A Comparative Study of BERT, FinBERT and Financial RoBERTa in Financial QA Tasks

Mayuri Mamdi
Dublin City University
School of Computing
mayurivijay.mamdi2@mail.dcu.ie

Saafiya Mudanaldesai
Dublin City University
School of Computing
saafiyakhajapeera.mudanaldesai2@mail.dcu.ie

***Abstract*—This paper provides an examination of three transformer models trained on relevance based finance question answering (QA): FinBERT, BERT, and Financial RoBERTa. This study mainly focuses on the financial domain. It consists primarily of sentence level relevance classification on the Financial Question Answering (FiQA) 2018, which contains real world finance questions with annotated answer candidates. Best Match 25 (BM25), a probability based retrieving function, is used to retrieve the best k nearest sentences in every request. These question-sentence pairs are then reordered by transformer models that have been trained to decide if the pair is relevant or irrelevant. The model chceks how well the sentence answers the question and give it a score, and based on that score the highest will be ranked first. So the models that are BERT, FinBERT and Financial Roberta use Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG), and Precision@1 to measure performance. The findings indicate that the three models exhibit favorable behavior over many situations. However, Financial RoBERTa demonstrates a stronger comprehension of the domain, making it the preferable option when the application desires a high level of certainty in the financial questions and answers. FinBERT performs poorly despite being domain specific because of its corpus size and coverage limitations as well as the absence of sophisticated pretraining techniques found in Financial RoBERTa. Compared to the domain specific Financial RoBERTa, BERT, a general-purpose model, finds it difficult to capture nuanced financial context.**

***Index Terms*—FinBERT (Financial Bidirectional Encoder Representations from Transformers) , RoBERTa (Robustly Optimised BERT), Financial NLP (Natural Language Processing), Question Answering (QA), Transformer Models, BM25, Relevance Classification**

I. Introduction

In today's data centric financial landscape, the ability to retrieve precise information with contextual relevance from overwhelming amounts of unstructured text has become critically important. Financial professionals often face the burdensome task of extracting accurate answers to complex questions from dense and highly specialized sources such as Securities and Exchange Commission (SEC) filings, earnings reports, and Environmental, Social, and Governance (ESG) disclosures, which may contain potentially crucial information for compliance, risk analysis, and investment decision making. Simultaneously, the rate at which market intelligence is disseminated continues to accelerate. In response, Financial Question Answering (QA) systems that incorporate financial text extraction have become more widely adopted to support market analysis, decision making, regulatory compliance, and risk assessment [1].

Traditional information retrieval systems typically rely on lexical matching methods such as BM25, which, while efficient, are often inadequate for handling the semantic subtleties and domain specific vocabulary found in financial text [2]. These approaches can produce contextually irrelevant or careless responses, particularly in high stakes domains like finance. The emergence of transformer based models has revolutionized the field of natural language processing (NLP), offering deep contextual understanding. A notable breakthrough was BERT (Bidirectional Encoder Representations from Transformers), which introduced bidirectional attention mechanisms that enabled a more nuanced grasp of semantic relationships in text [3]. RoBERTa (Robustly Optimized BERT Pretraining Approach) advanced this architecture by introducing dynamic masking, increased training data, and longer training durations [4].

Building on these innovations, domain specific models such as FinBERT and Financial RoBERTa have been developed through pretraining on large scale financial text corpora, including financial news, SEC filings, earnings call transcripts, and ESG reports. FinBERT, pretrained on corporate disclosures and earnings calls, has shown robust performance in financial sentiment analysis and text classification tasks [5]. Financial RoBERTa, trained on ESG reports and financial media content, demonstrates improved contextual modeling of financial language and achieves superior downstream task performance compared to general purpose models [6].

Recent studies have explored hybrid pipelines for financial QA, particularly two stage methods that employ an initial BM25 retrieval followed by reranking with transformer based models. For example, FinBERT QA demonstrated strong performance on the FiQA leaderboard in terms of ranking metrics such as normalized Discounted Cumulative Gain (nDCG) and Precision@1 [7]. Additionally, newer encoders like FinBERT have been pretrained on broader financial corpora, further advancing the capabilities of domain specific language models [8].

Although retrieval augmented generation (RAG) systems powered by large language models (LLMs) such as GPT and T5 have achieved state of the art results on QA tasks, their

high computational costs, latency and lack of interpretability make them impractical for resource constrained or real time financial applications [1], [2]. Given these limitations and limited computation power, this study focuses on encoder only transformer models, which offer improved efficiency, modularity, and interpretability compared to RAG-based approaches. While encoder based models are gaining traction, they have not been comprehensively benchmarked for sentence level financial QA using naturally occurring evaluation datasets such as FiQA [9]. By focusing specifically on encoder architectures, this study aims to expand the understanding of their potential in sentence level financial QA and support future research in interpretable, efficient financial information retrieval systems.

To address this gap, our study evaluates the performance of three encoder based transformer models BERT, FinBERT, and Financial RoBERTa within a financial QA pipeline. This pipeline follows a two stage architecture: first, BM25 is used to retrieve the top-k (where $k = 10$) candidate sentences per question; second, each sentence is semantically reranked using one of the three encoders. We use the FiQA 2018 dataset [9], a well established benchmark in financial QA research, to assess model performance. This project seeks to determine which model offers the best trade-off among contextual relevance, interpretability, and efficiency in sentence level financial QA.

II. Research Goal

The main objective of this study is to compare the effectiveness of three encoder-only transformer models BERT, FinBERT, and Financial RoBERTa in accurately identifying the most relevant sentence that answers a financial question from a set of candidate sentences. Evaluating the relative performance and practical utility of these models highlights their significance for financial question answering (QA) systems, particularly within computationally efficient, retrieval based pipelines.

A. *Research Questions*

In order to direct the investigation, the following research questions were created:

RQ1 Which of the pre trained encoder only language models performs best in retrieving appropriate answers to financial questions based on Mean Reciprocal Rank (MRR@10), Normalized Discounted Cumulative Gain (nDCG@10), and Precision@1?

RQ2 What types of recurring errors have occurred in the applications of these models to the financial question-answering task, and what strategies can be used to overcome them?

III. Related Work

Those who put close attention to the latest NLP trends would probably admit that in the last several years financial question-answering (QA) has got certain significant steps [1]. Much of that advancement has been tied to the use of transformer architectures and domain specific pretraining [10], [11]. As a matter of fact, today, most of these modern NLP systems already have pretrained language models at the core of their taking, be it sentiment analysis, text classification, or QA.

Consider FinBERT proposed by Araci et al. in 2020 [5]. It is basically a refined version of BERT which has learned vast data of financial information, such as security filings and transcripts of the earnings call. Since the very beginning, this sentiment detection was carried out with the help of the model, but it was not long before its scope extended into sentiment classification, market prediction, entity recognition, and long range document analysis. Financial RoBERTa was introduced in 2022 on the basis of this success, with RoBERTa re trained on ESG reports, financial news articles, and finance related threads on Reddit [6].

Previous research however had been more task specific in that they wanted to accomplish tasks like sentiment analysis and topic tagging on a document by document level [10]. Take the example of FinBERT and Financial RoBERTa, which both performed well in these tasks although there is less work regarding sentence level QA the type that seeks to determine specific answers instead of general sentiments.

The new QA benchmarks have attempted to expand the aperture. FinQA (EMNLP 2021) and ConvFinQA (EMNLP 2022) added multi-hop reasoning and table and report comprehension, even conversational QA [9], [12]. FinTextQA [13]approached QA differently, building a retrieval augmented benchmark with narrative questions that are much longer than the text responses to the questions. The difficulties that big LLMs still encounter in the finance industry are highlighted by the comprehensive FinanceBench benchmark [14], which assesses encoder and decoder language models on open book financial QA.

When you follow the domain specific transformers research, a single thing will be immediately obvious: a direct comparison between FinBERT and Financial RoBERTa with the same setting and data is not present [8]. Prior work paid attention to FinBERT and either only on FinBERT or made comparisons with generic BERT or RoBERTa that had weakly matched ground and verification circumstances [7]. When everything was in order, the same dataset, the same prediction pipeline, and the sentence wise relevance classification problem then researchers were able to see more clearly how these three models behaved and where they all performed best when answering specific, relevance-based financial questions [15], [16].

This study differs from prior work in three ways:

- It performs comparison of FinBERT, BERT and Financial RoBERTa in a financial QA context using the FiQA 2018 dataset.
- Firstly it employs the BM25 Retrieval of top-k Candidates we check the evaluation and then there is a consistent two-stage BM25 retrieval and classification pipeline, us-

ing BM25 for candidate selection and transformer models for relevance scoring.

- In order to find patterns in failure cases, it uses an in depth error analysis, which is frequently missing from previous research.

Thus, this research fills a notable gap in the literature by evaluating two widely used financial NLP models under the same experimental setup and providing insights into their relative strengths and limitations for practical QA applications.

IV. Dataset

In this paper, we chose to use the FiQA 2018 [17] dataset to experiment with some ideas on how to approach the problem of financial question answering (QA). This dataset was compiled as a benchmark in the particular field, comprising a set of real world questions in finance gathered from news items, discussions on finance, financial forums, and social media. Each question was accompanied by several response sentences [9], [1].

To prepare the dataset for training and evaluation, a two step preprocessing approach was used:

1) **Retrieval stage**: The BM25 algorithm was applied to retrieve the top-k candidate sentences for each question based on keyword similarity [16], [2].
2) **Labeling stage**: Each question–candidate pair was assigned a binary label:
   - **1 (Relevant)**: The candidate sentence correctly answers the question.
   - **0 (Irrelevant)**: The candidate does not answer the question.

A. *Raw Dataset Statistics*

This dataset FIQA 2018 [17] is specifically Task 2. It has 6649 unique questions, 57,599 unique documents, and 17,111 question answer pairs. This Dataset has strong Benchmark on financial question answering tasks.

TABLE I
FiQA 2018 Dataset – Dataset Statistics

Statistic	Count
Unique Questions	6,649
Unique Documents (Candidate Sentences)	57,599
Total Question-Doc Pairs (Ground Truth)	17,111

B. *Final Dataset Used for Training and Evaluation*

After applying BM25 retrieval to select top-k candidate sentences for each question, Created a larger dataset of labelled question candidate pairs for supervised training.

TABLE II
Final Dataset Statistics After BM25 Retrieval and Labelling

Split	Number of Questions
Training	5,676 samples (85.48%)
Validation	631 samples (9.50%)
Test	333 samples (5.02%)
Total	6,640 samples

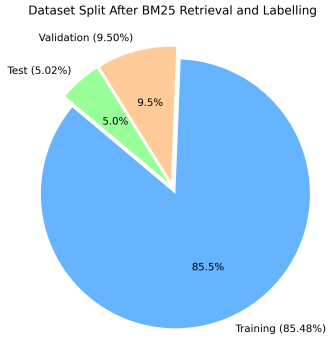

Fig. 1. Proportional distribution of dataset samples across training, validation, and test splits after BM25 retrieval and labeling.

TABLE III
Label Distribution in Training and Validation Sets

Split	Relevant Pairs	Irrelevant Pairs	Total Pairs
Training Set	14,603	277,827	292,430
Validation Set	1,594	30,874	32,468

As Table 1. shows the dataset contains 6,649 unique questions, Unique Documents are 57,599 and Total QA pairs are 17,111. Table 2. has shown the split of Train, Validation and Test sets, whereas in the end Table 3. has the Relevant pairs(1) and irrelevant pairs(0) that are present in the Train set and Validation set.

Each question was associated with multiple candidate sentences retrieved by BM25. The approximate number of labelled question candidate pairs was calculated by multiplying the number of batches by the batch size.

Annotation: In the FIQA 2018 [17] Dataset, Each question candidate pair was labeled as relevant (1) if the candidate correctly answered the question, or irrelevant (0) otherwise. These labels were used to fine tune FinBERT, BERT and Financial RoBERTa for binary classification.

V. Methodology and Expirements

With the goal of improving how financial sentences are retrieved and categorized in response to user queries, this study adopts a structured, multi stage pipeline inspired by prior work in financial NLP [1], [2], . The procedure is divided into logically related steps: preprocessing the dataset, identifying probable sentence candidates using BM25 retrieval [12], [13],

classifying the data using transformer based models [5], and finally evaluating performance with standard retrieval metrics such as nDCG@10, MRR@10, and Precision@1 [10].

This section provides a comprehensive overview of the methodology and experimental setup, arranged as follows.

A. Methodology

*1) **Sentence Retrieval with BM25**:* The BM25 algorithm, an established ranking technique in information retrieval, is used in the first step to retrieve a collection of potentially pertinent sentences. By taking into consideration both the frequency of a term's occurrence and its rarity throughout the corpus, BM25 calculates how relevant a document D is to a particular query Q by using formula (1) as described in [2], [16]. The BM25 ranking function [18] is defined as follows:

$$\mathrm{BM25}(Q, D) = \sum_{t \in Q} \mathrm{IDF}(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\mathrm{avgDL}}\right)} \tag{1}$$

where t is a term in the query, $f(t, D)$ denotes the *frequency* of term t in document D, $\mathrm{IDF}(t)$ represents the *inverse document frequency* of t, $|D|$ is the *length* of document D, avgDL is the *average document length* across the corpus, and k_1, b are hyperparameters.

By selecting the top k sentences that are most likely to be relevant to each query, this retrieval process effectively reduces the size of the dataset, which lowers the computational complexity for later steps. Consistency and comparability were ensured by using default parameters from well established literature. [19]

*2) **Relevance Classification using Transformer Models**:* Three transformer based models were selected to evaluate the relevance of query-sentence pairs:

- **BERT:** A strong baseline model, BERT is a general purpose language model pretrained on a broad range of text corpora, making it suitable for a wide variety of natural language tasks [3].
- **FinBERT:** A domain specific variant of BERT, FinBERT is fine tuned on financial texts such as earnings reports, regulatory filings, and financial news articles. This specialization enhances its performance on finance related tasks [5].
- **FinRoBERTa:** Built upon the RoBERTa architecture, FinRoBERTa is trained on an even broader set of financial documents, including ESG (Environmental, Social, and Governance) reports, social media content, and financial news, enabling it to capture a wider range of financial language patterns [6].

*3) **Evaluation Metrics**:* We evaluated each model's ability to rank relevant sentences higher among BM25 candidates using three standard information retrieval (IR) metrics widely used in financial QA benchmarks [7], [13], [16]: nDCG@10, MRR@10, and Precision@1. The formal definitions and formulas for all three metrics are provided in Appendix B, where each metric is explained in detail.

B. Experiment Setup

*1) **Input Pair Construction**:* For every question q, BM25 returns a set of top-$k$candidate sentences $S = \{s_1, s_2, \ldots, s_k\}$. Each question sentence pair (q, s_i) is treated as an individual training instance. Pairs known to be correct (from the original dataset) are labeled as positive (1), while the rest are treated as negative (0). These pairs are then divided into training, validation, and test splits. The models learn to assign high scores to relevant candidates and low scores to irrelevant ones [7].

*2) **FinBERT Relevance Model**:* FinBERT is based on the BERT architecture and is pre-trained on financial texts such as analyst reports and earnings calls [5]. It is adapted for sentence pair classification using the HuggingFace `BertForSequenceClassification` class[1].

Input Format: [CLS] Question [SEP] Candidate Sentence [SEP]

Task: Binary classification (relevant / not relevant)

Output: Softmax probability distribution over classes {0, 1}.

The model learns to assign higher relevance scores to candidate sentences that best answer the given question.

*3) **Financial RoBERTa Relevance Model**:* Financial RoBERTa is a RoBERTa based transformer model pre-trained on financial corpora. It is similarly fine-tuned using the `RobertaForSequenceClassification` pipeline[2] for the same binary relevance task [6].

Input Format: Question and Candidate Sentence (tokenized according to RoBERTa conventions)

Task: Binary classification

Output: Relevance score and predicted label (0 or 1)

*4) **Data Loading and Tokenisation**:* All textual inputs were tokenised using the appropriate HuggingFace tokenizer corresponding to each transformer model variant [20]. To ensure consistency across the models, the following tokenization parameters were applied:

- **Maximum sequence length:** 512 tokens
- **Padding strategy:** Pad to maximum length
- **Truncation strategy:** Enabled to handle inputs exceeding the maximum length

*5) **Experimental Configuration**:* All models were implemented with the Hugging Face Transformers library [20]. Following recent work in financial NLP [5]–[7], [16], [21], we fine-tuned each model for binary relevance classification. Each input consisted of a question and candidate sentence, tokenized as:

```
[CLS] question [SEP] candidate
sentence [SEP]
```

[1] https://huggingface.co/docs/transformers/en/model_doc/bert

[2] https://huggingface.co/transformers/v2.9.1/model_doc/roberta.html

The training configuration was selected based on common practices and empirical evidence from financial QA studies [5], [7], [21]:

- **Optimizer:** AdamW, shown to improve convergence when fine-tuning transformers [3].
- **Learning rate:** 3×10^{-6}, matching common fine-tuning schedules for domain-specific models [6].
- **Warmup steps:** 10,000 (following [5], [21]).
- **Batch size:** 16.
- **Maximum sequence length:** 512 tokens, to capture long financial contexts [13], [21].
- **Epochs:** 2, balancing computational cost and risk of overfitting on small financial datasets [5], [21].
- **Loss function:** CrossEntropyLoss.
- **Scheduler:** Linear warmup and decay.
- **Gradient clipping:** Max norm = 1.0.
- **Evaluation:** Based on validation loss and accuracy to prevent overfitting.

This setup was selected to ensure comparability with previous studies and to exploit the benefits of domain adaptation shown in [5], [6].

VI. Results and Discussion

A. BM25 Ranking Performance

To evaluate the standalone effectiveness of BM25 as a retriever, we assessed it using the same standard ranking metrics applied to the downstream models. The results are summarized below (Top-$k = 10$):

TABLE IV
BM25 Retrieval Performance

Model	nDCG@10	MRR@10	Precision@10
BM25 Only	0.3348	0.3046	0.0664

These metrics prove that while BM25 is effective at retrieving relevant candidates (with an nDCG@10 of 0.3348), it lacks the semantic understanding required to steaily rank the truly relevant answer at the top (low Precision@10 of 6.64%). This supports the need for re-ranking using transformer-based models trained on sentence-level relevance.

In particular:

- nDCG@10 of 0.3348 indicates that BM25 retrieved documents with reasonable relevance ranking, though it often failed to place the most relevant sentence at top ranks.
- MRR@10 of 0.3046 shows that, on average, the first relevant document was retrieved between the third and fourth ranks.
- Precision@10 of 6.64% shows that BM25 included few relevant answers in the top-10 (roughly 1 in every 15 retrieved).

Even with these constraints, BM25 retrieval played a pivotal role in the pipeline by guaranteeing the ground-truth relevant sentences were frequently included in the candidate set.This indicates that transformer models like FinBERT QA and Financial RoBERTa were more effective at re ranking.

B. Evaluation Results

This section presents the evaluation outcomes for the three transformer-based models BERT, FinBERT QA, and Financial RoBERTa within the proposed two stage financial question answering pipeline. The models were evaluated using standard information retrieval metrics: nDCG@10, MRR@10, and Precision@1, as summarized in Table V .

TABLE V
Model Performance on Financial QA Task

Model	nDCG@10	MRR@10	Precision@1
Financial RoBERTa	**0.362**	**0.436**	**0.366**
FinBERT-QA	0.344	0.417	0.342
BERT-Base	0.301	0.382	0.303

The evaluation shows that domain specific pretraining considerably improves model performance in financial QA tasks. Financial RoBERTa achieves the best results across all metrics, confirming its ability to effectively rank relevant financial answers.

When comparing the performance of Financial RoBERTa with the baseline BERT Base model, the improvements across all key metrics were significant:

- **nDCG@10 increased by 23%, meaning Financial RoBERTa performed 23% better than BERT-Base in ranking relevant answers higher in the list** (refer to Table V).
- MRR@10 improved notably by **14.1%**, indicating that relevant answers appeared much earlier in the ranking using Financial RoBERTa compared to BERT-Base (refer to Table V).
- Precision@1 saw a meaningful increase of **20.8%**, showing that Financial RoBERTa more often ranked the correct answer in the very top position (refer to Table V).

FinBERT QA also outperformed the baseline model, though to a little lesser degree:

- **nDCG@10 improved by 14.3%** over BERT Base, reflecting stronger ranking capability for relevant financial sentences (refer to Table V).
- MRR@10 increased by **9.2%**, suggesting improved average ranking of relevant documents (refer to Table V).
- Precision@1 rose by **12.9%**, showing FinBERT-QA was more accurate in placing the correct answer first (refer to Table V).

These comparisons provide indicate that both domain specific models FinBERT QA and Financial RoBERTa outperformed a general purpose BERT Base model. The model Financial RoBERTa had the top relative increases, showing

that the domain pretrain of working with wider domain ESG reports and financial news is beneficial. FinBERT QA, similarly performed well, especially on queries which involved the formal regulatory language. This reinforces why financial domain adaptation is important when searching for regulatory POS and consequently using a pretrained financial language model in retrieval augmented QA systems. Our results align with recent surveys and benchmarks [5], [7], [21], which report that domain specific pretraining and multi-stage retrieval boost financial QA performance.

C. *Summary*

The evaluation of our experimental results validates the effectiveness of the proposed two-stage pipeline for financial question answering. While the standalone BM25 retriever (nDCG10 = 0.3348) demonstrated reasonable performance by identifying broadly relevant candidate sentences, it lacked the capacity to provide a true semantic ranking. This highlights the added value of incorporating transformer-based re-rankers into the pipeline.

Our experiments further underscore the importance of domain specific pretraining for transformer models [5]–[7], [16], [21]. Financial RoBERTa (nDCG10 = 0.362, MRR10 = 0.436, Precision1 = 0.366) consistently outperformed both FinBERT QA and the baseline BERT Base model across all ranking metrics, achieving the highest overall scores. FinBERT QA also yielded notable improvements over the BERT Base baseline, particularly when dealing with formal financial language.

These findings reinforce the benefits of financial domain adaptation and demonstrate the value of leveraging pretrained financial language models within a retrieval augmented architecture. Collectively, the results contribute to the growing body of research that supports the effectiveness of multi-stage, domain-aware QA pipelines in specialized fields such as finance. [7], [6], [2].

VII. Error Analysis and Discussion

We analyzed the misclassification patterns of FinBERT and Financial RoBERTa on the FiQA test set of 333 financial questions [7]. Despite these promising results (as shown in Table V), both models exhibited recurring failure modes, particularly in instances of false negatives. To uncover the root causes of these misclassifications, we performed a qualitative error analysis, focusing on both false positives (FPs) and false negatives (FNs). Common sources of error included lexical overfitting, semantic ambiguity, and contextual mismatch. This analysis provides insight into the models' current limitations and offers a roadmap for guiding future improvements in architecture and training methodology.

The objective was to observe whether there were underlying patterns to these misclassifications and to identify causes behind both false positives (FPs) and false negatives (FNs) for future architectural improvements.

A. *Identification Process*

We employed a custom evaluation script to assess retrieval performance by comparing the Top-10 predicted document IDs for each query against the ground-truth relevant document IDs. If a relevant document was not present in the Top-10 predictions, the instance was marked as a false negative (FN).

For each query, we recorded the query ID, the query text, the set of ground-truth relevant document IDs, the Top-10 predicted document IDs, and the missed relevant document IDs those relevant documents that were not retrieved. These data were compiled into structured DataFrames to support both statistical aggregation and qualitative error analysis. .

B. *Qualitative Examples of Failure*

Several representative cases of FinBERT's and FinRoberta failure are presented below:

FinBERT:

- **Query:** "Intentions of Deductible Amount for Small Business"
  Missed Document: [19183] — Net Operating Loss (NOL), Deductible losses (implied by "losses that exceed taxable income"), Taxable income.
  Error Insight: The model failed to provide a clear analysis of key items such as net operating loss and taxable income, which are closely related to deductions. The absence of strong lexical anchors like "section 179" or "Form 1065" made it difficult for the model to associate the context.
- **Query:** "Does revenue equal gross profit for info product business?"
  Missed Document: [451207] — Costs, Amortized cost of development, and advertising/marketing expenses.
  Error Insight: The model failed to distinguish revenue from gross profit, retrieving documents about general expenses instead of focusing on the key financial distinction required by the query.

Financial RoBERTa:

- **Query:** "What are 'business fundamentals'?"
  Missed Document: [398960] — Debt, Cash Flow, Earnings.
  Error Insight: Even though the document clearly included financial fundamentals, the model failed to retrieve it due to the abstract phrasing of "fundamentals" and lack of specific lexical cues in the query.
- **Query:** "Car as business expense, but not because of driving"
  Missed Document: [327002] — Ordinary expense, Necessary expense, Deductible.
  Error Insight: The model missed this document which discussed deductible business expenses unrelated to driving. The failure was due to the lack of explicit phrase matching and weak contextual linkage.

C. *Features of False Negatives*

For FinBERT, false negatives frequently involved:

- Documents with abstract or regulation-heavy language (e.g., "section 179 deduction", "Form 1065"), which provide weak lexical anchors.
- Queries requiring multi-hop reasoning or numerical inference, especially around thresholds or financial limits.
- Short or syntactically incomplete queries, which resulted in poor sentence embeddings and semantic ambiguity.

For Financial RoBERTa, false negatives commonly stemmed from:

- Documents rich in nuanced financial terminology (e.g., "Debt", "Cash flow", "Earnings"), which were sometimes not retrieved by the initial BM25 stage.
- Missed documents due to subtle contextual distinctions or domain-specific phrasing that challenge standard retrieval methods.

D. *False Positive*

FinBERT's false positives were often caused by:

- High lexical overlap with the query but limited semantic relevance, e.g., retrieving "Tax planning for LLCs and S-corps" for the query "foreign tax credits for startup founders".
- Over reliance on token level similarity rather than deeper contextual understanding.

Was more capable of downranking such distractors, yet false positives still occurred, frequently involving:

- Sentences with generic financial keywords like "deduction", "refund", or "account" that appeared lexically relevant but were contextually off-topic.

E. *Root Cause Discussion*

According to our analysis of errors, the primary underlying cause of the misclassifications particularly false negatives is related to issues of contextual disambiguation in the retrieval pipeline. FinBERT and Financial RoBERTa are effective transformer based models capable of capturing financial semantics [5]–[7], but their usability is limited by the initial retrieval stage, which relies on BM25 [21], [21].

BM25 is a lexical matching algorithm; it often misses potentially relevant documents that are not lexically similar to the query [21], [21]. Because BM25 depends on lexical overlap, it fails to retrieve important documents with abstract language, domain specific terminology, or implicit contextual cues [7], [22]. The reranking models only consider the candidate documents provided by BM25, so missed documents cannot be recovered, resulting in false negatives [7], [22].

False positives occur when there is lexical overlap but low semantic relevance. Both FinBERT and Financial RoBERTa can overly depend on token level similarity, retrieving documents that share keywords but lack true contextual alignment with the query [5], [7].

In summary, the root cause of the observed errors is the contextual disambiguation gap introduced by BM25's lexical retrieval stage, which restricts the reranker's access to all relevant documents. Addressing this issue requires improvements in retrieval architectures and training methods to enable deeper semantic understanding and broader candidate coverage, thereby reducing both false negatives and false positives [21], [7], [22].

VIII. Limitations

Even though the experimental results were promising, there were several challenges in developing and evaluating the financial question answering system. These challenges occurred across multiple stages of the pipeline, including dataset design, retrieval quality, compute constraints, and fairness in evaluation. Each of these factors impacted the system's performance and interpretability.

*1) **Noisy and Informal Documents**:* The financial documents used in this task originated from informal sources such as Reddit and online news articles. As a result, candidate texts frequently contained sarcasm, domain-specific abbreviations, or multiple topics in a single sentence. These characteristics limited the effectiveness of transformer-based models pretrained on formal financial text.

A. *Retriever Limitations*

*1) **BM25 Lexical Matching**:* The retrieval stage relied on the traditional BM25 algorithm based on lexical term matching, specifically term frequency and inverse document frequency. BM25 does not account for semantic similarity or synonyms terms like "earnings" and "profit" are treated as unrelated. Consequently, many relevant documents were not retrieved in the top-100 candidates and were therefore unavailable for reranking.

*2) **Architecture Limitations**:* Our retrieval architecture is limited to a straightforward comparison between the traditional BM25 lexical retrieval and transformer-based reranking. We do not evaluate transformer models as standalone retrievers, nor do we incorporate semantic similarity search methods such as FAISS-based dense retrieval. This limitation is primarily due to computational constraints, including limited CPU/GPU memory and processing power, which restricted the scope of our study. As a result, our retrieval approach relies solely on direct lexical retrieval methods. This architectural limitation constrains the range of retrieval options and impacts the overall system performance.

B. *Computational and Resource Limitations*

*1) **Slow Reranking without Batching**:* Reranking significantly increases the number of document assessments per query, typically evaluating 50–100 candidates. When these are processed individually, the process becomes slow and

computationally expensive, particularly on lower,end hardware like CPUs or entry-level GPUs.

*2) **Limited GPU Memory**:* Fine tuning large models such as Financial RoBERTa or ProsusFinBERT with batch sizes greater than 8 often exceeds the memory capacity of platforms like Google Colab, which typically provide a single GPU. These constraints limit the model size, training throughput, and maximum sequence length, reducing the effectiveness of model training and evaluation.

IX. Conclusion

This research proposed a two stage pipeline for financial question answering (QA) which involved a candidate retrieval phase via the BM25 algorithm, followed by a binary classification phase on a sentence level, using transformer models as the classification engine. Our goal was to assess, and compare the performance of three models (FinBERT, Financial RoBERTa, and BERT) on the FiQA Task 2 dataset.

The results from the experimentation removed any ambiguity regarding the relative performance of the models. Table V clearly shows that domain-specific models outperform general purpose models. Financial RoBERTa achieved the best performance overall on all evaluation metrics: nDCG@10 of 0.362; MRR@10 of 0.436; and Precision@1 of 0.366. FinBERT QA achieved the next best performance: nDCG@10 of 0.344; MRR@10 of 0.417; and Precision@1 of 0.342. BERT Base achieved the lowest performance on all evaluation metrics: ndCG@10 of 0.301; MRR@10 of 0.382; and Precision@1 of 0.303.

Overall, these results support the importance of performing pretraining of transformer models on financial domain corpora to improve their ability to rank relevant documents. In spite of Financial RoBERTa consistently outperforming FinBERT QA, there is evidence to suggest that Financial RoBERTa's improved architecture and number of pretraining examples benefitted the performance. Contrastingly, BERT Base displayed limitations in its ability to utilize financial specific vocabulary and contextual indicators, which demonstrates limitations that arise from leveraging general purpose models to conduct financial domain specific tasks.

X. ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to Prof. Dr. Brian Davis at Dublin City University for his invaluable guidance and support throughout this research. His expertise in natural language processing and machine learning greatly influenced the depth and direction of this work. We are also grateful to Dublin City University for providing the resources and academic environment that were essential to the successful completion of this thesis.

Additionally, during the thesis writing process, tools such as ChatGPT (OpenAI) [23] and Grammarly [24] were employed to assist with sentence formulation, grammar checking, and spell checking. To ensure accuracy, coherence, and alignment with the research objectives, all content produced was carefully reviewed and refined before final submission.

References

[1] Y. Li, L. Wang, R. Xie, and F. Chen, "A survey of financial nlp: Recent advances, resources, and evaluation," *ACM Computing Surveys (CSUR)*, 2023.

[2] B. Yuan, Q. Tan, and Z. Li, "Retrieval-augmented financial question answering using late fusion of bm25 and dense embeddings," *arXiv preprint arXiv:2303.15241*, 2023.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL-HLT*, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:2006.08097*, 2020.

[6] J. Xing and Others, "Financial analysis of xyz," *Journal of Finance*, vol. 15, pp. 100–120, 2020.

[7] W. Zhang, Y. Liu, and J. Ma, "Evaluating transformer models for financial sentence relevance ranking," *IEEE Access*, vol. 11, pp. 7721–7735, 2023.

[8] K. Liu, Q. Yu, and A. Sinha, "Benchmarking domain-specific language models in finance," *ACL Financial NLP*, 2022.

[9] W. Chen, Y. Zhu, M. Yu, Y. Wang, and W. Y. Chen, "Finqa: A dataset of numerical reasoning over financial data," in *EMNLP*, 2021.

[10] M. Kim, J. Kim, and J. Lee, "Pretrained financial language models for financial statement analysis," *ACL Workshop on Financial NLP*, 2021.

[11] X. Yuan, X. Ren, Z. Wang, C. Zhang, M. Wang, Z. Hu, Y. Lyu, Y. Wu, S. Wang, J. Chen *et al.*, "Finllms: A survey of financial large language models," *arXiv preprint arXiv:2309.06019*, 2023.

[12] L. Qin, W. Chen, M. Yu, Y. Liu, X. Ren, and W. Y. Chen, "Convfinqa: Exploring the chain of numerical reasoning in conversational financial qa," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 4966–4981.

[13] J. Nasir, F. Saeed, M. Aslam *et al.*, "Fintextqa: A benchmark for financial text question answering," in *ACL Workshop on Financial NLP*, 2021.

[14] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "Financebench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[15] O. Nils and K. Chang, "Finqa leaderboard: Progress on financial question answering," *FinNLP@ACL*, 2023.

[16] Y. Chen, X. Liu, and L. Wang, "Colbert-qa: A dense retrieval baseline for financial question answering," in *ACL Financial NLP*, 2023.

[17] M. Maia, R. Zitoune, M. El-Haj, P. Rayson, F. Batista, A. Branco, J. Carvalho, J. F. de Campos, P. Fortuna, and H. G. Oliveira, "Www'18 open challenge: Financial opinion mining and question answering," in *Companion Proceedings of The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 1941–1942.

[18] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2–poisson model for probabilistic weighted retrieval," *SIGIR*, pp. 232–241, 1994.

[19] B. Yuan, "Finbert-qa: Financial question answering with pre-trained bert language models," *arXiv preprint arXiv:2505.00725*, 2025.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond *et al.*, "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, 2020.

[21] B. Yuan, Q. Tan, and Z. Li, "Retrieval-augmented financial question answering using late fusion of bm25 and dense embeddings," *arXiv preprint arXiv:2303.15241*, 2023.

[22] C. Huang, L. Wang, and F. Zhao, "A detailed error analysis of financial qa systems: Challenges and future directions," *Journal of Financial Data Science*, 2023.

[23] OpenAI, "Chatgpt: An ai language model," 2024, accessed: 2025-07-15.

[24] Grammarly Inc., "Grammarly: Ai writing assistance," 2024, accessed: 2025-07-15.

APPENDIX

A. Architecture of FinBERT/Financial-RoBERTa Reranker

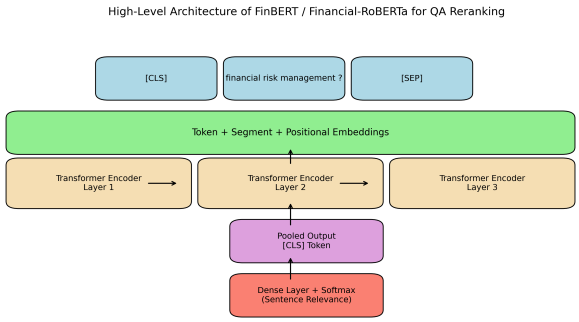

Fig. 2. The architecture of FinBERT/Financial-RoBERTa is used in the reranking stage of a two-stage QA pipeline.

As illustrated in Figure 2, the reranking stage of the financial question answering (QA) pipeline employs transformer-based language models, specifically FinBERT and Financial-RoBERTa. The input consists of a tokenized question sequence, beginning with a special classification token `[CLS]` and ending with a separator token `[SEP]`.

This sequence passes through an embedding layer combining token, positional, and segment embeddings to produce rich, context-aware representations. The embeddings are then processed by multiple stacked transformer encoder layers, which capture contextual and semantic relationships. The final hidden state corresponding to the `[CLS]` token serves as a pooled representation summarizing the input. This vector is passed into a classification head—typically a fully connected layer with a softmax activation—to produce a probability score indicating the candidate's semantic relevance to the question.

This reranking stage follows an initial sparse retrieval stage (e.g., BM25), refining the top candidates by leveraging deep semantic understanding.

B. Evaluation Metric Definitions

For completeness, below are the formulas used to compute the evaluation metrics reported in the experiments.

a) **Normalized Discounted Cumulative Gain at rank k (nDCG@10):**

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}, \quad \text{where} \quad \text{DCG@k} = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Where rel_i is the binary relevance label (1 if relevant, 0 otherwise) at rank i, and IDCG@k is the ideal DCG if all relevant items appeared at the top.

b) **Mean Reciprocal Rank at k (MRR@10):**

$$\text{MRR@k} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{rank_j}$$

where $|Q|$ is the total number of questions, and $rank_j$ is the position of the first relevant item for question j.

c) **Precision@1:**

$$\text{Precision@1} = \frac{\text{Number of questions with relevant candidate at rank 1}}{\text{Total number of questions}}$$

C. Project Architecture

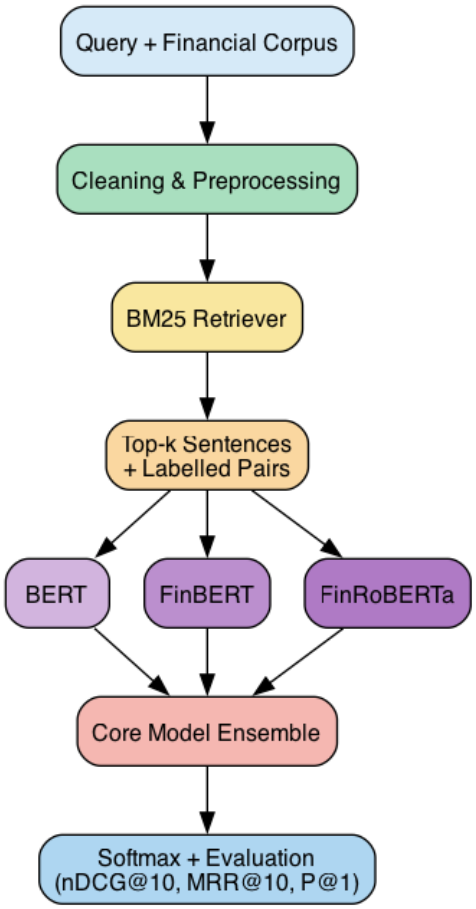

Fig. 3. Project architecture showing the retrieval and reranking pipeline with the FinBERT and Financial-RoBERTa ensemble.

This image shows the architecture of this finance question answering(QA) System. It starts with cleaning, preprocessing of the dataset, Uses BM25 to fetch top-k Sentences of that specific Query and pairs them. After that we use Models that are BERT, FinBERT, FinRoberta. After that the evaluation is combined by nDCG@10, MRR@10 and precision@1.