	22.9	22.8	17.0	41.0	40.0	50.4	66.1
Our Method (RFG) and Ablations																				
RFG Gen Doc □	17.6	29.4	35.1	41.2	39.9	5.9	16.8	29.7	45.0	44.5	4.0	13.3	**16.4**	23.0	22.8	17.4	41.8	43.8	51.7	63.4
RFG Gen Answer ♣	16.8	29.6	34.2	41.1	40.1	**6.9**	15.0	30.8	45.6	44.9	4.1	13.3	15.9	22.7	23.1	19.6	39.7	45.3	46.5	60.6
RFG (□ + ♣)	17.8	30.1	**35.3**	**41.4**	**40.6**	6.6	16.5	31.3	45.9	45.8	4.2	13.7	16.3	23.1	22.9	19.5	43.7	46.8	52.5	64.9
RFG (♢ + □ + ♣)	**18.1**	**30.5**	34.5	40.9	39.8	**6.9**	15.9	**31.6**	**46.4**	**47.0**	**4.4**	13.6	**16.4**	**23.3**	**23.5**	21.1	**49.9**	49.1	55.0	68.6

Table 2: Cost and Performance Analysis for FiQA-2018 and NFCorpus. Costs are calculated per 1,000 prompts (at $0.10/1M input tokens and $0.30/1M output tokens). Performance is measured by nDCG@10. Symbols denote the query components used: ♢ for the Original Query, □ for the Generated Document, and ♣ for the Generated Answer.

Method	FiQA-2018				NFCorpus			
	Input Tokens	Output Tokens	Cost/1k ($)	nDCG@10	Input Tokens	Output Tokens	Cost/1k ($)	nDCG@10
Original Query ♢	0.00	0.00	0.0000	44.3	0.00	0.00	0.0000	36.1
HyDE	35.67	2112.80	0.6374	45.9	27.33	1713.82	0.5169	38.1
Query2Doc	1197.96	461.42	0.2582	43.6	1713.82	444.05	0.3046	30.8
RFG Gen Doc □	1294.59	381.33	0.2439	44.5	2166.86	357.70	0.3240	39.9
RFG Gen Ans ♣	1294.59	140.91	0.1717	44.9	2166.86	177.03	0.2698	40.1
RFG (□ + ♣)	2589.18	522.24	0.4156	45.8	4333.72	534.73	0.5938	40.6
RFG (♢ + □ + ♣)	2589.18	522.24	0.4156	47.0	4333.72	534.73	0.5938	39.8

Expansion" baseline: greenish cells represent an improvement, while reddish cells indicate a performance degradation.

An analysis of Table 1 reveals the consistent superiority of the RFG framework. Variants of our method achieve the top nDCG@10 score in 14 out of the 20 evaluated configurations. For instance, on the NFCorpus dataset with the DPR model, our full RFG approach elevates the baseline nDCG@10 from 14.8 to 18.1, a relative increase of over 22%. Similarly, on FiQA-2018 with the high-performing BGE-L model, RFG reaches a score of 47.0, surpassing both the baseline (44.3) and HyDE (45.9).

This stability contrasts sharply with existing methods, which can be detrimental. Query2doc, for example, causes a performance drop on NFCorpus with GTE-L (from 37.9 to 28.6). In contrast, our RFG framework yields performance gains (greenish cells) in nearly all scenarios, demonstrating its robustness and effectiveness across different datasets and retrieval models.

5.2 Cost-Benefit Analysis

This analysis is crucial for understanding the practical and financial viability of each approach when deployed on a large scale. Table 2 details this cost analysis alongside the average token counts per prompt.

The analysis reveals a clear trade-off between cost and performance. On the **NFCorpus** dataset, the combination of our components **RFG (Gen Doc + Gen Ans)** achieves the highest nDCG@10 (40.6), albeit at the highest cost. Notably, **RFG Gen Ans** offers nearly identical performance (40.1) for less than half the cost, presenting itself as a highly efficient alternative.

For the **FiQA-2018** dataset, the highest performance is achieved by the full **RFG (Original Query + Gen Doc + Gen Ans)** strategy, which reaches an nDCG@10 score of 47.0. *HyDE* follows with a score of 45.9 but at a significantly higher cost ($0.6374). From a cost-benefit perspective, **RFG Gen Ans** provides the most favorable balance, delivering a solid outcome improvement to 44.9 (from a baseline of 44.3) for the lowest cost among the effective expansion strategies ($0.1717).

6 DISCUSSION

Our experimental results confirm that the RFG framework consistently enhances retrieval effectiveness. Its success is rooted in two core principles: **grounded generation** and **multi-perspective fusion**. Unlike ungrounded methods such as *HyDE* or *Query2doc*, which rely solely on an LLM's parametric knowledge, RFG grounds the query generation process on documents from an initial retrieval. This forces the

LLM to produce queries that are faithful to the corpus content, mitigating hallucinations and improving relevance. The synergy of our approach is evident when we fuse the original query with the generated document and answer using RRF (♢ + □ + ♣), which consistently yields the best results by combining multiple unique perspectives.

The robustness of this multi-perspective fusion is particularly clear in challenging scenarios like the ArguAna dataset. On this benchmark, where long, detailed queries make most expansion methods harm performance, our full framework significantly mitigated this degradation. By retaining the original query in the fusion set, RFG provides the necessary signal diversity for RRF to consolidate rankings effectively, outperforming other methods even under unfavorable conditions.

These findings have significant implications, directly challenging the conclusion from Weller *et al.* that query expansion offers diminishing returns for strong retrieval models. Our results demonstrate consistent improvements even for powerful models like GTE-large and BGE-large. We argue that the performance degradation observed in prior studies is not a flaw of expansion itself, but a symptom of using noisy, *ungrounded* generative methods. Our work shows that controlled, context-aware expansion remains a valuable technique, shifting the focus from *whether* to use expansion to *how* to implement it effectively.

Our study has two primary limitations. First, the evaluation was confined to a specific set of datasets from the BEIR benchmark, which limits its proven generalizability. Second, we did not analyze the computational overhead and latency of our multi-query process, a key factor for real-world production systems.

Future work will address these limitations and explore new directions. A key step will be to evaluate RFG on diverse domains beyond BEIR, such as conversational, legal, and other specialized fields (*e.g.*, Medicine, Oil & Gas, or Finance). We also plan to develop an adaptive version of RFG that dynamically selects pseudo-queries based on query complexity to reduce computational costs. Finally, we will investigate more advanced fusion algorithms beyond RRF.

7 CONCLUSION

We observe an unresolved challenge in IR query expansion regarding the risk of hallucinations and the questioned effectiveness of these models on high-performing retrieval systems. This study introduced **RFG**, a novel framework that effectively integrates pseudo-relevance feedback (PRF) to ground the LLM generation, thereby producing diverse and faithful multiple queries regarding the corpus content. Our comprehensive empirical evaluation, which relied on several datasets from the BEIR benchmark and spanned a broad spectrum of retrieval models, demonstrated that RFG consistently outperformed baseline methods. We found that our grounded expansion strategy benefits both low-performing and more robust, supervised retrievers. We demonstrated the synergistic effect of combining multiple retrievals using Reciprocal Rank Fusion (RRF), which confirms that the diversity of the generated queries is crucial for more comprehensive and accurate IR mechanisms. The RFG approach, based on grounding and diversity, represents a promising research path, opening new avenues for the creation of more accurate and reliable RAG systems. We contributed a robust and effective query expansion framework that reframes the debate on the utility of expansion in the era of LLMs.

ACKNOWLEDGEMENTS

This study was sponsored by Petróleo Brasileiro S.A. (PETROBRAS) within the project "*Application of Large Language Models (LLMs) for online monitoring of industrial processes*" conducted in partnership with the University of Campinas [01-P-34480/2024 - 62208].

REFERENCES

Boteva, V., Gholipour, D., Sokolov, A., and Riezler, S. (2016). A full-text learning to rank dataset for medical information retrieval. In Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., and Silvello, G., editors, *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (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*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Gao, L., Ma, X., Lin, J., and Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning.

Mackie, I., Chatterjee, S., and Dalton, J. (2023). Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2026–2031, New York, NY, USA. Association for Computing Machinery.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., and Balahur, A. (2018). Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Mistral AI (2025). Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1. Accessed: June 25, 2025.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: Massive text embedding benchmark. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Qdrant (2024). Qdrant documentation: Overview. https://qdrant.tech/documentation/overview/. Accessed: June 25, 2025.

Rackauckas, Z. (2024). Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Rashid, M. S., Meem, J. A., Dong, Y., and Hristidis, V. (2024). Progressive query expansion for retrieval over cost-constrained data sources.

Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 232–241, Berlin, Heidelberg. Springer-Verlag.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Wachsmuth, H., Syed, S., and Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Wang, L., Yang, N., and Wei, F. (2023). Query2doc: Query expansion with large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Weller, O., Lo, K., Wadden, D., Lawrie, D., Van Durme, B., Cohan, A., and Soldaini, L. (2024). When do generative query and document expansions fail? a comprehensive study across methods, retrievers, and datasets. In Graham, Y. and Purver, M., editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1987–2003, St. Julian's, Malta. Association for Computational Linguistics.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. (2024). C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J.-R. (2024). Large language models for information retrieval: A survey.