

FinTruthQA: A Benchmark for AI-Driven Financial Disclosure Quality Assessment in Investor–Firm Interactions

Peilin Zhou^{1*}, Ziyue Xu^{2*}, Xinyu Shi³, Jiageng Wu⁴, Yikang Jiang⁵,
Dading Chong⁶, Wang Dong², Jun Chen², Bin Ke⁵, Jie Yang^{4†}

¹New York University Abu Dhabi

²Zhejiang University

³Southwest University of Finance and Economics

⁴Harvard University

⁵National University of Singapore

⁶Peking University

{zhoupalin, zyxu1224, jieynlp}@gmail.com

Abstract

Accurate and transparent financial information disclosure is essential for market efficiency, investor decision-making, and corporate governance. Chinese stock exchanges’ investor interactive platforms provide a widely used channel through which listed firms respond to investor concerns, yet these responses are often limited or non-substantive, making disclosure quality difficult to assess at scale. To address this challenge, we introduce FinTruthQA, to our knowledge the first benchmark for AI-driven assessment of financial disclosure quality in investor–firm interactions. FinTruthQA comprises 6,000 real-world financial Q&A entries, each manually annotated based on four key evaluation criteria: *question identification*, *question relevance*, *answer readability*, and *answer relevance*. We benchmark statistical machine learning models, pre-trained language models and their fine-tuned variants, as well as large language models (LLMs), on FinTruthQA. Experiments show that existing models achieve strong performance on *question identification* and *question relevance* (F1>95%), but remain substantially weaker on *answer readability* (Micro F1~88%) and especially *answer relevance* (Micro F1~80%), highlighting the nontrivial difficulty of fine-grained disclosure quality assessment. Domain- and task-adapted pre-trained language models consistently outperform general-purpose models and LLM-based prompting on the most challenging settings. These findings position FinTruthQA as a practical foundation for AI-driven disclosure monitoring in capital markets, with value for regulatory oversight, investor protection, and disclosure governance in real-world financial settings.

Code — <https://github.com/bethxx99/FinTruthQA>

Datasets — <https://github.com/bethxx99/FinTruthQA>

*These authors contributed equally.

†Corresponding author. This work was conducted while the author was affiliated with Zhejiang University.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Introduction

In capital markets, listed firms increasingly communicate with investors through online interactive platforms, where the quality of their responses can shape investor confidence, disclosure governance, and market efficiency. The online Q&A platforms established by the Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE) provide a widely used channel through which listed firms respond to investor concerns in near real time. Unlike conventional disclosures (e.g., mandatory reports or conference calls), which follow relatively standardized formats, these platforms capture more spontaneous and fine-grained investor–firm interactions, making them a valuable setting for evaluating how effectively firms communicate substantive information to the market.

Despite their practical importance, these platforms present substantial challenges for large-scale analysis because investor–firm interactions are both massive in volume and highly variable in quality. Traditional disclosure measures primarily capture quantity rather than quality (Leuz and Wysocki 2016). Although natural language processing (NLP) has successfully automated the analysis of financial reports (Smailović et al. 2017) and news sentiment (Van de Kauter, Breesch, and Hoste 2015), there is currently no standard benchmark for systematically evaluating AI models on financial disclosure quality in real-world investor–firm interactions, especially for tasks requiring contextual and relevance-sensitive judgment.

To address this gap, we present FinTruthQA, a benchmark dataset for AI-driven financial disclosure quality assessment on Chinese stock exchanges’ interactive investor platforms. It is organized around four key evaluation criteria, namely *question identification*, *question relevance*, *answer readability*, and *answer relevance*, which together capture important dimensions of disclosure quality in investor–firm interactions and define the four benchmark tasks in our study. Using FinTruthQA, we systematically evaluate a broad range of NLP techniques across these tasks. Our results support

data-driven disclosure monitoring, regulatory oversight, and investor protection in real-world financial settings.

Using FinTruthQA, we obtain the following key findings:

- Among the four tasks, *answer readability* and especially *answer relevance* are substantially more difficult than *question identification* and *question relevance*, highlighting the nontrivial nature of fine-grained disclosure quality assessment.
- On the most challenging task, *answer relevance*, financial-domain PLMs, particularly FinBERT with continued pretraining, achieve the best performance, reaching a Micro F1 of 80.04%.
- Current general-purpose LLMs still underperform domain-adapted supervised models on fine-grained relevance judgment, indicating that stronger domain grounding is needed for trustworthy deployment in financial applications.

Related Work

Disclosure Measurement

Measuring disclosure quality has been a persistent challenge in empirical accounting research (Leuz and Wysocki 2016). Comprehensive measures traditionally quantify disclosure frequency, such as counting reports or conference calls (Lee and Zhong 2022). Other approaches rely on analyst surveys—which are often limited to specific, large U.S. firms (Nagar, Nanda, and Wysocki 2003)—or utilize keyword-based techniques to analyze the tone of traditional, structured disclosures like earnings forecasts (Bochkay et al. 2023; Frankel, Mayew, and Sun 2010). Crucially, these traditional methods primarily capture information quantity or basic sentiment, making it difficult to evaluate the substantive quality of dynamic, unstructured text.

Financial NLP

Due to the technical complexity of financial text (Chang et al. 2016), extensive research has developed domain-adapted pre-trained models (Yang, Uy, and Huang 2020; Liu et al. 2021). Early applications targeted sentiment analysis (Wan et al. 2021), distress prediction (Hajek and Munk 2024), and named entity recognition (Zhang et al. 2024). Recently, the focus has shifted toward question-answering (QA) involving textual and tabular numerical reasoning (Zhao et al. 2022; Li, Ye, and Zhao 2022), alongside the deployment of large language models (LLMs). Several domain-specific LLMs—such as FinTral (Bhatia et al. 2024), FinBERT2 (Xu et al. 2025), and Open-FinLLMs (Xie et al. 2024b)—have demonstrated robust capabilities across multimodal and text-based financial tasks. Concurrently, new datasets and benchmarks like CFBenchmark (Lei et al. 2023) and FinBen (Xie et al. 2024a) evaluate LLM performance on financial assistance, corporate QA, and data annotation (Aguda et al. 2024). However, despite these advancements, no prior dataset is dedicated to the automated quality assessment of financial Q&A interactions.

Dataset Construction

Data Collection

The data for this study was obtained from the interactive platforms for communication between investors and listed companies established by the Shanghai Stock Exchange (Shangzhen E-Hudong <https://sns.sseinfo.com/>) and the Shenzhen Stock Exchange (Hudongyi <https://irm.cninfo.com.cn/>). These two stock exchange sites cover almost all listed companies in China. These Q&A interactions are authentic, generated by human users and company representatives, and are regulated by the China Securities Regulatory Commission, ensuring the accuracy of disclosures. The companies involved are legally obligated to ensure the accuracy of their disclosures, and failure to do so could result in legal consequences. This regulatory oversight ensures that the data we used reflects genuine interactions. We collected over 900,000 Q&A entries, among which we randomly selected 6,000 samples for manual annotation, with 3,000 entries from each exchange’s platform. The data collection involved using web scraping techniques to extract the Q&A entries from the official platforms. The data were then manually reviewed and processed to ensure quality and consistency. The process was overseen by a PhD student from the Business School of National University of Singapore. The ethics statement, including considerations of general ethical conduct and potential negative societal impacts, can be found in our GitHub repository. We also give several examples in our GitHub repository to demonstrate the wide range of investor inquiries covered in our dataset.

Data Annotation

We focus on four key information disclosure quality evaluation criteria¹: *question identification*, *question relevance*, *answer relevance*, and *answer readability*. These criteria are widely recognized as crucial indicators of information quality and are important for investors to consider when evaluating Q&A information. Specifically, the criteria in FinTruthQA are defined as follows:

1. **Question identification:** This criterion assesses whether the question from an investor is a genuine question, as opposed to a suggestion or complaint. Each question is assigned a binary label: *Positive*, indicating it is a real question, or *Negative*, indicating it is not. Data entries with questions labeled as *Negative* are excluded from further evaluation under the remaining criteria.
2. **Question relevance:** This criterion evaluates whether a question is pertinent to the investment topic, labeling it as *Positive* if relevant and *Negative* if not. It assesses whether the question is related to the company’s shareholders, financial indicators, industry conditions, or other investment-related topics.

¹In this paper, the terms “criteria” and “tasks” are used interchangeably. “Criteria” refers to the specific standards used to evaluate the quality of the Q&A pairs, while “tasks” refers to the individual evaluation processes performed based on these criteria, particularly in the experimental sections.

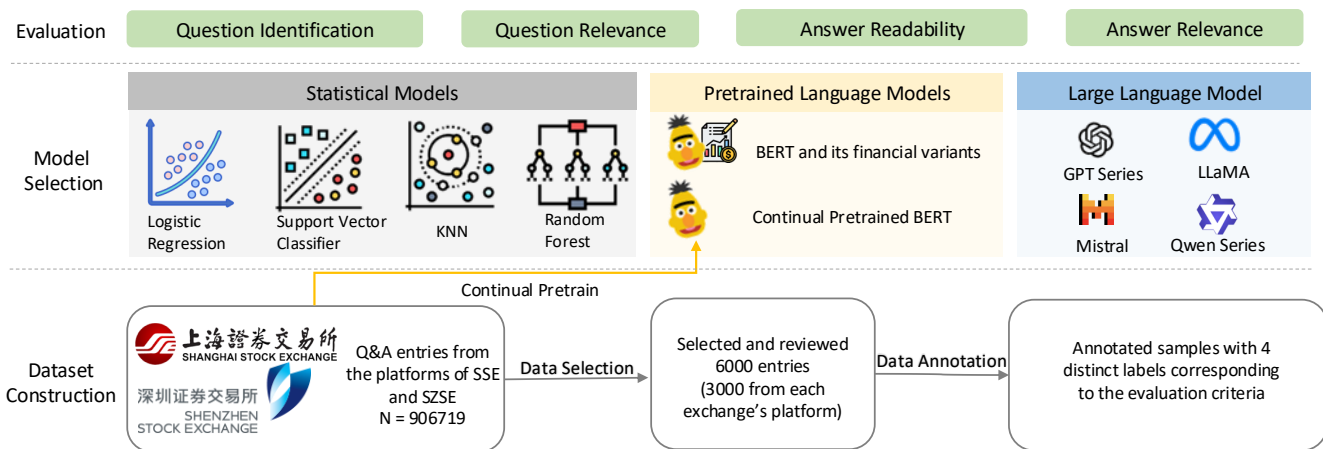


Figure 1: Overview of the FinTruthQA construction and evaluation pipeline. The process begins with dataset construction, where 906,719 Q&A entries were collected from the official platforms of the Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE). From these, a subset of 6,000 entries (3,000 per platform) was selected and manually annotated based on four evaluation criteria. For model evaluation, we include traditional machine learning models, pre-trained language models, and large language models (LLMs), enabling comparative analysis of performance across different modelling strategies.

3. **Answer readability:** This criterion evaluates how easily an answer can be understood and whether it is well-written. It considers multiple dimensions, including structural coherence, clarity of information, stylistic appropriateness, and grammatical correctness, ensuring a comprehensive evaluation of the answer’s quality. The readability score ranges from 1 to 4, with 4 indicating the highest level of readability—clear, concise, and free of ambiguity. A score of 3 indicates moderate readability with some issues, while a score of 2 reflects significant readability challenges. A score of 1 suggests the answer is unreadable, with major ambiguities, incoherent sentences, or content unrelated to the question.
4. **Answer relevance:** This criterion evaluates whether an answer is relevant and informative in response to the question. It assesses whether the answer provides sufficient detail and explanation to address the question. Answers are classified into three levels: Level 1 indicates the answer is completely unrelated to the question (none of the questions is answered); Level 2 indicates the answer is partially related (some questions are answered or it is difficult to judge); and Level 3 indicates the answer fully responds to the question (all questions are answered).

The importance of these criteria lies in their collective ability to enhance the quality and reliability of information on investor communication platforms. The annotation process was carried out in three phases: pre-annotation, formal annotation, and data verification and integration. In the pre-annotation phase, we recruited six students from the Accounting Department at Zhejiang University and conducted five rounds of annotation. After the last round, the inter-annotator agreement (IAA) rate was calculated, and the four students with the highest average IAA rates (85.52%)—two undergraduates and two master’s students—were selected

for the formal annotation phase. To ensure annotation quality, the four selected annotators were evenly divided into two groups, with each group annotating the same data instances. Discrepancies were first discussed and resolved within the group, with unresolved issues escalated to the project leader for final decisions. After the formal annotation, the project leader verified the data to ensure alignment with the latest annotation criteria and completeness, before integrating the data and summarising key information. Figure 2 illustrates an example of the annotation. A detailed description of the annotation process is listed in our GitHub repository.

Quality Control

For the *answer relevance* task, we observed that the raw annotations contained noise due to the subjective nature of this task. To address this, we introduced an additional quality control step to correct the raw annotations. We implemented an enhanced correction process inspired by Confident Learning principles (Northcutt, Jiang, and Chuang 2021). Sample data points demonstrating inaccuracies, along with implementation details, can be found in our GitHub repository.

Dataset Statistics

The complete annotated dataset comprises 6,000 samples labeled for *question identification*, *question relevance*, *answer relevance*, and *answer readability* tasks, with each data point having four distinct labels corresponding to these tasks. The dataset was randomly divided into training, validation, and testing sets in a ratio of 4:1:1. The label distributions for each task are detailed in Table 1. The provided statistics reflect the data after undergoing the quality control process. Note that our later analysis focuses exclusively on real questions. Only for the first task (*question identification*), we used the full set of 6,000 samples. For the remaining three tasks, we removed non-question samples, as they could skew the



QA Pair	Question	Answer
	 Investor: Hello, could you please tell me the number of shareholders in the company as of February 10? Thank you!	 Company: Hello, as of February 10, the number of shareholders in the company is 65,409.
Quality Evaluation	Criteria 1: Question identification Label: Positive. <i>Whether the question is genuine question, as opposed to a suggestion or complaint? Positive: Genuine question; False: Not a question</i>	Criteria 3: Answer readability Label: 4. <i>Whether an answer is easy to understand and well-written? Label set: [1, 2, 3, 4] Higher values indicate better readability.</i>
	Criteria 2: Question relevance Label: Positive. <i>Whether a question is relevant to the investment topic? Positive: Relevant; False: Irrelevant.</i>	Criteria 4: Answer relevance. Label: 3. <i>Whether the answer provides a relevant and informative response? Label set: [1, 2, 3] Higher values indicate stronger relevance.</i>

Figure 2: Example of a Q&A interaction between a retail investor and a listed company representative drawn from the Fin-TruthQA dataset (translated to English), along with the annotated labels for each of the four quality assessment criteria: question identification (whether the input is a genuine question), question relevance (whether the question is related to investment-related topics), answer readability (how clear and well-written the answer is, on a 1–4 scale), and answer relevance (how well the answer addresses the question, on a 1–3 scale).

Task	Label	# Train	# Valid	# Test	# Label	Freq.	Total
Q-ID	Negative	327	90	81	498	8.30%	6000
	Positive	3673	910	919	5502	91.70%	
Q-Rel	Negative	714	6	7	27	0.49%	5502
	Positive	3654	911	910	5475	99.51%	
A-Read	1	106	190	198	1094	19.88%	5502
	2	26	3	6	35	0.64%	
	3	68	19	20	107	1.94%	
	4	2868	705	693	4266	77.54%	
A-Rel	1	731	194	206	1131	20.63%	5502
	2	533	136	130	779	14.21%	
	3	2404	587	581	3572	65.16%	

Table 1: Dataset statistics for all tasks. Q-ID: question identification; Q-Rel: question relevance; A-Read: answer readability; A-Rel: answer relevance.

model’s understanding and performance, especially for tasks like *question relevance*, where the context of a real question is crucial for accurate assessment. It is worth noticing that the *question relevance* labels are highly imbalanced, with only 6 and 7 negative labels in the validation and test sets, respectively. We retained this task as the imbalance reflects the real-world data distribution, and the concept of *question relevance* is important in financial information disclosure. Figure 3 shows the length distribution of questions and answers, measured in characters, for all 6,000 QA pairs.

Benchmarks

Model Selection

To fully evaluate the performance of different methods in FinTruthQA, we benchmarked several models that cover Machine Learning (ML)-based, Pre-trained Language Model (PLM)-based, PLM with continued Pre-Training

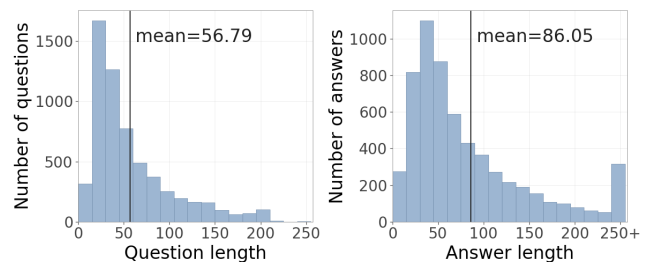


Figure 3: Length distributions of our dataset (in characters).

(CPT)-based, and LLM-based approaches. Given the distinct nature and varying difficulty of the tasks, we applied tailored approaches to each accordingly.

ML-based. We modeled these tasks as text classification problems and used various advanced machine learning models, including Logistic Regression (LR), Support Vector Machine (SVM), K-nearest neighbor (KNN), and Random Forest (RF). For semantic feature extraction, we first utilized the Jieba toolkit² for Chinese word segmentation and then extracted weighted text representations using TF-IDF. To achieve optimal performance, we used cross-validation and GridSearch to rigorously explore the best hyperparameters. The best-performing classifiers and hyperparameters varied across tasks. Therefore, we performed task-specific hyperparameter searches for each task individually.

PLM-based. In our experiments, we found that the tasks of *question identification* and *question relevance* were relatively straightforward, with BERT achieving excellent performance on these tasks. Therefore, our focus shifted to exploring the effectiveness of PLM-based methods on the more challenging tasks of *answer relevance* and *answer readability*. Specifically, we investigated a range of BERT-

²<https://github.com/fxsjy/jieba>

Category	Model	Question identification (%)				Question relevance(%)			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ML-based	TF-IDF + LR	93.70±0.86	94.11±0.75	99.35±0.18	96.66±0.48	99.42±0.13	99.42±0.13	100.00±0.00	99.71±0.07
	TF-IDF + RF	93.70±0.65	94.02±0.63	99.46±0.18	96.66±0.37	99.42±0.13	99.42±0.13	99.93±0.10	99.67±0.08
	TF-IDF + SVM	93.60±0.73	93.78±0.76	99.64±0.13	96.62±0.40	99.42±0.13	99.42±0.13	100.00±0.00	99.71±0.07
	TF-IDF + KNN	92.67±1.11	93.27±0.98	99.16±0.37	96.12±0.61	99.42±0.13	99.42±0.13	100.00±0.00	99.71±0.07
PLM-based	BERT	96.89±0.55	97.64±0.22	99.02±0.41	98.32±0.28	99.67±0.11	99.72±0.12	99.96±0.05	99.85±0.06

Table 2: Model evaluation results of machine learning and pre-trained language models on the *question identification* and *question relevance* task. Models include TF-IDF-based classifiers (logistic regression, random forest, SVM, KNN) and a pre-trained BERT model. The PLM-based BERT achieves the highest overall performance across both tasks.

based models, including those specifically trained on financial text data, such as financial news, research reports, and financial encyclopedia entries (*FinBERT*,³ *Mengzi-fin* (Zhang et al. 2021), and *DKPLM* (Zhang et al. 2022)). Furthermore, we examined whether fine-tuned PLMs on specific tasks, such as NLI (Natural Language Inference) and QA (Question Answering), would offer any benefits for these tasks (*SBERT-nli* (Reimers and Gurevych 2019) and *RoBERTa-extractive-qa* (Liu et al. 2019)). Finally, we also examined the performance of a large BERT model (*BERT (Large)*).⁴ Notably, unless otherwise specified, all PLMs employed in this study are the base and Chinese versions. Links for these PLMs can be found in our GitHub repository.

PLM with CPT-based. General domain PLMs, trained on broad texts, often lack the specialized knowledge required to handle the nuanced language and terminology in finance. This limitation is especially pronounced in tasks like *answer relevance*, where understanding domain-specific jargon and abbreviations is essential for accurate predictions. To bridge this gap, we curated a substantial volume of domain-specific texts and conducted continued pre-training on general PLMs. Specifically, we collected over 900,000 Q&A pairs from the SSE and SZSE platforms by web scraping. The top-performing models from previous experiments were then selected for continued pre-training using masked language modeling (MLM). This pre-training on a large-scale, domain-specific dataset enhances the PLM’s ability to understand financial contexts, capture more specialized information, and achieve better performance in downstream tasks.

LLM-based. We further evaluate a diverse set of advanced LLMs on the most challenging *answer relevance* task, including the OpenAI series (GPT-4 (Achiam et al. 2023), GPT-4o, GPT-5, and O3), GPT-OSS 120B, LLaMA-3.1 (Dubey et al. 2024), the Qwen series (Qwen2 (Yang et al. 2024) and Qwen3-Next variants), and Mistral (Jiang et al. 2023). Due to their in-context learning ability, LLMs can perform this task in a zero-shot manner, without requiring specific fine-tuning. Specifically, we provided the LLMs with task descriptions and evaluation criteria, instructing them to directly output a relevance score (*Direct*). Moreover, we incorporated the Chain of Thought (CoT) (Wei et al. 2022) technique, prompting the models to provide specific

reasoning alongside the score (*CoT*). Full details are provided in our GitHub repository.

Settings and Metrics

Implementation Details. For the ML-based models, we utilized the standard `scikit-learn` library. For all PLM-based methods, we used `PyTorch` and HuggingFace’s framework for downstream fine-tuning and evaluation. Detailed hyperparameter settings for our PLM-based experiments can be found in our GitHub repository. We conducted experiments comparing Q&A pairs with and without the company name as a prefix and found negligible differences, so we ultimately used data without the prefix. The tasks of *question identification* and *question relevance* only required encoding the questions as input, while *answer readability* and *answer relevance* required encoding both the question and the answer. The sequence started with a special classification token [CLS], and the question and answer were separated by the [SEP] token . During continued pre-training, we added two special tokens, [Q] and [A], into the vocabulary and prefixed them to the question and answer, respectively. This helps the PLMs better identify the components of each sentence. The continued pre-training was performed based on the UER framework (Zhao et al. 2019). All tests of LLMs were conducted using the transformer framework, except for the GPT-4 experiments, which were conducted by calling OpenAI’s API with version GPT-4-0613. Specifically, we used Llama-3.1 with 8 billion parameters, and both Qwen2 and Mistral with 7 billion parameters.

Metrics. For *question identification* and *question relevance*, which are binary classification tasks, we used accuracy, precision, recall, and F1-score as evaluation metrics. For *answer readability* and *answer relevance*, which are multi-class classification tasks, we calculated both micro and macro F1-scores, along with the Quadratic Weighted Kappa (QWK) (Dong, Zhang, and Yang 2017), which is well-suited for ordinal classification by penalizing misclassifications based on their distance from the true label. Each model was run in five times, and the mean and variance of the results are reported.

Results and Analysis

In this section, we present the evaluation results and discuss the performance of the benchmarked models. To provide a

³<https://github.com/valuesimplex/FinBERT>

⁴<https://huggingface.co/yechen/bert-large-chinese>

Model	F1 (Mi)	F1 (Ma)	F1 (W)	QWK
<i>ML-based</i>				
TF-IDF + LR	81.72 ±0.27	34.91 ±0.22	78.90 ±0.27	41.20 ±0.80
TF-IDF + RF	80.48±0.70	32.81±1.06	76.85±1.03	33.79±3.87
TF-IDF + SVM	81.64±0.37	33.92±0.38	78.15±0.40	37.98±1.48
TF-IDF + KNN	78.99±0.34	31.07±1.23	74.95±0.84	27.31±4.28
<i>PLM-based</i>				
BERT	86.04±0.44	41.97±0.67	84.89±0.44	63.52±1.43
RoBERTa	86.26±0.62	42.02±1.92	85.08±0.89	64.00±0.96
FinBERT	87.57 ±0.39	41.34±0.45	85.98 ±0.69	66.53 ±1.83
Mengzi-fin	86.99±0.80	41.93±1.42	85.77±1.20	66.33±1.81
DKPLM	87.24±0.76	41.90±1.17	85.76±0.72	65.78±0.63
SBERT-nli	85.31±1.05	39.28±0.81	83.48±1.38	58.19±2.77
RoBERTa-extractive-qa	87.21±0.52	41.22±0.11	85.75±0.64	65.97±0.80
BERT (Large)	87.24±1.03	42.17±1.66	85.78±1.30	65.92±2.74
RoBERTa (Large)	86.48±1.48	45.51 ±4.59	85.37±1.41	64.58±2.11

Table 3: Model evaluation results on the *answer readability* task. F1 (Mi), F1 (Ma), and F1 (W) denote micro-, macro-, and weighted-F1 scores, respectively. PLM-based models consistently outperform ML-based baselines across most metrics, with FinBERT achieving the best overall performance in F1 and QWK.

comprehensive understanding, our analysis is organized systematically across the four aforementioned tasks, highlighting the strengths and limitations of different modeling approaches.

Task 1: Question Identification In this task, we benchmarked several ML-based and PLM-based models, with their performances summarized in Table 2. The ML-based models showed consistent results with minor variations. Both TF-IDF + LR and TF-IDF + RF reached an accuracy of 93.70% and an F1 score of 96.66, with closely aligned precision and recall. TF-IDF + SVM exhibited slightly lower accuracy but achieved the highest recall among the ML models, demonstrating its strength in identifying relevant instances. TF-IDF + KNN, while performing well with an accuracy of 92.67%, had the lowest precision and F1 score in this group. On the other hand, the PLM-based BERT model outperformed all ML-based methods, achieving the highest accuracy, precision, and F1 score. This result highlights the PLM model’s ability to capture deeper contextual nuances, which traditional frequency-based methods like TF-IDF struggle to achieve.

Task 2: Question Relevance Table 2 presents the evaluation results for the question relevance task. The PLM-based model outperformed the ML-based models on all metrics except recall, achieving an accuracy of 99.67%, an F1 score of 99.85%, precision of 99.72%, and recall of 99.96%. This demonstrates its ability to effectively identify relevant questions while maintaining a strong balance in correctly classifying rare negative samples. In contrast, the TF-IDF-based models showed nearly identical performance across metrics, primarily predicting all samples as positive. Given the highly imbalanced label distribution (910 positive vs. 7 negative

Model	F1 (Mi)	F1 (Ma)	F1 (W)	QWK
<i>ML-based</i>				
TF-IDF + LR	69.61±0.49	49.74 ±0.53	65.07 ±0.60	38.16 ±2.13
TF-IDF + RF	69.97±0.34	44.79±0.84	63.07±0.76	36.63±1.10
TF-IDF + SVM	70.08 ±0.83	44.42±0.43	62.93±1.07	36.78±1.65
TF-IDF + KNN	66.66±1.03	38.63±2.05	57.87±1.42	23.22±3.36
<i>PLM-based</i>				
BERT	77.90±0.98	67.64±0.94	76.79±0.75	63.95±0.85
RoBERTa	77.10±1.34	65.95±2.64	75.77±1.35	62.22±2.46
FinBERT	78.19±0.96	68.18±0.79	77.25±0.81	65.49±0.54
Mengzi-fin	78.04±0.72	68.96 ±2.42	77.44 ±0.82	66.31 ±1.10
DKPLM	76.92±0.57	63.00±3.72	74.65±1.15	62.82±1.78
SBERT-nli	74.52±0.36	61.18±2.54	72.38±1.07	53.88±1.12
RoBERTa-extractive-qa	78.33 ±0.36	67.02±1.48	77.12±0.69	64.66±1.44
BERT (Large)	78.01±0.78	65.94±1.62	76.16±0.56	63.53±0.90
RoBERTa (Large)	78.30±0.94	68.50±2.09	77.40±1.54	64.52±1.56
<i>PLM with CPT-based</i>				
FinBERT	80.04 ±0.23	70.58 ±0.90	79.02 ±0.18	67.89 ±1.26
RoBERTa-extractive-qa	78.34±0.34	68.01±1.71	77.45±0.74	66.76±0.58
BERT (Large)	78.12±0.65	65.15±3.97	76.02±1.16	64.82±2.07
RoBERTa (Large)	74.37±1.78	61.40±1.28	72.46±0.73	53.91±3.42

Table 4: Model evaluation results on the *answer relevance* task. PLM-based models substantially outperform ML-based baselines, while CPT-enhanced PLMs further improve performance, with CPT-FinBERT achieving the best overall results across all evaluation metrics.

samples in the test set), these models achieve high recall by predicting positives but struggle to correctly identify the few negative instances, reflecting the challenge posed by the skewed data.

Task 3: Answer Readability The evaluation results for the task *answer readability* are shown in Table 3. Among the ML-based models, the LR classifier exhibited the highest performance, achieving an F1 Micro score of 81.72% and a QWK of 41.20%. While LR provided reasonable performance, the relatively low F1 Macro score highlights its difficulties in handling class imbalance effectively. In contrast, the PLM-based models demonstrated significant improvements over the ML-based models. For instance, BERT achieved an F1 Micro score of 86.04% and a QWK of 63.52%, showcasing its superior ability to capture the complexities of readability assessment through richer contextual understanding.

Detailed Evaluation of FinBERT’s Performance. FinBERT outperformed all other models with an F1 Micro score of 87.57%, an F1 Macro score of 41.34%, an F1 Weighted score of 85.98%, and the highest QWK of 66.53%. Its pre-training on financial text provided a distinct advantage for this domain-specific task, enabling it to achieve the highest accuracy and alignment with human judgments. In further analysis, we find that the model performs well in recognizing answers with the highest readability, but it struggles to accurately differentiate between the lower readability levels, highlighting the need for improvement in capturing finer dis-

Model	F1(Mi)	F1(Ma)	F1(W)	QWK
GPT-4	67.39 /59.76	63.11 /58.45	70.14 /64.28	59.06 /57.51
GPT-4o	54.53/54.96	53.02/54.73	59.76/58.25	53.34/47.97
GPT-5	60.52/59.32	59.45/58.32	62.66/61.10	51.91/49.86
O3	65.43/ 66.41	62.88/ 64.29	68.37/ 69.24	57.93/ 60.12
GPT-OSS 120B	62.27/61.94	60.55/59.94	65.35/64.81	55.76/53.87
Llama-3.1 8B	53.00/46.46	41.59/40.82	56.13/51.57	43.24/39.56
Qwen2 7B	60.52/59.00	31.33/38.97	52.83/57.36	23.68/42.35
Mistral 7B	54.53/59.32	30.40/23.03	50.99/60.20	29.57/38.90
Qwen3-Next 80B Inst.	59.98/60.09	56.71/57.09	64.51/61.38	56.23/43.20
Qwen3-Next 80B Think.	59.11/58.89	56.26/56.62	60.50/59.82	44.83/42.98

Table 5: Zero-shot evaluation results of LLMs on the *answer relevance* task. Each cell reports Direct/CoT prompting results.

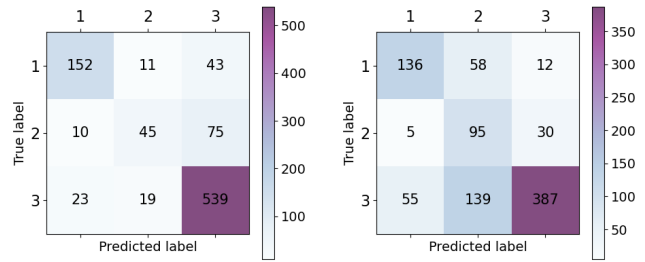
tinctions in readability.

Task 4: Answer Relevance The evaluation results for the *answer relevance* task are detailed in Table 4 for supervised models and Table 5 for zero-shot LLMs. As the most challenging criterion, assessing nuanced answer relevance revealed distinct performance variations among the different modeling strategies. Our key observations are summarized as follows:

PLM-based Methods Consistently Outperform ML-based ones. As shown in Table 4, all PLM-based methods can achieve better performance than ML-based methods across all metrics. For instance, RoBERTa-extractive-qa achieved the highest Micro F1 score of 78.33%, while Mengzi-fin excelled with the highest Macro F1 score of 68.96% and a strong QWK of 66.31%. Their superior performance can be attributed to pre-training on task-specific and domain-specific data: RoBERTa-extractive-qa was pre-trained on QA datasets, which enhanced its ability to comprehend and evaluate answer relevance, while Mengzi-fin was pre-trained on financial corpora, enabling it to better capture domain-specific nuances in the financial context.

Impact of Continued Pre-training on PLM-based Methods. We selected four PLM-based models and further pre-trained them using 900,000 Q&A pairs. As shown in Table 4, we observed that only FinBERT demonstrated significant performance improvements, achieving the highest Micro F1 score of 80.04% and a QWK of 67.89%, indicating its enhanced ability to assess answer relevance in financial contexts. In contrast, the other three models (RoBERTa-extractive-qa, BERT Large, and RoBERTa Large) did not exhibit consistent improvements. This suggests that domain alignment plays a crucial role in the effectiveness of continued pre-training, as FinBERT’s domain-specific pre-training on financial texts provided a strong foundation, whereas the other models may have experienced a mismatch between their learned representations and the specific requirements of the task, resulting in less notable gains.

LLMs Still Struggle with Fine-Grained Relevance Judgments in Chinese Financial QA. As shown in Table 5, we evaluated advanced LLMs on the answer relevance task un-



(a) FinBERT with continued pretraining for *answer relevance* task. (b) GPT-4 (direct score) for *answer relevance* task.

Figure 4: Confusion matrices for *answer relevance* predictions from FinBERT with continual pretraining and GPT-4 (direct scoring).

der both direct scoring and CoT prompting. GPT-4 achieves the best performance in the direct setting, while O3 performs best under CoT. However, the effect of CoT is highly model-dependent: it substantially improves O3, but does not benefit GPT-4 or GPT-5. This suggests that explicit reasoning is not universally helpful for fine-grained relevance assessment in Chinese financial disclosures. More broadly, even the strongest LLMs remain behind the best supervised PLM-based models, indicating that this task remains challenging for general-purpose LLMs. We attribute this difficulty to the need for domain-specific financial knowledge, the subtle distinction between adjacent relevance labels, and the prevalence of indirect or evasive disclosure patterns in real investor–company interactions.

We also experimented with Chinese financial LLMs, including CFGPT2 (Li et al. 2023) and DISC-FinLLM (Chen et al. 2023). However, these models frequently failed to follow the scoring instructions and often produced repetitive or malformed outputs, making their predictions unreliable. We therefore exclude them from the main comparison.

Error Analysis. To better understand the gap between supervised PLMs and LLMs, we further analyze their confusion patterns. Figure 4a presents the confusion matrix of the best-performing supervised model, FinBERT with continued pretraining. The model reliably identifies answers that are fully relevant to the question (label 3), but often confuses partially irrelevant answers (label 2) with fully relevant ones, suggesting that the boundary between these two categories is especially subtle. By contrast, the confusion matrix of GPT-4 shown in Figure 4b reveals more frequent confusion between completely irrelevant answers (label 1) and partially relevant ones (label 2), together with a tendency to over-predict the middle category. This pattern indicates weaker calibration in borderline cases, especially when an answer contains superficially related financial terminology without directly addressing the question. Additional qualitative analysis of representative failure cases can be found in our GitHub repository.

Conclusion

This study introduces FinTruthQA, a benchmark designed to evaluate advanced NLP techniques for the automatic quality assessment of information disclosure in financial Q&A data. We constructed a dataset of 6,000 real-world financial Q&A entries, each manually annotated based on four key accounting-oriented evaluation criteria. Experiments revealed the effectiveness of pre-trained BERT models in tackling these tasks. In particular, we observed that models pre-trained on financial corpora, as well as those designed for related tasks such as QA, outperformed other variants of BERT and GPT-4. Although existing NLP models showed reasonable performance on this benchmark, there is still significant room for improvement, especially in the task of *answer relevance*, which calls for more robust and business-aware NLP models. Beyond academic evaluation, FinTruthQA provides a practical foundation for AI-driven disclosure monitoring in capital markets, supporting disclosure governance, investor-facing communication, and data-driven decision-making in real-world financial settings. Our benchmark is useful not only for post-hoc regulatory assessment but also for real-time applications that promote financial transparency, reduce information asymmetry, and help auditors, financial analysts, regulators, and corporate communication teams identify inadequate or potentially misleading responses in a timely manner.

Limitations and Future Work. The Chinese cultural and regulatory context of SSE and SZSE may limit generalizability to other markets, and the manual annotation process inherently involves a degree of subjectivity. Furthermore, reliance on standard quantitative metrics may neglect interpretability, the *question relevance* dataset is highly imbalanced (though reflective of real-world distributions), and computational constraints restricted our LLM evaluations to zero-shot settings. Future work will explore fine-tuned LLMs, retrieval-augmented methods, and human-centered evaluation protocols, while also validating these methodologies across diverse global regulatory and business contexts.

References

- Achiam, J.; Adler, S.; Agarwal, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.
- Aguda, T. D.; Siddagangappa, S.; Kochkina, E.; et al. 2024. Large Language Models as Financial Data Annotators: A Study on Effectiveness and Efficiency. In *Proc. LREC-COLING*, 10124–10145.
- Bhatia, G.; Nagoudi, E. M. B.; Cavusoglu, H.; et al. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv:2402.10986*.
- Bochkay, K.; Brown, S. V.; Leone, A. J.; et al. 2023. Textual analysis in accounting: What’s next? *Contemp. Account. Res.*, 40(2): 765–805.
- Chang, C. Y.; Zhang, Y.; Teng, Z.; et al. 2016. Measuring the information content of financial news. In *Proc. COLING*, 3216–3225.
- Chen, W.; Wang, Q.; Long, Z.; et al. 2023. DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning. *arXiv:2310.15205*.
- Dong, F.; Zhang, Y.; and Yang, J. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proc. CoNLL*, 153–162.
- Dubey, A.; Jauhri, A.; Pandey, A.; et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.
- Frankel, R.; Mayew, W. J.; and Sun, Y. 2010. Do pennies matter? Investor relations consequences of small negative earnings surprises. *Rev. Account. Stud.*, 15: 220–242.
- Hajek, P.; and Munk, M. 2024. Corporate financial distress prediction using the risk-related information content of annual reports. *Inf. Process. Manag.*, 61(5).
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; et al. 2023. Mistral 7B. *arXiv:2310.06825*.
- Lee, C. M.; and Zhong, Q. 2022. Shall we talk? The role of information investor platforms in corporate communication. *J. Account. Econ.*, 74(2-3): 101524.
- Lei, Y.; Li, J.; Cheng, D.; et al. 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv:2311.05812*.
- Leuz, C.; and Wysocki, P. D. 2016. The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. *J. Account. Res.*, 54(2): 525–622.
- Li, C.; Ye, W.; and Zhao, Y. 2022. Finmath: Injecting a tree-structured solver for question answering over financial reports. In *Proc. LREC*, 6147–6152.
- Li, J.; Bian, Y.; Wang, G.; et al. 2023. CFGPT: Chinese financial assistant with large language model. *arXiv:2309.10654*.
- Liu, Y.; Ott, M.; Goyal, N.; et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.
- Liu, Z.; Huang, D.; Huang, K.; et al. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proc. IJCAI*, 4513–4519.
- Nagar, V.; Nanda, D.; and Wysocki, P. 2003. Discretionary disclosure and stock-based incentives. *J. Account. Econ.*, 34(1-3): 283–309.
- Northcutt, C.; Jiang, L.; and Chuang, I. 2021. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, 70: 1373–1411.
- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proc. EMNLP*.
- Smailović, J.; Žnidaršič, M.; Valentinčič, A.; et al. 2017. Automatic analysis of annual financial reports: A case study. *Computación y Sistemas*, 21(4): 809–818.
- Van de Kauter, M.; Breesch, D.; and Hoste, V. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Syst. Appl.*, 42(11): 4999–5010.
- Wan, X.; Yang, J.; Marinov, S.; et al. 2021. Sentiment correlation in financial news networks and associated market movements. *Sci. Rep.*, 11(1): 3062.

Wei, J.; Wang, X.; Schuurmans, D.; et al. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv:2201.11903*.

Xie, Q.; Han, W.; Chen, Z.; et al. 2024a. The FinBen: A Holistic Financial Benchmark for Large Language Models. *arXiv:2402.05263*.

Xie, Q.; Li, D.; Xiao, M.; et al. 2024b. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications. *CoRR*.

Xu, X.; Wen, F.; Chu, B.; et al. 2025. FinBERT2: A specialized bidirectional encoder for bridging the gap in finance-specific deployment of large language models. In *Proc. KDD*, 5117–5128.

Yang, A.; Yang, B.; Hui, B.; et al. 2024. Qwen2 technical report. *arXiv:2407.10671*.

Yang, Y.; Uy, M. C. S.; and Huang, A. 2020. Finbert: A pretrained language model for financial communications. *arXiv:2006.08097*.

Zhang, H.; Dang, Y.; Zhang, Y.; et al. 2024. Chinese nested entity recognition method for the finance domain based on heterogeneous graph network. *Inf. Process. Manag.*, 61(5).

Zhang, T.; Wang, C.; Hu, N.; et al. 2022. DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In *Proc. AAAI*, volume 36, 11703–11711.

Zhang, Z.; Zhang, H.; Chen, K.; et al. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv:2110.06696*.

Zhao, Y.; Li, Y.; Li, C.; et al. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. *Proc. ACL*.

Zhao, Z.; Chen, H.; Zhang, J.; et al. 2019. UER: An Open-Source Toolkit for Pre-training Models. *Proc. EMNLP*, 241–246.