Mitigating Echo Chambers

Team Members:

1. Ariele Mairani
2. Nicolò Vella
3. Luca Santagati

Abstract

This study addresses the mitigation of echo chambers in financial question answering. After indexing the entire collection the starting points were three initial experiments with baseline models such as BM25, TF-IDF and BM25+RM3 that established a maximum Mean Average Precision (MAP) of 0.21. We subsequently evaluated three advanced approaches: Cross-Encoder Re-Ranking, Sentiment-Diversified Re-Ranking, and Neural Query Expansion. Although performance varied across the experimental setups, the optimal configuration demonstrated significant efficacy, achieving a MAP of 0.29. This represents a substantial 39% improvement over the baseline, validating the proposed retrieval pipeline. Furthermore, sentiment diversification was increased by 6.26% with minimal relevance loss.

1 Introduction and Motivation

The main problems are to mitigate the vocabulary mismatch, echo chamber effect and poor quality queries. Standard IR systems typically optimize for topical relevance; in the financial sector, high relevance scores often correlate with the most popular or dominant viewpoint. This tendency creates an "echo chamber" effect, where subjective opinions are reinforced as objective truths, potentially obscuring alternative market sentiments.

The system is designed to be universally accessible, not just for financial experts but it is for all people interested in the financial field. Mitigating echo chamber effect is not only a technical objective but a requirement for effective risk management, because providing a diverse set of options is essential for users to make unbiased and well-informed decisions.
The primary objective of this project is to develop a robust IR system that addresses these domain-specific challenges in the Financial Question Answering domain. Specifically, our research focuses on the following question:
To what extent can a sentiment-diversified re-ranking framework mitigate the formation of echo chambers in financial question answering systems without compromising retrieval relevance?
We divided the work of the project evenly in three parts; in the first Phase every component developed a simple baseline model and we did the same for the second Phase by choosing one advanced experiment each.

2 Task and Dataset Description

The project addresses a specialized Financial Question Answering (QA) problem that is primarily opinion-based and multi-faceted. The system is designed to handle a heterogeneous financial corpus characterized by diverse data formats and varying levels of formalization: the dataset integrates information from microblogs, official financial reports, and news articles.

Its key features are the presence of both macro (real estate, market trends) and micro (specific asset sentiment) financial data, objective "factual" data and subjective "opinionated" content as well as a wide range of reliability, from authoritative institutional reports to noisy, informal social media posts.
We have identified several critical challenges inherent to the financial domain that our retrieval pipeline must overcome; namely the domain-specific vocabulary and subtle distinctions between technical terms, which is often arduous for generic embedding models to capture, the noisy and short text typical of microblogs, which often contain high levels of noise compared to official documentation, as well as the significant gap exists between the "natural language" used by laypersons in queries and the "expert vocabulary" found in professional financial corpora.

Metrics	Values
Total documents	57,638
Total postings	2,714,611
Total tokens	3,783,214
Unique terms	51,260
Average document length (tokens)	65.64

3 Methodology

3.1 Baseline Systems (Phase I)

To establish a performance benchmark for Financial Question Answering, we implemented three distinct baseline retrieval strategies. All baseline runs achieved a Mean Average Precision (MAP) score of 0.21, serving as the foundation for evaluating our advanced re-ranking pipeline.

- TF-IDF (Term Frequency-Inverse Document Frequency):
  We employed TF-IDF as a "naive" lexical baseline. This method ranks documents based strictly on the frequency of query terms, weighting them by their rarity across the corpus.
- BM25 (Best Matching 25):
  This was our primary Sparse Retriever. BM25 improves upon TF-IDF by incorporating probabilistic scoring and document length normalization.
- RM3 query expansion (Relevance Model):
  The language model is supposed to improve search results by first retrieving top documents using the original query (BM25), then extracting and scoring terms from these documents based on relevance.

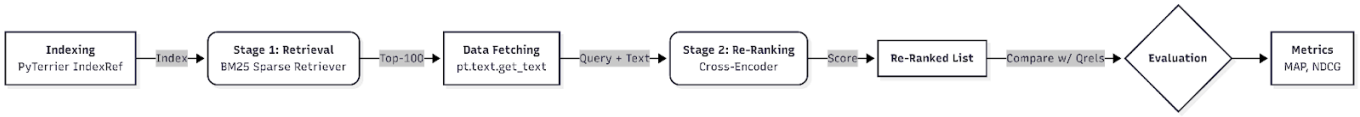
3.2 Advanced Systems (Phase II)

Cross-Encoder Re-Ranking

Our initial baseline relied on BM25, a sparse retrieval model that ranks documents based on keyword overlap and frequency. While efficient, this approach often struggles with the "vocabulary mismatch" problem common in finance, where specific jargon or synonyms do not textually match the user's query. To address this, our advanced experiment implements a Cross-Encoder Re-ranking pipeline. Unlike standard bi-encoders that process queries and documents separately, the Cross-Encoder jointly encodes the query and document pairs. As seen in our implementation, this allows the model to capture deep semantic interactions and fine-grained relevance that keyword matching misses.

We hypothesize that introducing a semantic re-ranking stage will significantly improve retrieval effectiveness (measured by MAP and NDCG) compared to the lexical baseline. Specifically, we posit that the Cross-Encoder will better identify relevant financial documents that use complex terminology, effectively bridging the semantic gap between layperson queries and expert content. This is a necessary precursor to our final goal of sentiment-diversified re-ranking, as we must first ensure high relevance before filtering for diversity.

Pipeline flowchart

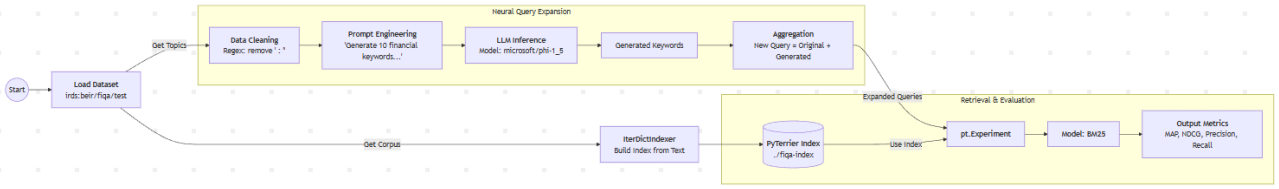


LLM Query Expansion

The baseline RM3 query expansion relies on a probabilistic approach that extracts and scores terms from pseudo-relevant documents based on frequency and co-occurrence patterns. This method is limited by the quality of the initial retrieval and may miss semantically relevant terms if they don't appear in top-ranked documents. This experiment aims to enhance retrieval relevance by leveraging semantically meaningful expansion terms generated directly from the original queries.

The chosen LLM model, Microsoft Phi-3-mini-4k-instruct, is fine-tuned for instruction-following tasks and is well-suited to provide enriched queries for our BM25 retrieval system. Unlike RM3's reliance on retrieved document statistics, this approach leverages the model's pre-trained semantic knowledge to identify contextually appropriate keywords for the financial domain. The LLM-enriched queries are then processed through our BM25 retrieval system.
We hypothesize that neural query expansion will generate semantically relevant terms more effectively than statistically frequent ones, leading to improved retrieval effectiveness. However, we are aware of the many challenges regarding Large Language Models, such as hallucination and over-expansion. To mitigate these issues, we tune generation parameters such as maximum token limits and design a targeted prompt that constrains the model to produce correctly formatted, focused output.

Pipeline flowchart



Sentiment Diversification

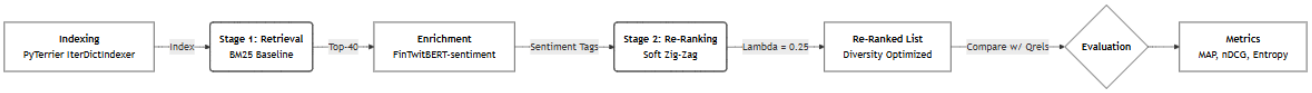
Following initial retrieval, documents undergo Sentiment Enrichment via the FinTwitBERT-sentiment model to classify

content as bullish, bearish, or neutral. These labels inform a Soft Zig-Zag Re-ranker that greedily selects documents by balancing normalized relevance scores against a diversity penalty. The objective function is defined as:

$$Score(d) = (1 - \lambda) * Relevance(d) + \lambda * Diversity(d)$$

Experimental results confirm that a conservative re-ranking parameter ($\lambda = 0.25$) achieves a 6.26% improvement in Shannon Entropy while maintaining a high nDCG@10, effectively diversifying the user's exposure to financial perspectives without introducing irrelevant content.

Pipeline flowchart



3.3 Implementation Details

Cross-Encoder Re-Ranking

This script establishes a two-stage information retrieval pipeline using PyTerrier on the FIQA dataset. The process begins with indexing, where an inverted index is created via IterDictIndexer, storing truncated document text directly in the metadata to facilitate downstream reranking. Before retrieval, queries undergo a preprocessing step using Regex to remove punctuation and non-alphanumeric characters.

The retrieval architecture combines sparse and dense methods. It starts with BM25 to fetch a candidate pool of the top 100 documents per query. These candidates are then processed by a custom CrossEncoderReranker class,which leverages the ms-marco-MiniLM-L-6-v2 model from Hugging Face and Sentence-Transformers. This model re-scores the pairs based on semantic relevance before the final performance is evaluated using metrics like NDCG@10 and MAP in a comparative experiment.

LLM Query Expansion

The LLM query expansion pipeline follows the same preprocessing and indexing structure as the RM3 baseline to ensure a direct comparison. Preprocessing involves cleaning query text to ensure compatibility with the language model and retrieval system. The indexing strategy utilizes IterDictIndexer to process the FiQA collection.

The key difference lies in the expansion process: while RM3 requires an initial retrieval stage to extract terms, the LLM approach generates expansion terms directly from the original query. Each query is processed through Microsoft Phi-3-mini-4k-instruct with a carefully designed prompt that instructs the model to generate semantically relevant financial keywords.

These LLM-generated terms are then concatenated with the original query to form an enriched query representation, which is fed into the same BM25 retrieval system used across all experiments. To enable direct comparison with RM3, system performance is evaluated using identical metrics: MAP, nDCG, Precision, and Recall.

Sentiment Diversification

Preprocessing involves cleaning query titles of problematic characters (colons and quotes) to prevent retrieval errors, and account for 38 empty documents identified during corpus inspection.

Our indexing strategy utilizes IterDictIndexer to process the FiQA collection, with metadata fields configured to retain 2000 characters of text per document for transformer-based analysis.

Following initial lexical retrieval via BM25, the top 40 candidates undergo Sentiment Enrichment using the Hugging Face transformers library, specifically the FinTwitBERT-sentiment model, optimized for financial terminology.

Next a Soft Zig-Zag Re-ranker uses a greedy selection loop to re-order documents by maximizing a weighted combination ($\lambda = 0.25$) of normalized relevance scores and a diversity penalty calculated against the running mean sentiment of selected results.

System performance is quantified through Shannon Entropy to measure echo chamber mitigation, while retrieval utility is monitored using nDCG, MAP, Recall and Precision.

4 Experiments and Results

The Cross-Encoder re-ranking stage yielded substantial performance gains across all tracked metrics compared to the BM25 baseline. The Normalized Discounted Cumulative Gain at 10 (NDCG@10), which measures the quality of the top-10 results shown to the user, increased from 0.253 to 0.350, marking a 38.7% improvement. This directly validates the objective of

enhancing the precision of the final ranked list. Simultaneously, Mean Average Precision (MAP) improved from 0.209 to 0.290, a 39.2% jump indicating that the system is successfully bubbling up more relevant documents across the entire retrieval list. Regarding user effort, the Reciprocal Rank improved from 0.321 to 0.435, suggesting that on average, the first relevant document appears significantly higher in the list compared to the baseline. Overall, this experiment successfully addressed the vocabulary mismatch problem, resulting in markedly better performance in retrieving relevant documents.

Regarding the disruption of echo chambers, we can instead note that the Soft Zig-Zag approach generated a 6.26% improvement in Shannon Entropy (from 1.184 to 1.258). This confirms that the re-ranker successfully injected diverse viewpoints into the top results, mitigating the filter bubble effect without "breaking" the retrieval quality. The improvement was achieved while maintaining a Mean Average Precision (MAP) of 0.2032, a negligible decrease from the baseline's 0.2058. This indicates that the system continues to retrieve highly relevant financial information.

4.1 Evaluation Setup

These are the metrics we used:

- The Mean Average Precision: how well the model finds all relevant documents.
- The Normalized Discounted Cumulative Gain at 10: give more points if relevant documents are at the very top compared to lower down.
- P_1 (Precision at 1): Measures if the very first document retrieved is relevant. This is the most critical metric for a Question Answering system, as users expect the immediate answer to be correct.
- P_5 / P_10 (Precision at k): Indicates the "density" of relevant information in your top results. For example, a P_10 score of 0.6 means that, on average, 6 out of the top 10 documents are relevant.
- recall_5 / recall_10 (Recall at k): Measures coverage: what percentage of all the relevant documents existing in the database were successfully found and placed in the top 5 or 10. High recall is essential in this project to ensure diverse viewpoints (e.g., both Bullish and Bearish opinions) are included.
- Shannon Entropy: quantifies the distribution of sentiments within the top-k results for a given query. A low entropy suggests the results are dominated by a single sentiment (an echo chamber), while high entropy indicates a diverse mix of opinions.

To ensure reproducibility, we utilized the pt.Experiment function to perform a side-by-side comparison of all runs, both baselines and advanced.All systems were executed against the same queries from the same dataset to maintain a controlled environment for metric comparison.
For the sentiment diversification run, both systems were evaluated at a retrieval depth of k=40. We implemented a RetrieverCache to store the output of the sentiment enrichment stage, ensuring that the exact same tagged document set was used for all re-ranking iterations and that results remained consistent across multiple executions.
The use of relevance judgments was centered on the official qrels provided by the FiQA dataset to determine the accuracy of each run. Retrieved documents were merged with the qrels using qid and docno as keys. These judgments facilitated the calculation of standard Information Retrieval metrics, including MAP, nDCG, Recall, P_5, and P_10, allowing for an objective assessment of how diversification impacted search utility.

4.2 Results

Cross-Encoder Re-Ranking

The addition of the Cross-Encoder reranker consistently outperforms the BM25 baseline because it moves beyond simple keyword matching to capture semantic relationships. This system combines the efficiency of sparse retrieval with the precision of dense ranking, significantly boosting metrics like NDCG@10 and MAP.

However, this method introduces a substantial computational bottleneck, as the Cross-Encoder must process every query-document pair individually, making it much slower than the initial retrieval. Additionally, there is a risk of domain mismatch; the model is pre-trained on the MS MARCO dataset (general web passages), which may not perfectly align with the specialized financial terminology found in the FIQA dataset. Furthermore, the model likely has a token limit, meaning that despite the code storing 4096 characters, the reranker may truncate the end of longer financial documents, potentially losing critical information located later in the text.

LLM Query Expansion

Method	MAP	P@1	P@5	P@10	Recall@5	Recall@10	NDCG@5	nDCG@10
RM3	0.206	0.221	0.107	0.068	0.242	0.303	0.224	0.245
LLM	0.180	0.182	0.089	0.061	0.219	0.287	0.194	0.219

RM3 consistently outperforms LLM-based query expansion across all metrics, achieving 14.6% higher MAP (0.206 vs 0.180) and significantly better early precision (P@1: 0.221 vs 0.182). The performance gap is largest in top-rank positions, with RM3

showing 21% better P@1, which is critical for question-answering tasks where users expect immediate accurate results.
RM3's way of extracting expansion terms from actual retrieved documents ensures keywords match the collection's terminology. In contrast, Phi-3's general knowledge produces less effective terms for this specific financial corpus, likely due to semantic drift.

Sentiment Diversification

As we previously mentioned, Shannon entropy was improved by 6.26%, effectively increasing the diversity. Nevertheless, this improvement might seem modest compared to, for example, the performance increase in relevancy given by the cross-encoder experiment, however it reflects a refinement of an already pluralistic system; with a high baseline of 1.18 bits, the retriever was already operating at approximately 75% of the theoretical maximum diversity ($H_{max} = 1.585$). Unlike personalization-driven Recommender Systems (RS) that frequently trigger "Filter Bubbles" through iterative feedback loops, Information Retrieval (IR) is anchored by objective query relevance, making it structurally more robust against extreme sentiment isolation. Consequently, we can conclude that the modest gain indicates that our framework effectively disrupts sentiment homogeneity while maintaining the system's primary utility as a reliable financial discovery tool.

Despite these gains, the system faces specific biases and architectural limitations, such as a minority exhaustion problem, where, because the dataset is often imbalanced, the algorithm can "run out" of minority sentiment documents, leading to a collapse of the diversity pattern at deeper ranks. This was highlighted by the choice of a lower λ. A "greedy" diversity strategy tended to exhaust the minority sentiment candidates (e.g., forcing a Bearish document when only low-quality ones remain), which drastically hurt relevance metrics. By using a soft weight, it effectively functions as a "tie-breaker," re-ordering documents with similar relevance scores to maximize diverse exposure.

Finally we report inconsistencies in the data as the indexing process identified 38 empty documents within the corpus, which provide no linguistic value for sentiment or relevance assessment.

5 Discussion and Conclusions

Cross-Encoder Re-Ranking

This experiment confirmed that semantic reranking effectively resolves the vocabulary mismatch inherent in keyword-based search, significantly elevating retrieval quality beyond what BM25 alone can achieve. However, the primary difficulty was the substantial increase in computational latency required for pairwise scoring, alongside the challenge of applying a general-purpose model to the specialized financial terminology of the FIQA dataset. Future iterations would benefit from fine-tuning the model on domain-specific data or implementing quantization techniques to reduce inference time without sacrificing accuracy. Ultimately, the system demonstrates that while hybrid architectures deliver superior relevance, optimizing the trade-off between precision and speed remains the critical hurdle for real-world deployment.

LLM Query Expansion

The LLM-based query expansion experiment produced lower retrieval effectiveness than traditional RM3 expansion, validating our concerns about the technology.
Given these results, we would not recommend integrating LLM-based expansion into a broader echo chamber mitigation pipeline, as it introduces complexity without enhancing the core objective.
Future work could involve a fine-tuned model and query reformulation, but the high computational expense might not be worth the marginal improvement.

Sentiment Diversification

The primary learning was that diversity functions best as a "tie-breaker" for documents with similar BM25 relevance. Using a low lambda avoids "bulldozing" the search results with irrelevant content simply because it possesses a dissenting sentiment.
To advance this system, future exploration should focus on dynamic lambda scaling, which would adjust the diversity weight based on the initial sentiment density of the top-k results.

AI Tools and Assistance Declaration

Regarding the BM25 basic experiment, and the sentiment diversification experiment we disclose the usage of AI tools for the following: Google's Gemini was used to explore the possibilities of implementation, strictly at a theoretical level, aiding in determining the usage of a soft zigzag function, as well as Shannon entropy as metric; further, given that the entropy calculation is already described by a mathematical equation, its translation into code was requested to the AI model, generating the core functionality of the "calculate_shannon_entropy" function.
Regarding the rest, AI was also used for formalizing some report parts and to have a suggestion for some code errors and imputation.