Investigating Polarity Manipulation in LLM-Based Query Expansion for Financial Opinion Search

Beatrice Camera - Daria Miele - Zofia Pempera

Academic Year: 2025/2026
Bachelor in Artificial Intelligence

Abstract

Financial information retrieval often involves opinion-oriented queries that require nuanced interpretations of risk and market behavior rather than simple fact-finding. While Large Language Models (LLMs) offer promising capabilities for query enrichment, the effects of sentiment-induced expansion in financial IR remain underexplored.

In this study, we investigate the impact of sentiment-controlled LLM-based query expansion within a hybrid retrieval framework combining BM25 lexical retrieval and a Bi-Encoder neural re-ranker. Using the FinQA dataset, we generate neutral, positive (bullish), and negative (bearish) query expansions through prompt engineering and verify induced sentiment using a finance-specific classifier (FinBERT).

Experimental results show that all expansion strategies degrade retrieval effectiveness compared to original queries, primarily due to query drift in a sparse-relevance setting. Notably, negative expansions are more resilient than positive ones, suggesting that risk-oriented vocabulary aligns better with financial opinion content.

These findings highlight the sensitivity of hybrid retrieval systems to linguistic steering and emphasize the need for precision-preserving integration of LLMs in financial information retrieval.

1. Introduction

Financial information retrieval poses challenges that extend beyond traditional fact-based search. Financial queries often involve **technical concepts and implicit judgments** about risk or market behavior, and users typically seek explanations or interpretations rather than isolated facts. As a result, keyword-based retrieval models may fail to capture the true intent of domain-specific financial questions.

This problem is especially relevant for users such as retail investors, financial analysts, and students, for whom retrieval quality directly impacts understanding and decision-making. While neural re-ranking models and Large Language Models (LLMs) can enrich short queries through semantic expansion, LLM-based query expansion may implicitly introduce sentiment or bias, potentially altering retrieval behavior in opinion-oriented financial domains.

In this project, we study the impact of **sentiment controlled LLM query expansion** on a hybrid financial retrieval system. Using a **BM25 + Bi-Encoder pipeline**, we compare original queries with neutral, positive (bullish), and negative (bearish) expansions to assess whether induced polarity systematically affects ranking effectiveness and retrieval metrics.

2. Task and Dataset Description

This project addresses an **opinion-oriented financial information retrieval task**. Given a short natural language query, the system must retrieve documents that provide explanations, evaluations, or perspectives on financial topics, rather than simple factual answers.

Experiments are conducted on the FinQA retrieval dataset, which contains 57,638 financial documents and 648 queries with binary relevance judgments. The corpus is heterogeneous, including short microblog posts, news articles, and longer analytical reports, leading to substantial variation in document length and writing style.

The dataset presents several retrieval challenges. **Queries are very short** (≈ 5 tokens on average) and often underspecified, causing severe **vocabulary mismatch** with relevant documents. The corpus is dominated by noisy microblog content, which increases lexical variability, while domain-specific financial terminology requires semantic understanding beyond keyword matching. Moreover, **relevance judgments are sparse** (2.6 relevant documents per query on average), making ranking errors particularly costly. Although sentiment is rarely explicit, many queries implicitly express **opinion-driven information needs**, motivating the investigation of sentiment-aware query expansion strategies.

FinQA Dataset Statistics	
Number of documents	57,638
Number of queries	648
Average query length (tokens)	5.14
Maximum query length (tokens)	13
Average document length (tokens)	55.29
Microblog documents	36,612 (≈64%)
News documents	19,710 (≈34%)
Report documents	1,316 (≈2%)
Total relevance judgements	1,706
Average relevant judgements per query	2.63
Relevance labels	Binary (relevant / non-relevant)

3. Methodology

Our methodology follows a controlled, **two-phase structure**. Phase I establishes lexical and hybrid baselines, while Phase II investigates the strategic impact of sentiment-controlled LLM-based query expansion within a fixed hybrid retrieval architecture.

3.1 Baseline Systems (Phase I)

We first implemented a set of classical and hybrid baseline retrieval systems to establish reference performance on the FinQA dataset.

The lexical baselines consist of **TF-IDF** and **BM25**, both implemented using PyTerrier and applied to an index built over the full document collection. These models rely exclusively on exact term matching and serve to quantify the limitations caused by vocabulary mismatch, short queries, and heterogeneous document sources.

In addition, we implemented **BM25 with RM3** pseudo-relevance feedback to evaluate whether classical query expansion mitigates lexical mismatch in this domain. RM3 expands the original query using terms extracted from the top-ranked documents retrieved by BM25.

To move beyond purely lexical matching, we implemented a **hybrid baseline** combining **BM25** candidate retrieval with **neural semantic reranking**. BM25 is used to retrieve a fixed top-k candidate set for each query, after which a bi-encoder based on a Sentence-BERT architecture reranks these candidates according to cosine similarity between query and document embeddings. This hybrid system serves as the primary reference point for all subsequent experiments involving query expansion.

All baseline systems are evaluated using the same experimental setup and metrics to ensure fair comparison.

3.2 Advanced System (Phase II)

The second phase focuses on evaluating the impact of Large Language Model based query expansion under controlled sentiment conditions.

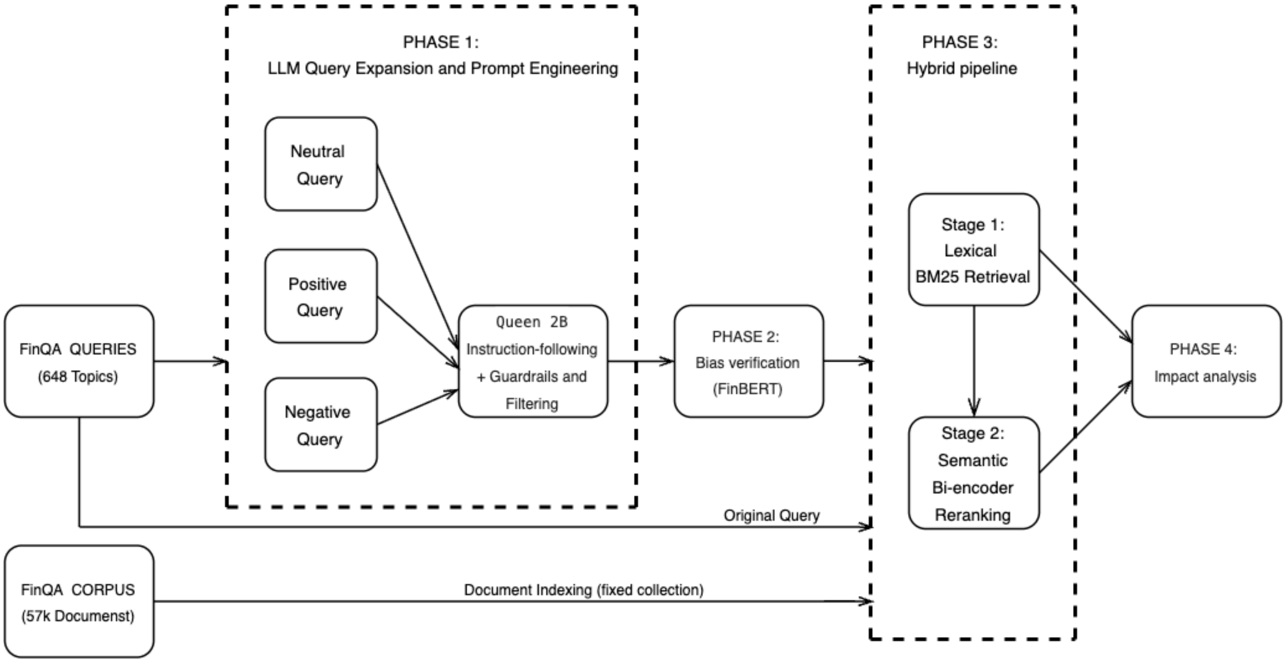
For each original query, we generate three expanded variants using an **instruction following LLM** through prompt engineering:

- **Neutral query expansion**, which enriches the query with domain-relevant financial terms without introducing sentiment.
- **Positive query expansion**, which injects optimistic or bullish financial language.
- **Negative query expansion**, which injects pessimistic or risk-oriented financial language.

Each expansion produces a fixed number of keyword phrases, which are appended to the original query to ensure consistent query length across conditions. All expansions are generated offline, cached to disk, and reused across experiments to guarantee reproducibility.

To verify that sentiment manipulation is effective, we perform a **bias verification** step using a financial-domain sentiment classifier (**FinBERT**). Each generated expansion term is scored, and aggregate polarity statistics are computed to confirm that positive expansions exhibit higher polarity than neutral ones, and neutral ones higher than negative. This step ensures that any observed changes in retrieval performance can be attributed to controlled sentiment differences rather than random variation.

All four query variants (original, neutral, positive, negative) are then passed through the same **hybrid retrieval pipeline**. BM25 retrieves the initial candidate set, which is reranked using the bi-encoder. Retrieval effectiveness is measured independently for each variant, allowing a direct comparison of how sentiment-induced query expansion alters ranking behavior while holding the retrieval architecture constant.



3.3 Implementation Details

All experiments are implemented in Python using Google Colab as the execution environment. PyTerrier is used for indexing, retrieval, and evaluation. Hugging Face Transformers and Sentence-

Transformers libraries are used for LLM inference, sentiment classification, and bi-encoder embedding generation.

LLM-generated query expansions are produced offline, cached in a JSON file, and version-controlled via a GitHub repository maintained by one group member to ensure reproducibility across runs.

The document collection is indexed without manual token-level preprocessing to avoid introducing retrieval bias. Query cleaning is applied only to prevent syntax errors in the retrieval engine. Neural models are used strictly as rerankers on BM25 candidate sets to maintain computational feasibility and experimental control.

4. Experiments and Results

This section reports the empirical evaluation of the proposed hybrid retrieval system and analyzes the impact of sentiment-controlled LLM-based query expansion on ranking effectiveness.

4.1 Evaluation Setup

Retrieval effectiveness is measured using standard Information Retrieval metrics: **P@1, P@5, P@10, R@5, R@10, nDCG@5, nDCG@10,** and **MAP**. These metrics jointly assess both ranking quality at top positions and overall retrieval performance.

All experiments are conducted using the **PyTerrier** framework with the official FinQA binary relevance judgments. To ensure fair comparison and reproducibility, the document index, retrieval pipeline, and neural re-ranking architecture are held **constant across all runs**. The only varying factor is the query formulation (Original, Neutral QE, Positive QE, Negative QE).

4.2 Results and Discussion

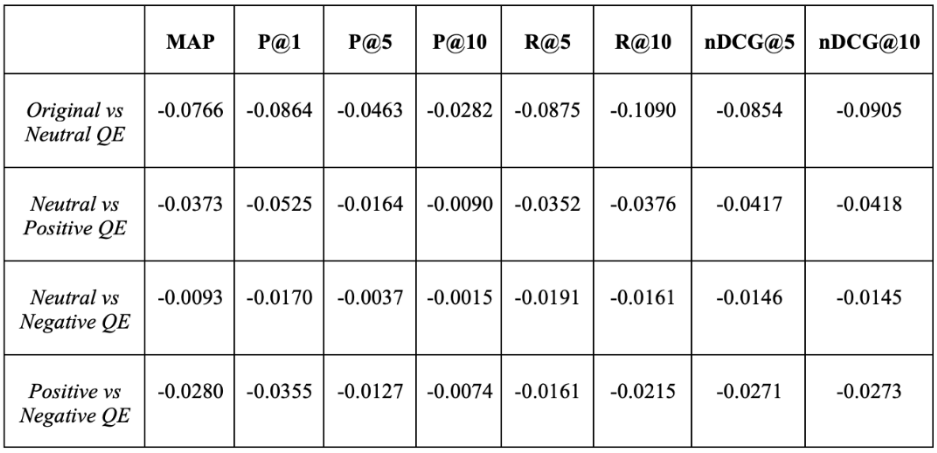
Table 1 reports retrieval effectiveness for the hybrid system under all query variants.

	MAP	**P@1**	**P@5**	**P@10**	**R@5**	**R@10**	**nDCG@5**	**nDCG@10**
Original	0.2991	0.3503	0.1608	0.0974	0.3557	0.4239	0.3402	0.3606
Neutral QE	0.2225	0.2639	0.1145	0.0691	0.2682	0.3149	0.2547	0.2700
Positive QE	0.1852	0.2114	0.0981	0.0602	0.2330	0.2773	0.2130	0.2282
Negative QE	0.2132	0.2469	0.1108	0.0676	0.2491	0.2988	0.2401	0.2555

Table 1: Hybrid Retrieval Performance Across Query Variants

The hybrid baseline using **the original queries achieves the highest scores** across all metrics. All query expansion strategies result in a performance degradation, indicating that LLM-generated expansions introduce semantic noise rather than improving ranking quality.

Among the expansion strategies, **Neutral QE** is the most stable but still underperforms the original queries. **Positive (bullish) QE** produces the largest performance drop, while **Negative (bearish) QE** consistently performs better than Positive QE. Table 2 presents pairwise performance differences, confirming a systematic penalty for all expansion variants.



	MAP	P@1	P@5	P@10	R@5	R@10	nDCG@5	nDCG@10
Original vs Neutral QE	-0.0766	-0.0864	-0.0463	-0.0282	-0.0875	-0.1090	-0.0854	-0.0905
Neutral vs Positive QE	-0.0373	-0.0525	-0.0164	-0.0090	-0.0352	-0.0376	-0.0417	-0.0418
Neutral vs Negative QE	-0.0093	-0.0170	-0.0037	-0.0015	-0.0191	-0.0161	-0.0146	-0.0145
Positive vs Negative QE	-0.0280	-0.0355	-0.0127	-0.0074	-0.0161	-0.0215	-0.0271	-0.0273

Table 2: Pairwise Performance Differences

This behavior is primarily explained by **query drift**: appending multiple expansion terms shifts the semantic focus away from the original information need. In the FinQA dataset, characterized by short queries and sparse relevance judgments, this dilution effect is particularly harmful.

A notable finding is the **polarity asymmetry** between sentiment-controlled expansions. Negative QE is more resilient than Positive QE, suggesting that risk-oriented vocabulary aligns better with the technical and caution-driven nature of financial opinion content. Thus, the results indicate that in financial opinion search, preserving concise query intent is more effective than introducing sentiment-driven semantic breadth.

5. Discussion and Conclusion

This study highlights that in financial information retrieval, **semantic enrichment is not inherently beneficial**. While LLM-based query expansion can inject domain knowledge and controlled sentiment, we learned that altering the linguistic structure of short financial queries often weakens the retrieval signal rather than strengthening it. Precision in intent representation proves more valuable than semantic breadth.

The main difficulties stem from the nature of the task and data. Financial queries are short and underspecified, relevance judgments are sparse, and the domain contains substantial lexical ambiguity. In this setting, even minor shifts in query formulation can produce large ranking effects, revealing the fragility of hybrid retrieval systems to linguistic manipulation.

Rather than expanding queries blindly, future work should explore **more selective uses of LLMs**, such as intent-preserving query rewriting, gated expansion, or pre-retrieval filtering. These directions would better exploit generative models without compromising retrieval focus.

Ultimately, the value of this system lies in its ability to expose how sentiment steering interacts with ranking behavior. By stress-testing hybrid retrieval under controlled polarity manipulation, this work provides insight into the limits of LLM-assisted search and underscores the need for **bias-aware integration** in financial IR systems.