UNIVERSITÀ
DI PAVIA

Mitigating Echo Chambers

Ariele Mairani
Nicolò Vella
Luca Santagati

Introduction

The context

Financial IR systems optimize for topical relevance, often favoring dominant viewpoints.

"Echo Chamber"

Vocabulary mismatch, poor query quality, lack of diverse sentiment

Research Objective

Research Question:

To what extent can a sentiment-diversified re-ranking framework mitigate echo chambers in financial QA systems without compromising retrieval relevance?

Methodology

Phase I: Development of simple Baseline Models.

Phase II: Advanced Experiments

Baseline Systems (Phase I)

BM25

We employed BM25 as current and future reference as the “de facto” golden standard for sparse retrieval.

It represents the control group

TF-IDF

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.

RM3

We employed RM3 as a query expansion baseline. The language model creates the most pseudo-relevant words from an initial retrieval to create the expanded queries.

Advanced Systems (Phase II)

Cross-Encoder Re-Ranking

- Solve the vocabulary mismatch
- Encode query and document pairs
- Optimize the trade-off

LLM Query Expansion

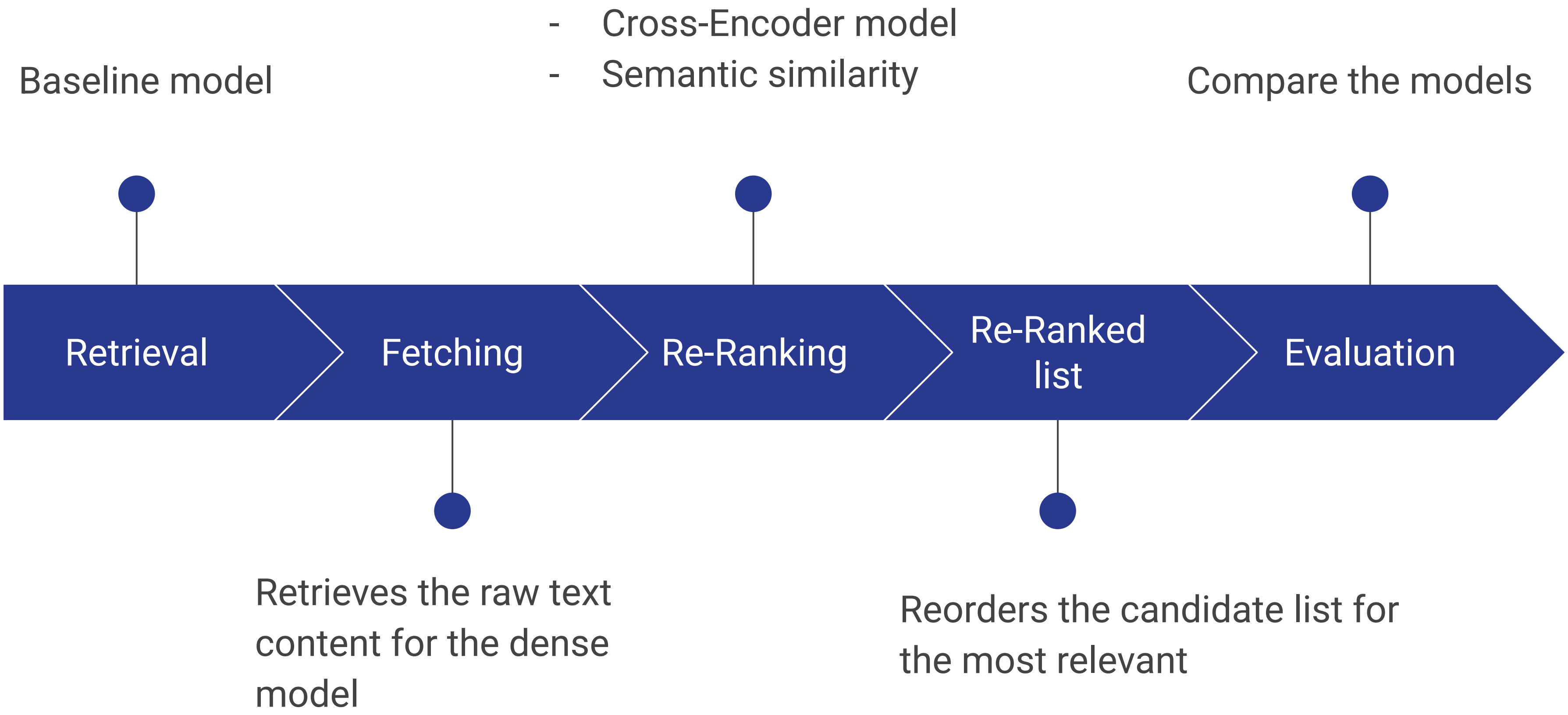
- Leverage pre-trained semantic knowledge
- Capture financial concepts beyond corpus terms

Sentiment Diversification

- classify documents based on sentiment
- determine the current amount of diversity
- increase it through a soft zig-zag approach

Implementation

Cross-Encoder Re-Ranking Pipeline



Retrieval

Baseline model

Fetching

Retrieves the raw text content for the dense model

Re-Ranking

- Cross-Encoder model
- Semantic similarity

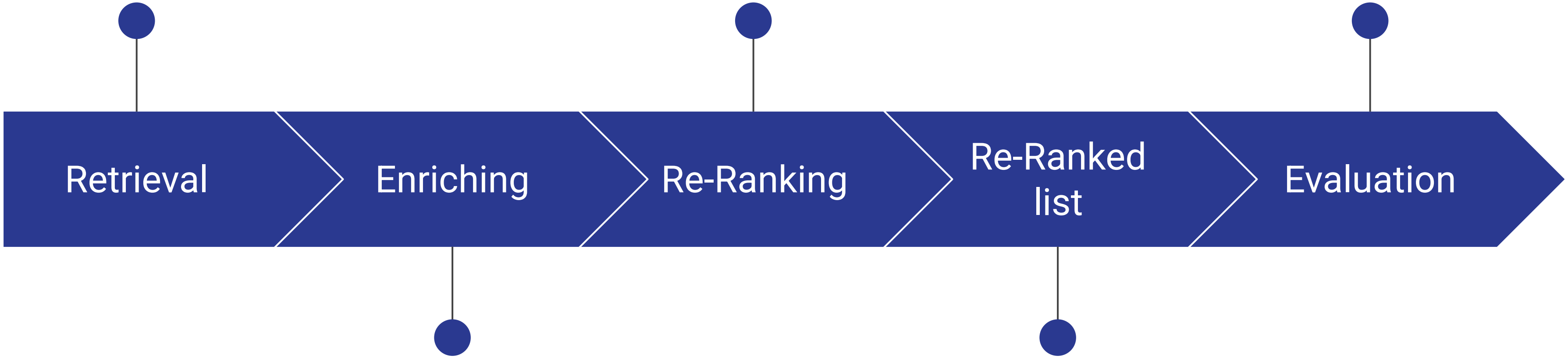
Re-Ranked list

Reorders the candidate list for the most relevant

Evaluation

Compare the models

Sentiment Diversification Pipeline



- BM25
- Query cleaning

soft zig-zag

6.26% Shannon
entropy increase

Retrieval

Enriching

Re-Ranking

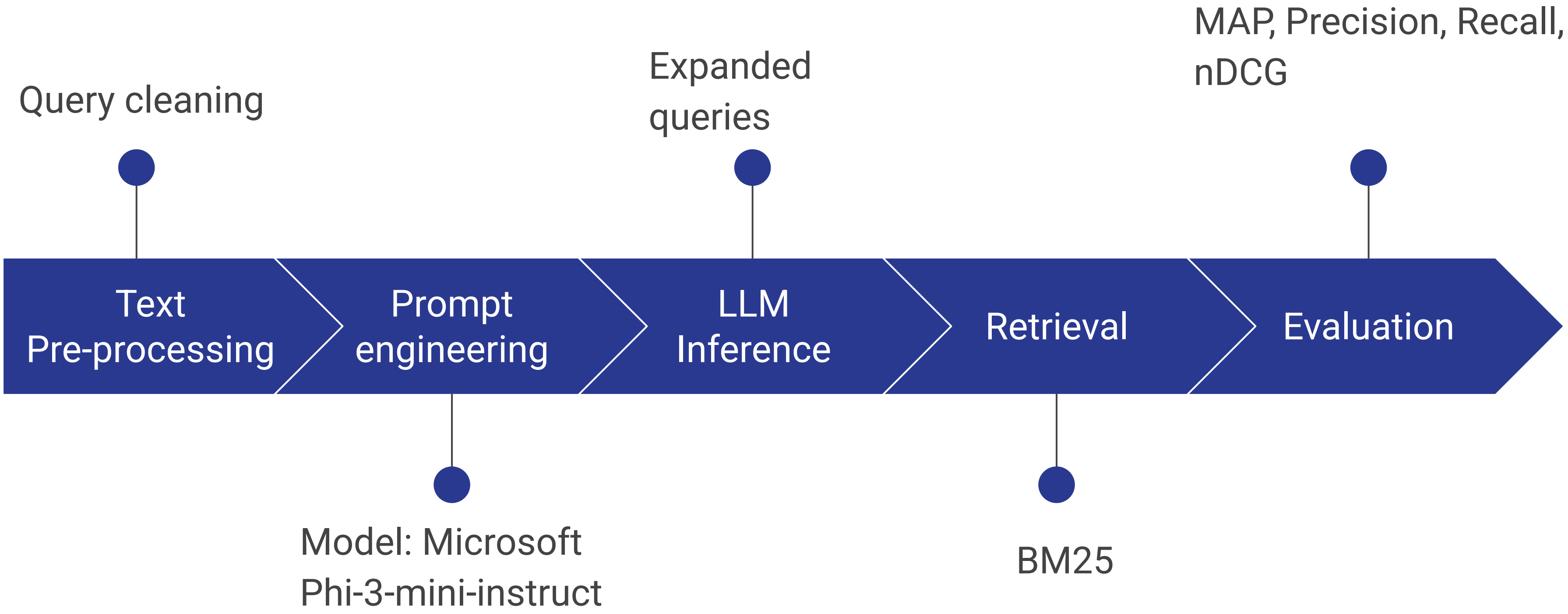
Re-Ranked list

Evaluation

FinTwitBERT-sentiment

diversity optimized

LLM Query Expansion



Query cleaning

Text Pre-processing

Prompt engineering

Model: Microsoft Phi-3-mini-instruct

Expanded queries

LLM Inference

Retrieval

BM25

MAP, Precision, Recall, nDCG

Evaluation

Results and conclusions

Sentiment Diversification

Results

We obtained a 6.26% increase in entropy, thus successfully addressing the bias blindness of sparse retrieval model

Discussion

The increase was modest as the data was already somewhat diversified, and as such the pipeline worked best as a “tie-breaker” for documents with similar relevance, rather than a complete re-ranker

Limitations

Character length was limited due to the model only accepting 512 tokens.

This could have caused interference with the classification, though none was found empirically

Cross-Encoder Re-Ranking

Results

NDCG@10: From 0.253 → 0.350 (**+38.7%**).

MAP: From 0.209 → 0.290 (**+39.2%**).

Significance: $p < 0.05$

Discussion

Semantic Shift: Deep semantic interactions

Hybrid Power: Successfully combines sparse retrieval efficiency with dense ranking precision.

Limitations

Latency Bottleneck: Query-document pairs.

Domain Mismatch: MS Marco is a standard model

LLM Query Expansion

Results

LLM-based query expansion underperformed RM3 across all metrics (MAP: 0.180 vs 0.206)

Largest gap in early precision (P@1: -21%).

Discussion

General pre-trained knowledge proved less effective than corpus-grounded term extraction, RM3's document-based expansion better captures collection-specific financial terminology.

Limitations

Phi-3-mini lacks domain-specific financial fine-tuning.

Small model size (4B parameters) limits specialized knowledge.

No retrieval context during generation (unlike RM3).