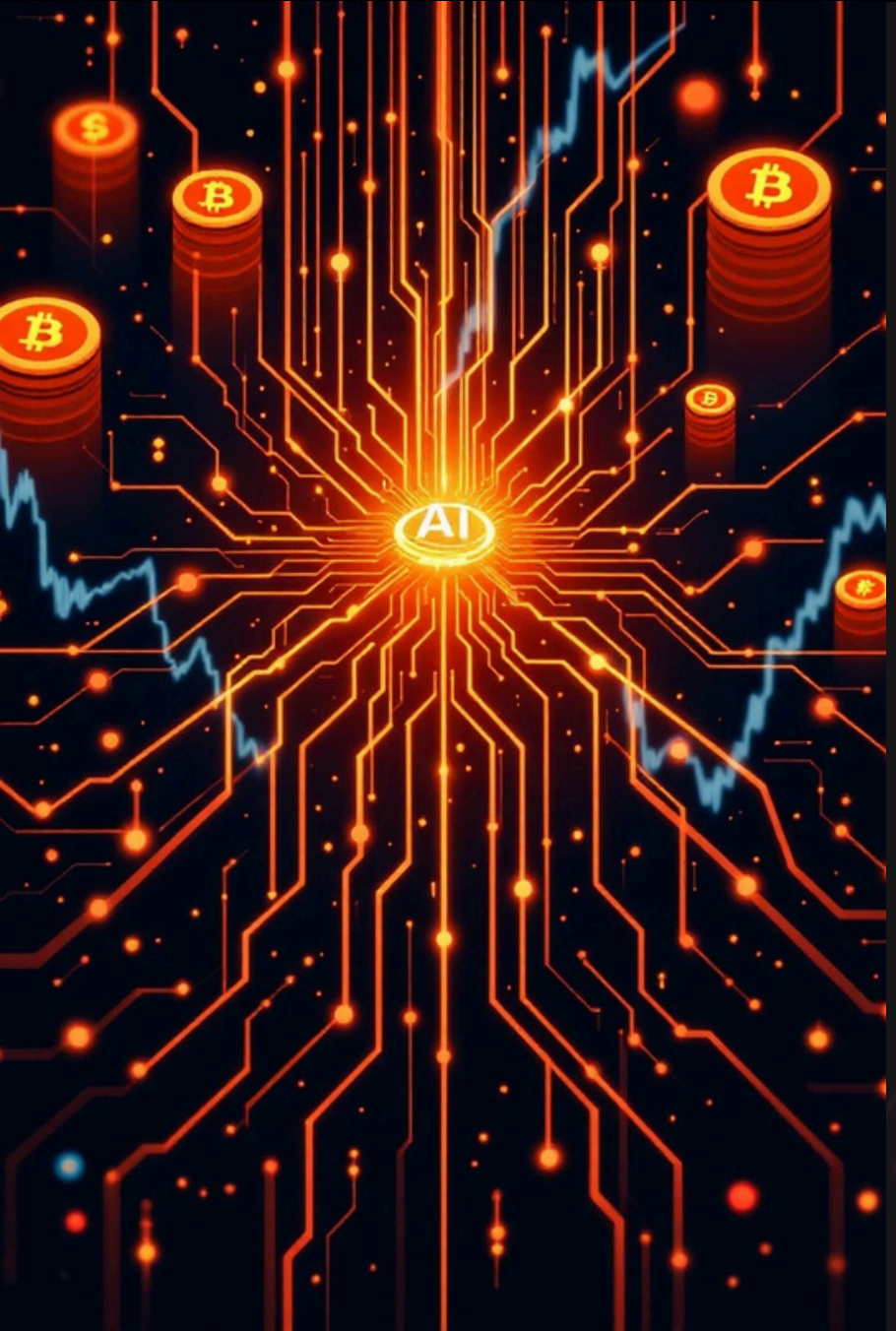
FinAlign

Bridging the Financial Vocabulary Gap

Matteo Pietro DiPilato, Lorenzo Cinquemani, Lorenzo Marchioni



Abstract: Hybrid Retrieval for FiQA

This project tackles Financial Question Answering (FiQA) with a hybrid retrieval system.
We integrate lexical (BM25) anddense (BGE-base) retrieval, exploring advanced query expansion like "Hypothetical Document Embeddings" (HyDE) and LLM-driven keyword expansion.

Key Finding

LLM-driven query rewriting significantly outperforms baselines, bridging the domain-knowledge gap for non-experts.

Performance Boost

MonoT5 reranker further improved performance, highlighting multi-stage pipelines' importance in domain-specific retrieval.

Introduction & Motivation

Traditional search struggles with complex financial documents due to keyword reliance. Our system, leveraging the FiQA dataset, moves beyond simple word-matching to understand user intent, primarily for non-experts.

> *Does the combination of query rewriting, prefix-based encoding, and document chunking improve the effectiveness of a hybrid (BM25 + dense) retrieval system?*

We hypothesize these techniques mitigate semantic mismatch and improve retrieval quality, empowering individuals to make better financial decisions.

Task & Dataset Description

Task Definition

Opinion-based and explanatory retrieval: retrieve documents that help answer financial questions, offering advice, explanations, or interpretations of financial concepts.

Dataset: FiQA (BEIR Benchmark)

- **Corpus: ~57,000 documents(news,microblogs, reports).**
- **Queries:** Natural language questions (e.g., "What is the difference between a REIT and a real estate stock?").
- **Qrels:** Graded relevance annotations for evaluation.

Challenges

- **Heterogeneity:** Formal reports mixed with informal social media text.
- **Vocabulary Gap:** Mismatch between user terms and document content.
- **Sparsity:** Few relevant documents per query, making recall difficult.

Methodology: Baseline Systems

We established benchmarks with traditional models to measure initial performance.

TF-IDF & BM25

PyTrier's standard implementation. BM25 served as our primary strong baseline for exact term matching.

BM25 + RM3

Pseudo-relevance feedback (RM3) tested statistical query expansion to address the vocabulary gap without external models.

Methodology: Advanced Systems

Our advanced pipeline integrates multiple stages for enhanced financial question answering.

Advanced Systems: Key Components

1

Query Rewriting & HyDE

LLM-based rewriting generates enriched queries and hypothetical answers to guide dense retrieval, combining original, rewritten, and HyDE representations for improved recall.

2

Hybrid BM25 + BGE

Combines BM25 for lexical recall and BGE bi-encoder for semantic similarity. Query-only prefix-based encoding guides the BGE model.

3

Document Chunking

Documents split into overlapping fixed-length chunks to reduce semantic dilution. Dense retrieval at chunk level, with max pooling for document scores.

4

Neural Re-ranking (MonoT5)

MonoT5cross-encoderre-rankstop candidates, enabling fine-grained relevance modeling for improved accuracy.

Experiments & Results

Evaluation on a cleaned BEIR/FIQA testset (57,600documents, 648 queries) using standard IR metrics.

name	map	P@1	P@5	P@10	R@5	R@10	nDCG@5	nDCG@10
TF-IDF	0.209578	0.234568	0.108642	0.071142	0.249805	0.313278	0.231198	0.253309
BM25	0.210418	0.236111	0.106790	0.070370	0.247691	0.309708	0.230294	0.252634
BM25_RM3	0.206510	0.220679	0.107099	0.067593	0.243568	0.303302	0.225140	0.245356
BGE	0.334822	0.388889	0.171605	0.110185	0.373849	0.457247	0.363043	0.391509
BGE (with Instruction)	0.347504	0.395062	0.184259	0.112191	0.404397	0.480580	0.382735	0.406301
Hybrid (Linear Normalized)	0.331388	0.367284	0.172222	0.111420	0.380137	0.472197	0.360711	0.392579
Hybrid (RRF)	0.314226	0.348765	0.165123	0.107099	0.367176	0.460231	0.343518	0.374995
BGE (with Instruction) + LLM query rewriter	0.357263	0.413580	0.181481	0.113735	0.397200	0.484324	0.386135	0.414345
Hybrid (Linear Normalized) + LLM query rewriter	0.334057	0.387346	0.175926	0.112346	0.387294	0.482136	0.366782	0.398061
BGE (with Instruction) + LLM query rewriter + chunking	0.359115	0.412037	0.183333	0.114506	0.403627	0.488636	0.389035	0.417210
Hybrid (Linear Normalized) + LLM query rewriter + chunking	0.338980	0.396605	0.177160	0.111574	0.392649	0.478734	0.373354	0.401661
Hybrid (Linear Normalized) + LLM query rewriter + chunking + monoT5 rerank	0.356564	0.410494	0.186111	0.115586	0.415902	0.496233	0.396234	0.421273

The advanced pipeline with BGE, LLM query rewriting, chunking, and monoT5 reranking achieved the highest performance.

Discussion & Conclusions

What We Learned

- Combinin gretrieval models is complex.
- Dense retrievers benefit from document chunking.
- Query rewriting quality is critical.
- Instruction-based query prefixing for BGE improves performance.
- Computational constraints significantly impact feasibility.

Main Difficulties

- High computational cost for experiments.
- Challenges in integrating multiple models, especially LLM-based rewriting.
- Data heterogeneity.

Future Work & Takeaway

Future Work

- Larger/fine-tuned models for dense retrieval and query rewriting.
- Systematic hyperparameter tuning for fusion weights.
- Difficulty-aware re-ranking strategies.
- Extend to Retrieval-Augmented Generation (RAG) for natural language answers.

Takeaway

- Effectivefinancial information retrieval requires careful pipeline design, semantic alignment, and practical engineering.
- Hybrid retrieval, document chunking, and instruction-aware dense encoders offer clear benefits when thoughtfully combined.