FinTextQA: A Dataset for Long-form Financial Question Answering

Jian Chen[1,2] **Peilin Zhou**[2] **Yining Hua**[3]
Yingxin Loh[1] **Kehui Chen**[1] **Ziyuan Li**[1] **Bing Zhu**[1*] **Junwei Liang**[2*]
[1]HSBC Lab [2]Hong Kong University of Science and Technology (Guangzhou)
[3]Harvard University
{alex.j.chen, bing1.zhu}@hsbc.com, {jchen524, pzhou460}@connect.hkust-gz.edu.cn,
yininghua@g.harvard.edu, junweiliang@hkust-gz.edu.cn

Abstract

Accurate evaluation of financial question-answering (QA) systems necessitates a comprehensive dataset encompassing diverse question types and contexts. However, current financial QA datasets lack scope diversity and question complexity. This work introduces *FinTextQA*, a novel dataset for long-form question answering (LFQA) in finance. *FinTextQA* comprises 1,262 high-quality, source-attributed QA pairs extracted and selected from finance textbooks and government agency websites.Moreover, we developed a Retrieval-Augmented Generation (RAG)-based LFQA system, comprising an embedder, retriever, reranker, and generator. A multi-faceted evaluation approach, including human ranking, automatic metrics, and GPT-4 scoring, was employed to benchmark the performance of different LFQA system configurations under heightened noisy conditions. The results indicate that: (1) Among all compared generators, Baichuan2-7B competes closely with GPT-3.5-turbo in accuracy score; (2) The most effective system configuration on our dataset involved setting the embedder, retriever, reranker, and generator as Ada2, Automated Merged Retrieval, Bge-Reranker-Base, and Baichuan2-7B, respectively; (3) models are less susceptible to noise after the length of contexts reaching a specific threshold. The dataset is publicly available [1].

1 Introduction

The growing demand for financial data analysis and management has led to the expansion of artificial intelligence (AI)-driven question-answering (QA) systems (Wu et al., 2023). These systems not only enhance customer service but also assist in risk management and personalized stock recommendations (Yuan et al., 2021, 2023c,b). The intricate nature of financial data, with its domain-specific terminologies, concepts, and the inherent uncertainty of the market and decision-making processes, demands a deep understanding of the financial domain to generate accurate and informative responses (Confalonieri et al., 2021). In this context, long-form question answering (LFQA) scenarios become particularly relevant as they require models to demonstrate a broad spectrum of sophisticated skills, including information retrieval, summarization, data analysis, comprehension, and reasoning (Fan et al., 2019).

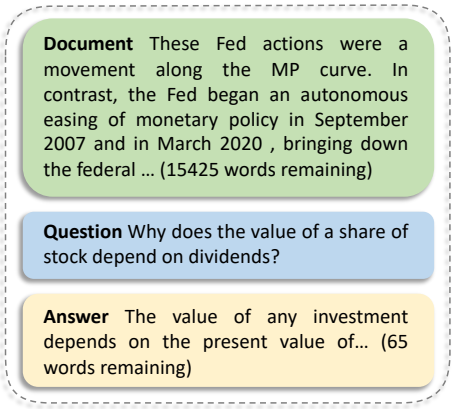

Figure 1: An LFQA sample in *FinTextQA*. Models are expected to generate paragraph-length answers when given questions and documents.

In the general domain, there are several LFQA datasets available, including ELI5 (Fan et al., 2019), WikiHowQA (Bolotova-Baranova et al., 2023) and WebCPM (Qin et al., 2023). However, it is important to note that there is currently no LFQA dataset specifically tailored for the finance domain. Existing financial QA benchmarks often fall short in addressing question complexity and variety by primarily on sentiment analysis and numerical calculation, as comprehensive paragraph-length responses and relevant document retrievals are often required to answer intricate, open-domain questions (Han et al., 2023). To address these challenges, we intro-

* Co-corresponding Author
[1]https://huggingface.co/datasets/GPS-Lab/FinTextQA

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6025–6047

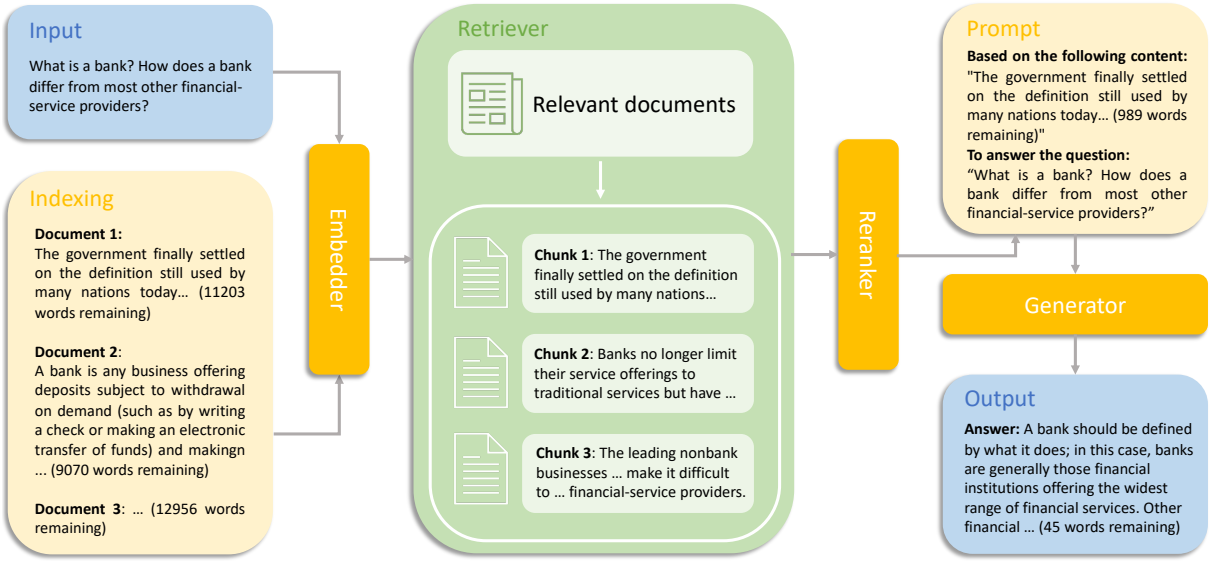

Figure 2: The workflow of our proposed RAG-based LFQA system. Embedder aims to encode documents and user's question into semantic vectors. Retriever retrieves relevant document chunks based on the encoded question. Reranker removes less-similar chunks. With a prompt which combines question and chunks, the generator finally output desired answer.

duce a new dataset, *FinTextQA*, which comprises LFQAs from finance-related textbooks and government agency websites to assess QA models on general finance and regulation or policy-related questions. *FinTextQA* consists of 1,262 high-quality, source-attributed question-answer pairs and associated document contexts. It contains six question types with an average text length of 19.7k words, curated from five rounds of human screening. This dataset is pioneering work in integrating financial regulations and policies into LFQA, challenging models with more demanding content.

In addition to introducing the dataset, we conduct comprehensive benchmarking of state-of-the-art (*sota*) models on *FinTextQA* to provide baselines for future research. Current LFQA systems frequently solely rely on fine-tuning pre-trained language models such as GPT-3.5-turbo, LLaMA2 (Touvron et al., 2023), Baichuan2 (Yang et al., 2023a), etc., which often fail to provide detailed explanations or effectively handling complicated finance questions (Yuan et al., 2023a). In response, we opt for the Retrieval-augmented generation (RAG) framework, as illustrated in Figure 2. By processing documents in multiple steps, RAG systems can pre-process and provide the most relevant information to LLMs, enhancing their performance and explanation capabilities (Guu et al., 2020).

We believe this work, by introducing the first LFQA financial dataset and conducting comprehensive benchmark experiments on the dataset, marks a milestone in advancing the comprehension of financial concepts and enhancing assistance in this field: *FinTextQA* offers a rich and rigorous framework for building and assessing the capabilities of general finance LFQA systems. Our experimental analysis not only highlights the efficacy of various model configurations but also underscores the critical need for enhancing current methodologies to improve both the precision and explicability of financial question-answering systems.

2 Related Work

2.1 Long-Form Question Answering (LFQA)

Compared to the conventional question-answering task (Bouziane et al., 2015; Mao et al., 2023; Yang et al., 2023b; Cao et al., 2021; Cheng et al., 2023d, 2024), the goal of LFQA is to generate comprehensive, paragraph-length responses by retrieving and assimilating relevant information from various sources (Fan et al., 2019). This poses a significant test for current Natural Language Processing (NLP) and Artificial Intelligence (AI) models, given their limited understanding and learning capacities (Thompson et al., 2020; Cao et al., 2022; Zhang et al., 2023; Cheng et al., 2023b,c; Cheng, 2021).

Several LFQA datasets are available in general domain, including ELI5 (Fan et al., 2019), WikiHowQA (Bolotova-Baranova et al., 2023), and We-

bCPM (Qin et al., 2023). In the financial domain, some QA datasets have been developed. However, none of them addresses LFQA. While these datasets like FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) address specific scopes such as numerical reasoning, they do not touch upon general LFQA tasks. In addition, FIQA (Maia et al., 2018) only provides short-context documents, which may not adequately represent real-life scenarios and have limited industry applicability. Although FinanceBench (Islam et al., 2023) does cover a wider scope, it only offers 150 open-source question-answer pairs, while the question complexity and answer do not satisfy real-life LFQA scenarios.

2.2 Retrieval-Augmented Generation (RAG)

RAG frameworks represent a significant advancement in LFQA, incorporating external knowledge sources and In-Context Learning (ICL) for efficient information retrieval and application. By combining diverse documents into comprehensive prompts, RAG enables language models to generate contextually informed responses without task-specific retraining (Lewis et al., 2020).

The evolution of RAG involves three principal stages: Naive RAG, Advanced RAG, and Modular RAG. Naive RAG offers improvements over traditional language models by providing cost-efficient indexing, retrieval, and generation, albeit with certain constraints. Advanced RAG addresses these limitations by integrating refined indexing and retrieval techniques, optimizing data handling, and introducing strategic Retrieval and post-retrieval processes. Its capabilities include fine-tuning domain-specific embedders, employing dynamic ones for improved context comprehension, and applying reranker and prompt compression during post-retrieval processes (Ilin, 2023). Modular RAG represents a further advancement from traditional NLP frameworks by introducing specialized modules for similarity-based retrieval, fine-tuning, and problem-solving. It also incorporates innovative modules like Search and Memory (Cheng et al., 2023a), Fusion (Rackauckas, 2024), and Routing to customize RAG for specific applications and improve search and retrieval operation (Gao et al., 2023). The ongoing evolution of RAG demonstrates its potential to revolutionize information retrieval and adaptability in language model systems within the fast-evolving field of computational linguistics. In this study, we choose to assess the effectiveness of Modular RAG with the rewrite, retrieve, re-rank, and read modules following previous work (Gao et al., 2023).

3 *The FinTextQA Dataset*

3.1 Data Sources

The data in *FinTextQA* are sourced from well-established financial literature and government agencies, such as expert-authored question-answer pairs from recognized finance textbooks: *Bank Management and Financial Services* (BMFS), *Fundamentals of Corporate Finance* (FCF), and *The Economics of Money, Banking, and Financial Markets* (EMBFM). Additionally, crucial information regarding financial regulations and policies is incorporated from esteemed websites such as the Hong Kong Monetary Authority (HKMA) [2], European Union (EU)[3], and the Federal Reserve (FR)[4]. Question types encompass various domains, spanning concept explanation and numerical calculation to comparative analysis and open-ended opinion-based queries.

3.2 Selection of Policy and Regulation Data

In textbooks, questions are typically straightforward, and evidence (i.e., citations) can be easily found within each chapter. However, in policies and regulations, some of the questions draw from multiple sources and may not directly align with the documents in our dataset. This poses a challenge for the model to provide accurate answers. Additionally, policy and regulation data often require deeper analytical thinking and interpretation, demanding a robust reasoning ability from the QA system. Given the complexity and importance of financial regulations and policies, we have implemented a thorough two-step verification process to ensure the relevance and accuracy of the QA pairs and the associated regulation and policy documents:

1. **Evidence identification**: Initially, annotators are tasked with locating relevant evidence (aka citations and references) for each question-answer pair within the dataset. Any questions that cannot be feasibly linked to a valid citation or reference were promptly excluded from consideration;

[2] https://www.hkma.gov.hk/eng
[3] https://european-union.europa.eu/index_en
[4] https://www.federalreserve.gov

Source	# of Document	# of Question
EU	1	12
FR	8	190
HKMA	6	38
BMFS	19	319
EMBFM	26	472
FCF	20	231

Table 1: Disribution of numbers of documents and questions from different sources.

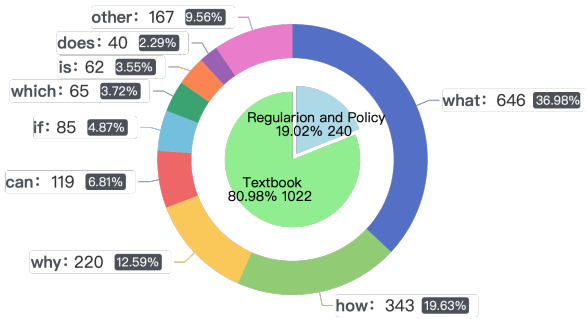

Figure 3: Distribution of data sources and interrogative words in *FinTextQA*.

2. **Relevance evaluation**: Another distinct group of annotators evaluates the coherence and connectedness between the question, context, and answer for each entry. Using a grading scale from 1 to 5, they ensure high standards of relevancy. Only entries with a score exceeding 2 across all three variables are included in the final dataset.

Initially, we collected 300 regulation and policy question-answer pairs with related document contexts. After careful data quality control, 240 pairs were retained. The data selection process resulted in a dataset demonstrating strong relevance among answer-context (3.91), question-answer (4.88), and question-context (4.54), indicating its high quality and dependability. Further details of human evaluation can be found in Appendix A.1.

3.3 Dataset Statistics

FinTextQA contains 1,262 QA pairs, with 1,022 pairs from finance textbooks, accounting for 80.98% of the dataset, and 240 pairs from policies and regulations, accounting for 19.02% of the dataset. We randomly split the dataset into training, validation, and test sets following a 7:1:2 ratio for model fine-tuning and evaluation. Table 1 presents data distribution across different sources.

Table 1 illustrates the distribution of these questions, representing various aspects of financial regulations and policies. The European Commission subset comprises 12 questions focused on transaction regulation and its interpretations. The Federal Reserve subset, containing over 190 questions, addresses topics such as banking regulations, monetary policy strategies, and international banking operations. The Hong Kong Monetary Authority subset contains 38 questions covering anti-money laundering, counter-terrorist financing ordinance, and credit card business regulations, etc.

FinTextQA consists mainly of compound questions, where each primary question includes 2-3 related sub-questions. This hierarchical format introduces more complexity for question understanding and reasoning. These sub-questions come in different forms, leading to a variety of interrogative words as illustrated in Figure 3. Our analysis shows that 36.98% of the questions start with "what", making it the most common starting word, followed by "how" at 19.63%, "why" at 12.59%, and "can" at 6.81%. The diversity in the types of interrogative words enriches the dataset, providing a more thorough test of large language models' ability to read and understand text.

3.4 Comparison to Existing Datasets

Table 2 shows a comparison of LFQA datasets, not limited to finance. *FinTextQA* stands out with an average question length of 28.5 words, answers of 75 words, and notably extended document contexts, averaging 19,779.5 words. These extensive contexts, segmented into chapters or sessions, are designed to enhance retrieval tasks. Furthermore, *FinTextQA* covers a broad scope, including multi-turn, numerical, finance domain, and open-ended questions. It contains the most complex questions and longest answers alongside the widest scope, as compared with other finance QA datasets. Further details of question types can be found in Appendix A.4.

4 Benchmarks on *FinTextQA*

4.1 RAG-based LFQA system

We employ the modular RAG as discussed in (Gao et al., 2023) and follow the guidelines outlined in LlamaIndex[5] to construct the RAG-based LFQA

[5] https://www.llamaindex.ai

Dataset	Average # of Words			Scope					
	Question	Document	Answer	Multi-turn	Comparative	Numerical	Domain	Open-minded	Cause and Effect
FIQA (Maia et al., 2018)	12.8	136.4	-				✓	✓	
TAT-QA (Zhu et al., 2021)	12.4	42.6	4.3			✓	✓		
FinQA (Chen et al., 2021)	16.6	628.1	1.1	✓	✓	✓	✓		
FinanceBench (Islam et al., 2023)	27.0	**65,615.6**	12.66			✓	✓	✓	
FinTextQA (ours)	**28.5**	19,779.5	**75**	✓	✓	✓	✓	✓	✓

Table 2: Comparison of various financial QA datasets. *FinTextQA* offers substantially longer questions and answers. Meanwhile, has a wider scope compared with other finance QA datasets.

Generator	Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
GPT-3.5-turbo	AMR	Ada2	LLMRerank	4.411	**0.346**	**0.134**	**0.224**	**0.062**
	AMR	Ember-v1	LLMRerank	4.365	0.341	0.130	0.221	0.060
	AMR	Ember-v1	Bge-Reranker-Base	4.439	0.339	0.131	0.221	0.062
Baichuan2-7B	AMR	Ada2	LLMRerank	4.578	0.340	0.124	0.219	0.057
	AMR	Ada2	Bge-Reranker-Base	**4.612**	0.338	0.123	0.217	0.054
	AMR	Ember-v1	Bge-Reranker-Base	4.513	0.333	0.120	0.215	0.053
Solar-10.7B	AMR	Ember-v1	Bge-Reranker-Base	4.348	0.329	0.119	0.205	0.052
	AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.310	0.329	0.118	0.205	0.051
	AMR	Ada2	Bge-Reranker-Base	4.378	0.327	0.119	0.204	0.051
Qwen-7B	AMR	Bge-Small-en-v1.5	LLMRerank	4.414	0.341	0.125	0.217	0.059
	AMR	Ada2	Bge-Reranker-Base	4.405	0.337	0.120	0.216	0.056
	AMR	Ember-v1	LLMRerank	4.432	0.339	0.121	0.215	0.056
LLaMA2-7B	SWR	Ada2	All-Mpnet-Base-v2	4.184	0.233	0.078	0.152	0.030
	AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.268	0.239	0.078	0.151	0.031
	AMR	Bge-Small-en-v1.5	LLMRerank	4.287	0.233	0.076	0.149	0.031
Gemini-Pro	AMR	Ember-v1	Bge-Reranker-Base	3.970	0.304	0.118	0.211	0.048
	AMR	Ember-v1	LLMRerank	3.990	0.306	0.119	0.211	0.052
	AMR	Bge-Small-en-v1.5	LLMRerank	3.989	0.303	0.119	0.210	0.051

Table 3: Systematic performance comparison of RAG-based LFQA system with different configurations.

Module	Model	Average Score
Embedder	Ada2	**4.586**
	Ember-v1	4.486
	Bge-Small-en-v1.5	4.455
	Gte-Large	4.261
Retriever	AMR	**4.492**
	SWR	4.466
	Vector Retrieval	4.358
Reranker	Bge-Reranker-Base	**4.489**
	LLMRerank	4.469
	All-Mpnet-Base-v2	4.383
Best-perfoming Configurations	AMR+Ada2+LLMRerank	**4.622**
	SWR+Ada2+Bge-Reranker-Base	4.620
	SWR+Ada2+All-Mpnet-Base-v2	4.620

Table 4: Performance comparison of embedders, retrievers, rerankers, and best-performing configuration. GPT-4 scores are generated regarding question-evidence relevance.

system. As shown in Figure 2, this LFQA system consists of four modules: embedder, retriever, reranker, and generator. The first three modules together serve to find relevant information (aka *evidence* or *citations*) from contexts. The last module synthesizes responses using the retrieved information. Each module can be implemented by different models, and the combinations of models for all modules constitutes the system's configurations. The selection of the models for each module is summarzied as follows:

The Embedder Module The role of the embedder module is to convert human language into a vector representation that can be understood and processed by computers. In our experiments, we adopt four popular embedding models that have achieved high rankings on the Hugging Face leaderboard, including (1) BAAI's Bge-small-en-v1.5 (Xiao and Liu, 2023), (2) NLPer's Gte-large (Li et al., 2023), (3) LLMRails' Ember-v1[6], and (4) OpenAI's Ada2[7].

The Retriever Module The retriever module forms the backbone of our experiment by searching and retrieving relevant context related to a given question. We explore three retriever methods, including Auto Merging Retriever (AMR) (Liu, 2023), (2) Sentence Window Retriever (SWR) (llamaindex, 2023), and a simple vector-based retriever approach. AMR organizes documents into a hierarchical tree system with parent nodes' contents distributed among child nodes. This enables users to determine the relevance of the parent node based on its child nodes' relevance to the query. SWR fetches context from a custom knowledge base by

[6] https://huggingface.co/llmrails/ember-v1

[7] https://platform.openai.com/docs/guides/embeddings

Model	# of Unanswered Questions		GPT-4 Score		Answer&Evidence Relevance		ROUGE-L		BLEU	
	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned
GPT-3.5-turbo	21	**0**	4.30	4.08	4.34	4.39	**0.21**	0.19	**0.05**	0.03
Baichuan2-7B	**0**	**0**	**4.50**	**4.51**	**4.73**	**4.73**	0.20	**0.20**	**0.05**	**0.04**
Qwen-7B	13	10	4.43	4.43	4.59	4.35	0.19	0.19	0.04	**0.04**
Solar-10.7B	13	12	4.38	4.38	4.50	4.50	0.19	0.18	0.04	**0.04**
LLaMA2-7B	**0**	**0**	4.14	4.27	4.22	4.28	0.10	0.10	0.02	0.02
Gemini-Pro	61	-	2.46	-	1.85	-	0.15	-	0.02	-

Table 5: Performance comparison of generators. # of Unanswered Questions is the number of "can not provide answer based on the content" generated by different generators. We use the same embedder, retriever, and reranker in this experiment. An example of unanswered questions is shown in Appendix Table 24.

considering a broader context and retrieving sentences around the most relevant sentence. This leads to the generation of higher-quality context. Finally, the vector-based retriever approach simply searches for related context through a vector index.

The Reranker Module The primary objective of rerankers is to refine the retrieved information by repositioning the most pertinent content towards the prompt edges. To accomplish this, we examine the influence of three rerankers on the overall system performance: (1) LLMRerank (Fajardo, 2023), (2) Bge-Ranker-Base[8], and (3) All-Mpnet-Base-v2(Song et al., 2020).

The Generator Module The generator module first consolidates the query and relevant document context prepared by the former modules into a well-structured and coherent prompt. These prompts are then fed to a LLM to generate final responses. To evaluate the performance of various LLMs, we include six *sota* models, including (1) Qwen-7B (Bai et al., 2023), (2) Baichuan2-7B (Yang et al., 2023a), (3)LLaMA2-7B (Touvron et al., 2023), (4) GPT-3.5-turbo, (5) Solar-10.7B (Kim et al., 2023a), and (6) Gemini-Pro (Team et al., 2023).

4.2 Experimental Settings

To ensure a thorough understanding of each model within every module in a controlled manner, we systematically tested all configurations of models in each module in the RAG-based LFQA system to determine the optimal one. All configurations are evaluated on two sets of experiments - one where the generators were fine-tuned using the training set of *FinTextQA*, and another without such fine-tuning. Note that Gemini-Pro remains a private model and is thus excluded from the fine-tuning process.

To understand the robustness of the best systems, we select the three highest-ranking configurations based on their performance with generators in their base form. This criterion ensures a fair comparison with Gemini-Pro. We then evaluate the performance of these systems under conditions of increased noise by incrementally adding numbers of documents from one to three.

Hyperparameter settings involved in the experiments are set as follows:

Retrievers For AMR, we define three levels of chunk sizes: 2048 for the first level, 512 for the second, and 128 for the third. For the SWR method, we set the window size to 3. For all retrievers, the similarity top k value was set to 6.

Rerankers We set the LLMRerank batch size to 5, and the top n values of LLMRerank, Bge-Reranker-Base, and All-Mpnet-Base-v2 to 4.

Generator We use AzureOpenai's API [9] to access GPT-3.5-turbo and GPT-4-0314 for GPT series models. Google VertexAI API[10] is used to access Gemini-Pro. The LLaMA2, Baichuan2, and Qwen models are all used in their 7B versions, while the Solar model is accessed in the 10.7B version. Fine-tuning of open-source models is carried out on the training set of *FinTextQA*. For GPT-3.5-turbo, we adopt the fine-tuning methods in Azure AI Studio[11], setting the batch size to 2, learning rate multiplier to 1, and epochs to 5.

GPTQConfig (Frantar et al., 2022) is used to load the Qwen-7B model in 4-bit, the GenerationConfig (Joao Gante, 2022) for Baichuan2-7B in 4-bit, and the BitsAndBytesConfig (Belkada, 2023) for LLaMA2-7B and Solar-10.7B in 4-bit. We employ LoRA for LLaMA2-7B, Baichuan2-7B, Qwen-7B,

[8] https://huggingface.co/BAAI/bge-reranker-base

[9] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[10] https://cloud.google.com/vertex-ai/docs/reference/rest

[11] https://oai.azure.com

and Solar-10.7B, with the rank set to 1, alpha set to 32, and dropout at 0.1. Prefix token lengths are set to 2048, learning rate to 1.0e-3, batch size to 2, and maximum input and target length to 2048. All fine-tuning efforts are performed using 12 NVIDIA RTX3090 GPUs for 10 epochs.

4.3 Evaluation Methods

4.3.1 Evaluation of Individual Modules

Embedders, Retrievers, and Rerankers. In recent studies, such as (Sottana et al., 2023; Cao et al., 2024; Liu et al., 2023; Ye et al., 2023) and (Kim et al., 2023b), the GPT-4 evaluator has been extensively tested. To evaluate the performance of these modules and their combined performance in evidence generation, we use GPT-4 to analyze the relevance between questions and retrieved citations (aka. evidence). In detail, GPT-4 is asked to grade the question-evidence relevance on a five-point Likert scale. The average score, referred to as the 'GPT-4 score', is calculated for overall performance evaluation. The prompt used for GPT-4-aided evaluation is shown in Appendix A.2.

Generators. To evaluate the performance of generators, we employ automatic metrics comprising matching-based measures such as ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), as well as the BLEU score (Papineni et al., 2002). However, since prior research (Zheng et al., 2023) shows that matching-based metrics may overestimate performance in long sequences, we also use the GPT-4 evaluation method mentioned above to assess evidence-answer relevance. In addition, we report the ratio of unanswered questions in the responses (e.g., cases when models return "can not provide answer based on the content).

4.3.2 Overall Evaluation of the RAG-based LFQA System

ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)) are used to automatically measure the overall system performance. The GPT-4 scoring method is used to evaluate the answers from helpfulness, relevance, accuracy, depth, and creativity. Additionally, we invite three annotators to rank top-performing answers from all tested models and compare them with the ground truth answers, capturing human perception and assessing subjective response quality. Further details of human evaluation can be found in Appendix A.1.

	Answer One	Answer Two	Answer Three	Answer Four
Average Ranking	2.19	2.11	3.10	2.60

Table 6: Comparison of the average rankings of four answers generated by top RAG systems.

4.4 Results

Embedders, Retrievers, and Rerankers Table 4 shows the GPT-4 score of different embedders, retrievers, and rerankers, which constitute the evidence generation pipeline. It also shows at the end the best-performing evidence-generation module combinations. We observe that the highest-performing embedding model is Ada2, achieving a score of 4.586, followed by Ember-v1 (4.486) and Bge-Small-en-v1.5 (4.455) with similar scores. Gte-Large lagged with a noticeable gap with a score of 4.261. Among the retrievers we assess, AMR outperforms the rest with an impressive score of 4.492. SWR ranks second at 4.466, while the simple vector-based approach has the lowest performance with a score of 4.358.

Among the rerankers, Bge-Reranker-Base performs the best, achieving a competitive score of 4.489. LLMRerank ranks second with a score of 4.469, followed by All-Mpnet-Base-v2 with a score of 4.383.

The evidence generation modules together, we observe that the combination of AMR, Ada2, and Bge-Reranker-Base yields the highest score of 4.622, followed by the combination of SWR, Ada2, and Bge-Reranker-Base/All-Mpnet-Base-v2, with a score of 4.620. The marginal differences in performance among these leading combinations indicate that a variety of configurations are capable of yielding satisfactory outcomes for evidence generation.

Generators Table 5 shows the comparison of different generators, contrasted by their base form and fine-tuned form. Although the fine-tuned models have a decreasing loss (from 2.5 to 0.1), they do not have significant improvement. They demonstrate slightly lower performance in terms of GPT-4 score, ROUGE-L, and BLEU scores. However, fine-tuned models have less unanswered questions, showing better understanding capabilities than their base forms.

We also observe that while Gemini-Pro shows high numeric scores, it struggles the most in generating contextually relevant responses. Conversely, Baichuan2-7B demonstrates the best prompt com-

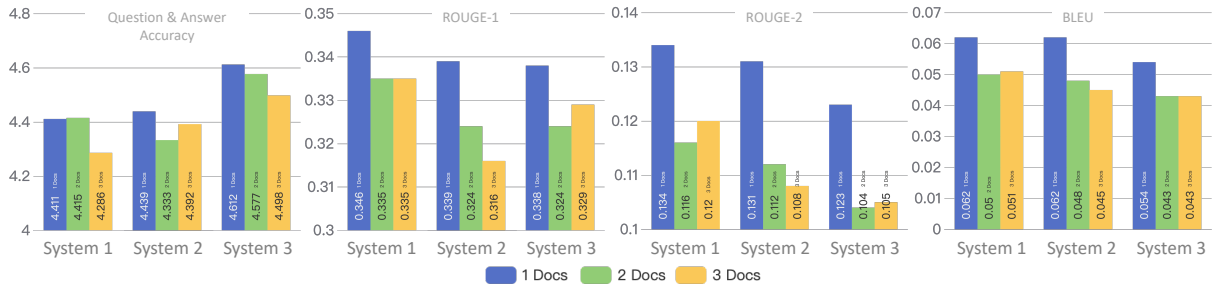

Figure 4: Evaluation of Three Best-performing system configurations with Different Numbers of Input Documents

prehension ability. GPT-3.5-turbo experiences more difficulty with contextual understanding, affecting its overall performance, while LLaMA2 has minimal context-related problems. However, LLaMA2 generates instances of simply rephrasing prompts, resulting in reduced accuracy scores.

RAG-based LFQA System Table 3 shows the performance comparison of RAG-based LFQA system with generators in their base forms. We observe that system with the top-3 performing configurations incorporate GPT-3.5-turbo and Baichuan2-7B as generators. In contrast, system configurations using the Gemini-Pro generator yield suboptimal performance in terms of accuracy. Meanwhile, we observe that system employing LLaMA2-7B as generators show the lowest ROUGE and BLEU scores among all the configurations tested. More results on the performances of different generators in the RAG systems are provided in Appendix A.3.

The top-scoring system configurations comprise (1) GPT-3.5-turbo, AMR, Ada2, and LLM-RERanker (noted as system 1); GPT-3.5-turbo, AMR, Ember-v1, and Bge-Reranker-Base (system 2); (3) Baichuan2-7B, AMR, Ada2, and Bge-Reranker-Base (system 3).

Table 6 shows the annotator-ranked preference of these top system configurations. We notice that some model-generated answers obtain higher average rankings than corresponding ground truths. For instance, Answer 2, produced by system 3, attains an average ranking of 2.11, outperforming the ground truth (2.19). Further investigation into annotator feedback reveals that annotators favor Answer 2 because it gives accurate responses while providing additional details. Answer 3, generated by system 1 performs the worst with the highest average ranking (3.10). Answer 4, generated by system 2, achieves an average ranking of 2.60.

Best System Configuration in Multi-Document Settings Figure 4 shows the performance of the three best-performing system configurations when given different numbers (n = 1 to 3) of documents. We observe a consistent pattern from the results: as the number of input documents increases, all system performance tend to decline. However, exceptions are also noted. For instance, the scores for certain instances with three documents marginally surpasses those with two documents in system 2 when compared to the accuracy score. Further investigation shows that the performance is dependent on the total context words of the input. When the number of context words reaches about 34k words, adding more input documents exerts a less marginal effect on system performance.

Performance of Generators on Regulation and Textbook Dataset Table 11 presents the performance results of all the generators on regulation-based and textbook-based questions. For the regulation and textbook datasets, the Rouge-1 scores are 0.317 and 0.270, respectively; the Rouge-2 scores are 0.146 and 0.084, respectively; the Rouge-L scores are 0.215 and 0.170; the BLEU scores are 0.074 and 0.031; and the GPT-4 scores are 4.238 and 4.293. These results indicate that the LFQA system performs better on the policy/regulation dataset. Since the textbook dataset consists of questions derived from textbook exercises, each question is closely tied to the text content. After annotating the data, 60 policy/regulation data items were removed to ensure that each remaining question had a corresponding answer in the document. Therefore, the quality of the dataset does not affect the performance of the LFQA system. Based on our current findings, the textbook dataset is more challenging.

Performance of Generators on Different Question Types Concerning overall performance on different question types, the models excelled at answering open-ended questions but struggled with numerical ones (Table 10). Specifically, GPT-3.5

Turbo demonstrated the strongest performance in open-ended questions, while Llama2-7B underperformed. Baichuan2-7B showed versatility in handling various question types, but its competence in addressing numerical questions was subpar, placing it third. Remarkably, among all models, Qwen-7B excelled in numerical ability, securing the top position across all metrics.

5 Conclusion

This study presents *FinTextQA*, an LFQA dataset specifically designed for the financial domain. The dataset is comprehensive, covering complex financial question systems and including queries on financial regulations and policies. This makes it a valuable resource for further research and evaluation of RAG modules and large language models. We also introduce a robust evaluation system that leverages human ranking, automatic metrics, and GPT-4 scoring to assess various facets of model performance. Our results suggest that the most effective combination of models and modules for finance-related LFQA tasks includes Ada2, AMR, Bge-Reranker-Base, and Baichuan2-7B.

Limitations

Despite its expert curation and high quality, *FinTextQA* contains a relatively smaller number of QA pairs compared to larger AI-generated datasets. This limitation could potentially affect the generalizability of models trained on it when applied to broader real-world applications. High-quality data are challenging to acquire, and copyright restrictions often prevent sharing. Therefore, future research should concentrate on data augmentation and the development of innovative methods to address data scarcity. Expanding the dataset by incorporating more diverse sources and exploring advanced RAG capabilities and retrieval frameworks could also be beneficial.

Ethical Statement

In this study, we uphold rigorous ethical standards and endeavor to mitigate any potential risks.

- While constructing our dataset, we meticulously ensure that all data are acquired through lawful and ethical means. Adhering to the Fair Use principle, the dataset is exclusively utilized for academic research purposes and is strictly prohibited from commercial exploitation.
- We bear the responsibility of openly sharing the interface, dataset, codes, and trained models with the public. Nonetheless, there exists a possibility of malicious misuse of these resources. For instance, our models could be employed to generate responses without appropriately crediting the information source. We are committed to ensuring their ethical use and guarding against any malicious or harmful intent.
- We are dedicated to mitigating bias, discrimination, or stereotypes during annotation by systematically excluding any questionable examples. To achieve this, we provide thorough training to annotators using 20 samples until they achieve an average accuracy score of 3.8 out of 5. We continually assess their performance throughout the annotation process. Additionally, we provide compensation of $114 per day to annotators until the completion of the annotation task.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62306257) and the Guangzhou Municipal Science and Technology Project (No. 2024A04J4390). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Natural Science Foundation, or the Guangzhou Government.

This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

Finally, we are also grateful to HSBC Lab and HSBC Global Payment Solutions Department for their substantial financial support.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Younes Belkada. 2023. Bitsandbytes.

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023.

Wikihowqa: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314.

Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki. 2015. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375.

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823.

Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. 2024. Rap: Efficient text-video retrieval with sparse-and-correlated adapter. *arXiv preprint arXiv:2405.19465*.

Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, pages 38–56. Springer.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Tingting Cheng. 2021. A multidimensional analysis: speaking style of learner speech across proficiency levels. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 149–159, Shanghai, China. Association for Computational Linguistics.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023a. Lift yourself up: Retrieval-augmented text generation with self memory. *arXiv preprint arXiv:2305.02437*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023c. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, Xianwei Zhuang, and Yuexian Zou. 2024. Towards multi-intent spoken language understanding via hierarchical attention and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17844–17852.

Xuxin Cheng, Zhihong Zhu, Wanshi Xu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023d. Accelerating multiple intent detection and slot filling via targeted knowledge distillation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. 2021. A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391.

Andrei Fajardo. 2023. Llm reranker demonstration (great gatsby).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Wookje Han, Jinsol Park, and Kyungjae Lee. 2023. Prewome: Exploiting presuppositions as working memory for long form question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8312–8322.

Ivan Ilin. 2023. Advanced rag techniques: an illustrated overview.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Sylvain Gugger Joao Gante. 2022. [link].

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023a. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023b. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jerry Liu. 2023. Auto merging retriever.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

llamaindex. 2023. Sentence window retriever.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Yangjun Mao, Jun Xiao, Dong Zhang, Meng Cao, Jian Shao, Yueting Zhuang, and Long Chen. 2023. Improving reference-based distinctive image captioning with contrastive rewards. *arXiv preprint arXiv:2306.14259*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*.

Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*.

Shitao Xiao and Zheng Liu. 2023. [link].

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Bang Yang, Meng Cao, and Yuexian Zou. 2023b. Concept-aware video captioning: Describing videos with effective prior information. *IEEE Transactions on Image Processing*.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023a. Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv:2308.01240*.

Zixuan Yuan, Hao Liu, Renjun Hu, Denghui Zhang, and Hui Xiong. 2021. Self-supervised prototype representation learning for event-based corporate profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4644–4652.

Zixuan Yuan, Junming Liu, Haoyi Zhou, Denghui Zhang, Hao Liu, Nengjun Zhu, and Hui Xiong. 2023b. Lever: Online adaptive sequence learning framework for high-frequency trading. *IEEE Transactions on Knowledge and Data Engineering*.

Zixuan Yuan, Yada Zhu, Wei Zhang, and Hui Xiong. 2023c. Earnings call analysis using a sparse attention based encoder and multi-source counterfactual augmentation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 331–339.

Peiyan Zhang, Haoyang Liu, Chaozhuo Li, Xing Xie, Sunghun Kim, and Haohan Wang. 2023. Foundation model-oriented robustness: Robust image model evaluation with pretrained models. *arXiv preprint arXiv:2308.10632*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Appendix

A.1 Evaluate Human Performance

In our study, we've made it a priority to closely consider the human aspect of performance, covering everything from how we gather data to how we evaluate the answers produced. Our team of annotators, all of whom hold master's degrees with their education conducted in English, play a crucial role in this process.

To accurately identify citations and references, we've laid out a detailed five-step annotation process. At the outset, we provide our team of three annotators with a benchmark example of what a correct citation looks like (shown in Table 9). This serves to clarify several critical criteria: how well the answer fits with the context (Groundedness), how relevant the answer is to the question asked (Answer Relevance), and how the question relates to the context provided (Context Relevance).

After this initial step, the annotators first conduct a practice round involving 20 data samples. Here, they compare their citation identifications against gold standard annotations. We then score their findings and give them feedback to help refine their skills. Once they've shown they've got a good handle on the process, they move on to four more rounds, each with an increasing number of data samples to work through. By the end of this process, after completing 300 tasks, our annotators are well-versed in our annotation standards, ensuring a high level of accuracy in our data collection and analysis.

In each round, we randomly select 10% of the samples for evaluating annotator performance (Table 8). To minimize potential biases, we engage another three annotators to rate relevance and accuracy using a 5-point Likert Scale. Table 7 presents the performance of annotators, revealing that the average scores are above 4 after the first round's training. Context Relevance and Answer Relevance scores are above 3.

The relevance scores in the 5th round are comparatively lower due to the difficulty in finding citations in many pairs. Ultimately, we remove 60 pairs with relevance scores lower than 2 or those lacking citations in the document, retaining 240 pairs for further analysis. This rigorous evaluation and annotation process ensures the quality of *FinTextQA*.

During the answer evaluation phase, annotator competence is measured through their performance in three TOEFL reading tests, ensuring strong reading comprehension skills. Proceeding to the ranking of generated answers, several responses — including ground truth answers — are presented without revealing their origin. If ground truth answers rank too low, the evaluation is considered unsuitable; if ranked within the top two among four responses, the evaluation is considered appropriate.

Annotators analyzed 253 question-answer (QA) pairs by comparing the correct answers to the top 3 answers generated by leading long-form question answering (LFQA) systems, specifically Baichuan2-7B and GPT-3.5 Turbo. This analysis produced four different types of answers for preference ranking. Taking individual differences into account, the agreement scores (Krippendorff's Alpha) for these rankings are: 0.728, 0.647, 0.679,

Round	Average Score	Context Relevance	Answer Relevance
1st - 20 pairs	3.83	4.05	3.44
2nd - 40 pairs	4.08	4.32	3.75
3rd - 60pairs	4.17	4.13	3.65
4th - 80 pairs	4.04	4.08	3.50
5th - 100pairs	4.13	3.51	3.13

Table 7: Performance of human annotation

Ground Truth Citation 1. Below the notification thresholds, the Commission should be able to require the notification of potentially subsidised concentrations that were not yet implemented or the notification of potentially subsidised bids prior to the award of a contract, if it considers that the concentration or the bid would merit ex ante review given its impact in the Union. 2. The Commission may request the prior notification of any concentration which is not a notifiable concentration within the meaning of Article 20 at any time prior to its implementation where the Commission suspects that foreign subsidies may have been granted to the undertakings concerned in the three years prior to the concentration. Such concentration shall be deemed to be a notifiable concentration for the purposes of this Regulation. 3. By way of derogation from paragraph 2 of this Article, Articles 21 and 29 shall apply from 12 October 2023.
Annotator Citation "1. This Regulation shall enter into force on the twentieth day following that of its publication in the Official Journal of the European Union. 2. It shall apply from 12 July 2023. 3. By way of derogation from paragraph 2 of this Article, Articles 47 and 48 shall apply from 11 January 2023and Article 14(5), (6) and (7) shall apply from 12 January 2024. 4. By way of derogation from paragraph 2 of this Article, Articles 21 and 29 shall apply from 12 October 2023."
Score: 3/5
Feedback: Only select part of the citations, which can not fully answer the question.

Table 8: An example of scoring evidence found by annotators

and 0.706, resulting in an overall agreement score of 0.750. This indicates a high level of consistency.

A.2 Prompt of GPT-4-aided Evaluation

Figure 5 shows the prompt we use to ask GPT-4 to evaluate the relevance and accuracy of model-generated answers in our experiments.

A.3 Experiment Results of Different Generators in RAG Systems

Table 12 - 17 shows a systematic performance comparison of RAG systems with different models in each module.

A.4 Example of Question Types

Table 18 - 23 shows the samples of QA pairs in each question type.

```
SYSTEM: "You are a helpful AI assistant who is good at analyze the text content”

USER: f'''
the user instruction {question_prompt} includes the question and content. the response {generated_answer} is
generated by GPT model.

Please act as an impartial judge and evaluate the quality of the response provided by AI assistant to the
content and question displayed upper!

You should give three scores to the response. the highest score is 5 and the lowest score is 1. the scores
include:

Score 1: the total score considering factors helpfulness, relevance, accuracy, depth, creativity, and level
of detail of the response.
Score 2: the relevancy of question and content.
Score 3: the relevancy of content and response.

Avoid any position biases and ensure that the order in which the responses were presented does not influence
your decision. Do not allow the length of the responses to influence your evaluation,

Do not favor certain names of the assistants.

Be as objective as possible.

Directly output the score and strictly follow the format:
### Score 1: number ### Score 2: number ### Score 3: number

'''
```

Figure 5: Prompt of the GPT-4 Scoring Evaluation Method.

Institute	European Union
Document	REGULATION (EU) 2022/2560 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
Question	Are transactions signed between 12 July 2023 and 12 October 2023 (and implemented on 12 October 2023 or later) subject to mandatory notification under Regulation EU 2022/2560?
Answer	Notifiable concentrations under Article 20 of Regulation EU 2022/2560 for which the agreement was concluded on 12 July 2023 or later but which have not yet been implemented on 12 October 2023, will need to be notified pursuant to Article 21 of Regulation EU 2022/2560 and are subject to the standstill obligation under Article 24 of Regulation EU 2022/2560. By contrast, the notification obligation does not apply to concentrations for which the agreement was concluded on 12 July 2023 or later but which are implemented before 12 October 2023. Notifying Parties are encouraged to engage in pre-notification contacts, in principle as of September 2023, in advance to facilitate the submission of notifications as from 12 October 2023.
Citation	Notifiable concentrations under Article 20 of Regulation EU 2022/2560 for which the agreement was concluded on 12 July 2023 or later but which have not yet been implemented on 12 October 2023, will need to be notified pursuant to Article 21 of Regulation EU 2022/2560 and are subject to the standstill obligation under Article 24 of Regulation EU 2022/2560. By contrast, the notification obligation does not apply to concentrations for which the agreement was concluded on 12 July 2023 or later but which are implemented before 12 October 2023. Notifying Parties are encouraged to engage in pre-notification contacts, in principle as of September 2023, in advance to facilitate the submission of notifications as from 12 October 2023.
Groundedness	5
Answer Relevance	5
Context Relevance	5

Table 9: A Sample of Ground Truth Annotations

Generator	Question Type	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	GPT-4 Score
Qwen-7B	Multi-turn	0.308	0.101	0.192	0.043	4.434
	Finance Domain Knowledge	0.310	0.102	0.193	0.044	4.437
	Comparative Analysis	0.311	0.101	0.194	0.044	4.441
	Open-minded	0.312	0.103	0.194	0.045	4.427
	Cause and Effect Analysis	0.307	0.100	0.191	0.043	4.426
	Numerical	0.305	0.100	0.193	0.043	4.423
Baichuan2-7B	Multi-turn	0.313	0.105	0.204	0.044	4.494
	Finance Domain Knowledge	0.310	0.106	0.199	0.043	4.496
	Comparative Analysis	0.312	0.103	0.196	0.041	4.495
	Open-minded	0.359	0.137	0.236	0.070	4.514
	Cause and Effect Analysis	0.319	0.103	0.198	0.042	4.512
	Numerical	0.274	0.078	0.169	0.025	4.399
LLaMA2-7B	Multi-turn	0.152	0.046	0.100	0.016	4.134
	fFinance Domain Knowledge	0.159	0.049	0.102	0.017	4.142
	Comparative Analysis	0.168	0.050	0.105	0.018	4.143
	Open-minded	0.169	0.058	0.109	0.024	4.132
	Cause and Effect Analysis	0.159	0.047	0.100	0.017	4.146
	Numerical	0.143	0.037	0.094	0.009	4.159
Solar-10.7B	Multi-turn	0.293	0.099	0.185	0.041	4.357
	Finance Domain Knowledge	0.297	0.101	0.183	0.041	4.375
	Comparative Analysis	0.300	0.100	0.183	0.040	4.371
	Open-minded	0.348	0.133	0.221	0.062	4.385
	Cause and Effect Analysis	0.305	0.098	0.184	0.038	4.392
	Numerical	0.253	0.081	0.159	0.025	4.295
GPT-3.5-turbo	Multi-turn	0.326	0.120	0.214	0.054	4.232
	Finance Domain Knowledge	0.324	0.118	0.209	0.052	4.295
	Comparative Analysis	0.329	0.116	0.207	0.050	4.322
	Open-minded	0.377	0.152	0.249	0.076	4.351
	Cause and Effect Analysis	0.329	0.114	0.207	0.048	4.298
	Numerical	0.302	0.090	0.186	0.030	3.943
Gemini-Pro	Multi-turn	0.249	0.091	0.177	0.032	3.823
	Finance Domain Knowledge	0.260	0.093	0.180	0.035	3.924
	Comparative Analysis	0.253	0.084	0.170	0.027	3.979
	Open-minded	0.286	0.121	0.203	0.051	4.019
	Cause and Effect Analysis	0.254	0.084	0.172	0.028	3.908
	Numerical	0.226	0.059	0.146	0.014	3.499
Overall	Multi-turn	0.274	0.094	0.179	0.038	4.246
	Finance Domain Knowledge	0.277	0.095	0.177	0.039	4.278
	Comparative Analysis	0.279	0.092	0.176	0.037	4.292
	Open-minded	0.308	0.117	0.202	0.055	4.305
	Cause and Effect Analysis	0.279	0.091	0.175	0.036	4.280
	Numerical	0.250	0.074	0.158	0.024	4.120

Table 10: Detailed Experiment Results of Generator on Question Types.

Generator	Document Type	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	GPT-4 Score
LLaMA2-7B	Regulation	0.194	0.081	0.126	0.037	4.108
	Textbook	0.152	0.042	0.097	0.013	4.157
Baichuan2-7B	Regulation	0.362	0.170	0.246	0.086	4.569
	Textbook	0.306	0.093	0.191	0.036	4.490
Solar-10.7B	Regulation	0.355	0.151	0.224	0.076	4.223
	Textbook	0.289	0.091	0.177	0.034	4.416
GPT-3.5-turbo	Regulation	0.367	0.178	0.252	0.093	4.208
	Textbook	0.321	0.107	0.202	0.044	4.330
Gemini-Pro	Regulation	0.261	0.131	0.201	0.061	3.818
	Textbook	0.258	0.083	0.172	0.028	3.951
Qwen-7B	Regulation	0.363	0.166	0.241	0.088	4.504
	Textbook	0.296	0.086	0.181	0.033	4.413
Overall	Regulation	0.317	0.146	0.215	0.074	4.238
	Textbook	0.270	0.084	0.170	0.031	4.293

Table 11: Detailed Experiment Results of Generator on Textbook and Regulation Questions.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.420	0.340	0.130	0.218	0.062
		LLMRerank	4.399	0.335	0.126	0.214	0.058
		All-Mpnet-Base-v2	4.261	0.320	0.116	0.205	0.048
	Ada2	Bge-Reranker-Base	4.359	0.338	0.129	0.218	0.060
		LLMRerank	4.411	0.346	0.134	0.224	0.062
		All-Mpnet-Base-v2	4.291	0.327	0.121	0.209	0.053
	Ember-v1	Bge-Reranker-Base	4.439	0.339	0.131	0.221	0.062
		LLMRerank	4.365	0.341	0.130	0.221	0.060
		All-Mpnet-Base-v2	4.278	0.328	0.120	0.211	0.052
	Gte-Large	Bge-Reranker-Base	4.319	0.332	0.125	0.213	0.056
		LLMRerank	4.312	0.325	0.121	0.207	0.052
		All-Mpnet-Base-v2	4.252	0.312	0.108	0.197	0.045
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.426	0.322	0.114	0.208	0.050
		LLMRerank	4.378	0.325	0.114	0.209	0.050
		All-Mpnet-Base-v2	4.367	0.327	0.116	0.211	0.051
	Ada2	Bge-Reranker-Base	4.464	0.320	0.113	0.208	0.048
		LLMRerank	4.462	0.321	0.116	0.209	0.052
		All-Mpnet-Base-v2	4.361	0.320	0.116	0.206	0.052
	Ember-v1	Bge-Reranker-Base	4.405	0.331	0.120	0.216	0.053
		LLMRerank	4.384	0.328	0.118	0.214	0.053
		All-Mpnet-Base-v2	4.394	0.323	0.113	0.210	0.049
	Gte-Large	Bge-Reranker-Base	4.183	0.312	0.108	0.200	0.046
		LLMRerank	4.268	0.309	0.109	0.201	0.047
		All-Mpnet-Base-v2	4.255	0.311	0.108	0.201	0.046
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	4.255	0.325	0.121	0.209	0.053
		LLMRerank	4.255	0.330	0.126	0.212	0.057
		All-Mpnet-Base-v2	4.215	0.339	0.125	0.216	0.058
	Ada2	Bge-Reranker-Base	4.218	0.331	0.125	0.213	0.057
		LLMRerank	4.289	0.326	0.124	0.210	0.057
		All-Mpnet-Base-v2	4.249	0.332	0.124	0.213	0.058
	Ember-v1	Bge-Reranker-Base	4.243	0.330	0.124	0.212	0.056
		LLMRerank	4.253	0.327	0.119	0.208	0.053
		All-Mpnet-Base-v2	4.278	0.327	0.124	0.210	0.057
	Gte-Large	Bge-Reranker-Base	4.065	0.316	0.110	0.200	0.048
		LLMRerank	4.034	0.320	0.115	0.204	0.051
		All-Mpnet-Base-v2	4.099	0.314	0.112	0.199	0.049

Table 12: Detailed Experiment Results of GPT-3.5-turbo in RAG Systems.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.536	0.331	0.119	0.213	0.052
		LLMRerank	4.544	0.336	0.120	0.213	0.052
		All-Mpnet-Base-v2	4.527	0.320	0.110	0.201	0.047
	Ada2	Bge-Reranker-Base	4.612	0.338	0.123	0.217	0.054
		LLMRerank	4.578	0.340	0.124	0.219	0.057
		All-Mpnet-Base-v2	4.521	0.320	0.112	0.201	0.049
	Ember-v1	Bge-Reranker-Base	4.513	0.333	0.120	0.215	0.053
		LLMRerank	4.618	0.334	0.124	0.215	0.058
		All-Mpnet-Base-v2	4.549	0.331	0.116	0.209	0.050
	Gte-Large	Bge-Reranker-Base	4.540	0.323	0.112	0.208	0.050
		LLMRerank	4.532	0.328	0.119	0.210	0.053
		All-Mpnet-Base-v2	4.536	0.314	0.102	0.198	0.044
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.605	0.294	0.090	0.183	0.034
		LLMRerank	4.593	0.304	0.097	0.192	0.038
		All-Mpnet-Base-v2	4.616	0.302	0.097	0.189	0.038
	Ada2	Bge-Reranker-Base	4.606	0.305	0.101	0.195	0.041
		LLMRerank	4.586	0.309	0.099	0.191	0.039
		All-Mpnet-Base-v2	4.540	0.307	0.100	0.193	0.039
	Ember-v1	Bge-Reranker-Base	4.584	0.311	0.105	0.197	0.043
		LLMRerank	4.568	0.311	0.104	0.194	0.044
		All-Mpnet-Base-v2	4.571	0.308	0.104	0.195	0.042
	Gte-Large	Bge-Reranker-Base	4.589	0.307	0.101	0.193	0.041
		LLMRerank	4.553	0.305	0.101	0.193	0.042
		All-Mpnet-Base-v2	4.576	0.300	0.098	0.188	0.039
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	4.458	0.316	0.109	0.203	0.045
		LLMRerank	4.481	0.314	0.106	0.201	0.045
		All-Mpnet-Base-v2	4.422	0.317	0.108	0.202	0.044
	Ada2	Bge-Reranker-Base	4.477	0.319	0.113	0.206	0.050
		LLMRerank	4.481	0.318	0.113	0.204	0.050
		All-Mpnet-Base-v2	4.470	0.317	0.112	0.202	0.049
	Ember-v1	Bge-Reranker-Base	4.513	0.324	0.113	0.207	0.049
		LLMRerank	4.464	0.329	0.115	0.210	0.051
		All-Mpnet-Base-v2	4.501	0.319	0.109	0.204	0.047
	Gte-Large	Bge-Reranker-Base	4.416	0.317	0.106	0.201	0.048
		LLMRerank	4.454	0.321	0.110	0.205	0.050
		All-Mpnet-Base-v2	4.409	0.312	0.108	0.202	0.045

Table 13: Detailed Experiment Results of Baichuan2-7B in RAG Systems.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.310	0.329	0.118	0.205	0.051
		LLMRerank	4.418	0.323	0.115	0.200	0.050
		All-Mpnet-Base-v2	4.331	0.312	0.106	0.193	0.045
	Ada2	Bge-Reranker-Base	4.378	0.327	0.119	0.204	0.051
		LLMRerank	4.350	0.322	0.116	0.200	0.050
		All-Mpnet-Base-v2	4.357	0.323	0.111	0.197	0.045
	Ember-v1	Bge-Reranker-Base	4.348	0.329	0.119	0.205	0.052
		LLMRerank	4.388	0.330	0.117	0.204	0.052
		All-Mpnet-Base-v2	4.317	0.319	0.110	0.196	0.046
	Gte-Large	Bge-Reranker-Base	4.328	0.318	0.113	0.198	0.048
		LLMRerank	4.338	0.318	0.110	0.197	0.045
		All-Mpnet-Base-v2	4.262	0.302	0.098	0.185	0.039
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.431	0.296	0.094	0.184	0.037
		LLMRerank	4.462	0.298	0.095	0.184	0.037
		All-Mpnet-Base-v2	4.458	0.300	0.094	0.185	0.037
	Ada2	Bge-Reranker-Base	4.424	0.296	0.095	0.183	0.037
		LLMRerank	4.460	0.297	0.096	0.182	0.037
		All-Mpnet-Base-v2	4.431	0.293	0.093	0.180	0.037
	Ember-v1	Bge-Reranker-Base	4.456	0.305	0.097	0.190	0.039
		LLMRerank	4.376	0.305	0.099	0.190	0.039
		All-Mpnet-Base-v2	4.394	0.307	0.100	0.190	0.041
	Gte-Large	Bge-Reranker-Base	4.344	0.297	0.094	0.184	0.037
		LLMRerank	4.354	0.296	0.094	0.184	0.037
		All-Mpnet-Base-v2	4.361	0.297	0.095	0.184	0.038
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	4.375	0.300	0.105	0.183	0.043
		LLMRerank	4.384	0.300	0.105	0.184	0.042
		All-Mpnet-Base-v2	4.414	0.301	0.106	0.184	0.043
	Ada2	Bge-Reranker-Base	4.422	0.298	0.106	0.183	0.044
		LLMRerank	4.409	0.297	0.105	0.183	0.044
		All-Mpnet-Base-v2	4.359	0.299	0.107	0.184	0.045
	Ember-v1	Bge-Reranker-Base	4.441	0.299	0.106	0.185	0.045
		LLMRerank	4.384	0.301	0.106	0.186	0.045
		All-Mpnet-Base-v2	4.420	0.297	0.106	0.184	0.045
	Gte-Large	Bge-Reranker-Base	4.253	0.285	0.096	0.172	0.040
		LLMRerank	4.298	0.287	0.097	0.174	0.040
		All-Mpnet-Base-v2	4.304	0.290	0.096	0.174	0.040

Table 14: Detailed Experiment Results of Solar-10.7B in RAG Systems.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.445	0.331	0.118	0.211	0.055
		LLMRerank	4.414	0.341	0.125	0.217	0.059
		All-Mpnet-Base-v2	4.414	0.320	0.107	0.198	0.047
	Ada2	Bge-Reranker-Base	4.405	0.337	0.120	0.216	0.056
		LLMRerank	4.538	0.335	0.119	0.211	0.055
		All-Mpnet-Base-v2	4.420	0.333	0.115	0.209	0.052
	Ember-v1	Bge-Reranker-Base	4.456	0.341	0.120	0.215	0.056
		LLMRerank	4.432	0.339	0.121	0.215	0.056
		All-Mpnet-Base-v2	4.361	0.328	0.110	0.204	0.050
	Gte-Large	Bge-Reranker-Base	4.399	0.336	0.118	0.210	0.051
		LLMRerank	4.424	0.331	0.117	0.210	0.052
		All-Mpnet-Base-v2	4.368	0.315	0.103	0.195	0.044
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.529	0.314	0.101	0.194	0.044
		LLMRerank	4.517	0.320	0.104	0.198	0.043
		All-Mpnet-Base-v2	4.490	0.322	0.105	0.198	0.047
	Ada2	Bge-Reranker-Base	4.540	0.326	0.110	0.206	0.050
		LLMRerank	4.548	0.322	0.104	0.199	0.049
		All-Mpnet-Base-v2	4.525	0.323	0.110	0.204	0.051
	Ember-v1	Bge-Reranker-Base	4.473	0.324	0.106	0.202	0.046
		LLMRerank	4.511	0.326	0.109	0.203	0.048
		All-Mpnet-Base-v2	4.508	0.321	0.106	0.198	0.045
	Gte-Large	Bge-Reranker-Base	4.483	0.314	0.102	0.193	0.045
		LLMRerank	4.424	0.316	0.104	0.196	0.046
		All-Mpnet-Base-v2	4.430	0.310	0.099	0.190	0.042
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	4.424	0.310	0.101	0.194	0.046
		LLMRerank	4.416	0.302	0.098	0.190	0.040
		All-Mpnet-Base-v2	4.388	0.290	0.093	0.180	0.038
	Ada2	Bge-Reranker-Base	4.430	0.299	0.102	0.190	0.044
		LLMRerank	4.382	0.298	0.100	0.190	0.042
		All-Mpnet-Base-v2	4.378	0.298	0.096	0.184	0.042
	Ember-v1	Bge-Reranker-Base	4.405	0.311	0.102	0.194	0.045
		LLMRerank	4.468	0.312	0.099	0.194	0.041
		All-Mpnet-Base-v2	4.489	0.294	0.095	0.182	0.039
	Gte-Large	Bge-Reranker-Base	4.302	0.288	0.086	0.175	0.037
		LLMRerank	4.357	0.289	0.088	0.179	0.037
		All-Mpnet-Base-v2	4.369	0.294	0.094	0.183	0.039

Table 15: Detailed Experiment Results of Qwen-7B in RAG Systems.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.268	0.239	0.078	0.151	0.031
		LLMRerank	4.287	0.233	0.076	0.149	0.031
		All-Mpnet-Base-v2	4.240	0.199	0.058	0.127	0.021
	Ada2	Bge-Reranker-Base	4.220	0.218	0.073	0.138	0.029
		LLMRerank	4.250	0.219	0.074	0.141	0.030
		All-Mpnet-Base-v2	4.203	0.206	0.062	0.131	0.022
	Ember-v1	Bge-Reranker-Base	4.215	0.230	0.076	0.146	0.031
		LLMRerank	4.287	0.223	0.075	0.142	0.029
		All-Mpnet-Base-v2	4.272	0.215	0.063	0.134	0.021
	Gte-Large	Bge-Reranker-Base	4.279	0.221	0.074	0.142	0.029
		LLMRerank	4.272	0.212	0.068	0.135	0.025
		All-Mpnet-Base-v2	4.181	0.213	0.059	0.133	0.019
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	4.141	0.214	0.067	0.137	0.025
		LLMRerank	4.202	0.225	0.072	0.145	0.027
		All-Mpnet-Base-v2	4.216	0.215	0.066	0.136	0.025
	Ada2	Bge-Reranker-Base	4.222	0.220	0.070	0.141	0.026
		LLMRerank	4.230	0.224	0.071	0.144	0.028
		All-Mpnet-Base-v2	4.184	0.233	0.078	0.152	0.030
	Ember-v1	Bge-Reranker-Base	4.215	0.218	0.070	0.139	0.025
		LLMRerank	4.196	0.203	0.065	0.131	0.023
		All-Mpnet-Base-v2	4.295	0.214	0.067	0.138	0.026
	Gte-Large	Bge-Reranker-Base	4.181	0.206	0.064	0.135	0.022
		LLMRerank	4.181	0.198	0.058	0.127	0.021
		All-Mpnet-Base-v2	4.259	0.206	0.062	0.131	0.022
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	4.193	0.175	0.049	0.110	0.016
		LLMRerank	4.243	0.168	0.049	0.108	0.016
		All-Mpnet-Base-v2	4.193	0.162	0.043	0.102	0.013
	Ada2	Bge-Reranker-Base	4.246	0.178	0.054	0.111	0.019
		LLMRerank	4.179	0.176	0.055	0.110	0.022
		All-Mpnet-Base-v2	4.247	0.164	0.047	0.104	0.014
	Ember-v1	Bge-Reranker-Base	4.151	0.180	0.054	0.113	0.016
		LLMRerank	4.256	0.177	0.052	0.111	0.017
		All-Mpnet-Base-v2	4.193	0.176	0.048	0.109	0.015
	Gte-Large	Bge-Reranker-Base	4.229	0.155	0.042	0.098	0.013
		LLMRerank	4.215	0.172	0.049	0.106	0.019
		All-Mpnet-Base-v2	4.245	0.166	0.046	0.106	0.016

Table 16: Detailed Experiment Results of LLaMA2-7B in RAG Systems.

Retriever	Embedder	Reranker	GPT-4 Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
AMR	Bge-Small-en-v1.5	Bge-Reranker-Base	3.887	0.298	0.115	0.206	0.045
		LLMRerank	3.989	0.303	0.119	0.210	0.051
		All-Mpnet-Base-v2 2	3.567	0.272	0.099	0.187	0.036
	Ada2	Bge-Reranker-Base	4.063	0.302	0.118	0.208	0.049
		LLMRerank	3.983	0.300	0.119	0.208	0.048
		All-Mpnet-Base-v2	3.698	0.278	0.102	0.190	0.037
	Ember-v1	Bge-Reranker-Base	3.970	0.304	0.118	0.211	0.048
		LLMRerank	3.990	0.306	0.119	0.211	0.052
		All-Mpnet-Base-v2	3.667	0.279	0.102	0.191	0.039
	Gte-Large	Bge-Reranker-Base	3.840	0.294	0.111	0.199	0.042
		LLMRerank	3.793	0.284	0.104	0.195	0.042
		All-Mpnet-Base-v2	3.344	0.255	0.093	0.178	0.030
SWR	Bge-Small-en-v1.5	Bge-Reranker-Base	3.894	0.257	0.091	0.178	0.028
		LLMRerank	3.922	0.257	0.094	0.177	0.030
		All-Mpnet-Base-v2	3.894	0.259	0.092	0.179	0.033
	Ada2	Bge-Reranker-Base	3.996	0.254	0.088	0.177	0.028
		LLMRerank	3.975	0.253	0.089	0.176	0.030
		All-Mpnet-Base-v2	4.023	0.261	0.095	0.183	0.032
	Ember-v1	Bge-Reranker-Base	3.938	0.272	0.102	0.189	0.036
		LLMRerank	4.037	0.268	0.101	0.188	0.036
		All-Mpnet-Base-v2	4.035	0.269	0.094	0.184	0.030
	Gte-Large	Bge-Reranker-Base	3.743	0.246	0.086	0.174	0.026
		LLMRerank	3.731	0.249	0.093	0.176	0.033
		All-Mpnet-Base-v2 2	3.772	0.246	0.091	0.174	0.030
Vector Retriever	Bge-Small-en-v1.5	Bge-Reranker-Base	3.914	0.262	0.091	0.177	0.032
		LLMRerank	3.948	0.256	0.091	0.177	0.036
		All-Mpnet-Base-v2	3.819	0.257	0.094	0.177	0.034
	Ada2	Bge-Reranker-Base	3.930	0.262	0.097	0.180	0.040
		LLMRerank	3.863	0.266	0.100	0.183	0.041
		All-Mpnet-Base-v2	3.981	0.254	0.089	0.173	0.033
	Ember-v1	Bge-Reranker-Base	3.946	0.265	0.094	0.184	0.038
		LLMRerank	3.873	0.258	0.090	0.176	0.033
		All-Mpnet-Base-v2	4.044	0.262	0.091	0.179	0.036
	Gte-Large	Bge-Reranker-Base	3.521	0.227	0.076	0.157	0.027
		LLMRerank	3.544	0.237	0.078	0.163	0.027
		All-Mpnet-Base-v2	3.542	0.229	0.078	0.158	0.027

Table 17: Detailed Experiment Results of Gemini-Pro in RAG Systems.

Question Type	Numerical Reasoning
Amount	63
Question	Critics have charged that compensation to top managers in the United States is simply too high and should be cut back. For example, focusing on large corporations, Mark Parker, CEO of Nike, earned about $47.6 million in 2016. Are such amounts excessive? In answering, it might be help- ful to recognize that superstar athletes such as LeBron James, top entertainers such as Taylor Swift and Dwayne Johnson, and many others at the top of their respective fields earn at least as much, if not a great deal more.
Answer	"How much is too much? Who is worth more, Mark Parker or LeBron James? The simplest answer is that there is a market for executives just as there is for all types of labor. Executive compensation is the price that clears the market. The same is true for athletes and performers. Having said that, one aspect of executive compensation deserves comment. A primary reason executive compensation has grown so dramatically is that companies have increasingly moved to stock-based compensation. Such movement is obviously consistent with the attempt to better align stockholder and management interests. In recent years, stock prices have soared, so management has cleaned up. It is sometimes argued that much of this reward is due to rising stock prices in general, not managerial performance. Perhaps in the future, executive compensation will be designed to reward only differential performance, that is, stock price increases in excess of general market increases.

Table 18: An Example of Numerical Reasoning QA Pairs

Question Type	Multi-turn
Amount	356
Question	Who owns a corporation? Describe the process whereby the owners control the firm's management. What is the main reason that an agency relationship exists in the corporate form of organization? In this context, what kinds of problems can arise?
Answer	In the corporate form of ownership, the shareholders are the owners of the firm. The shareholders elect the directors of the corporation, who in turn appoint the firm's management. This separation of ownership from control in the corporate form of organization is what causes agency problems to exist. Management may act in its own or someone else's. best interests, rather than those of the shareholders. If such events occur, they may contradict the goal of maximizing the share price of the equity of the firm.

Table 19: An Example of Multi-turn QA Pairs

Question Type	Finance Domain Knowledge
Amount	795
Question	What is a pro forma statement of cash flows and what is its purpose?
Answer	A pro forma statement of cash flows estimates the borrower's future cash flows. It is supposed to provide insight into the future cash flows of the borrower and its ability to repay the loan.

Table 20: An Example of Finance Domain Knowledge QA Pairs

Question Type	Comparative Analysis
Amount	392
Question	Suppose a company has a preferred stock issue and a common stock issue. Both have just paid a $2 dividend. Which do you think will have a higher price, a share of the preferred or a share of the common?
Answer	The common stock probably has a higher price because the dividend can grow, whereas it is fixed on the preferred. However, the preferred is less risky because of the dividend and liquidation preference, so it is possible the preferred could be worth more, depending on the circumstances.

Table 21: Example of Comparative Analysis QA Pair

Question Type	Open-minded
Amount	102
Question	Suppose you were the financial manager of a not-for-profit business (a not-for-profit hospital, perhaps). What kinds of goals do you think would be appropriate?
Answer	Such organizations frequently pursue social or political missions, so many different goals are conceivable. One goal that is often cited is revenue minimization; that is, provide whatever goods and services are offered at the lowest possible cost to society. A better approach might be to observe that even a not-for-profit business has equity. Thus, one answer is that the appropriate goal is to maximize the value of the equity.

Table 22: Example of Open-minded QA Pair

Question Type	Cause and Effect Analysis
Amount	688
Question	Last month, Central Virginia Power Company, which had been having trouble with cost overruns on a nuclear power plant that it had been building, announced that it was temporarily suspending payments due to the cash flow crunch associated with its investment program. The company stock price dropped from $28.50 to $25 when this announcement was made. How would you interpret this change in the stock price (that is, what would you say caused it)?
Answer	The stock price dropped because of an expected drop in future dividends. Since the stock price is the present value of all future dividend payments, if the expected future dividend payments decrease, then the stock price will decline.

Table 23: Example of Cause and Effect Analysis QA Pair

Question	Refer to the previous exercise. Which type of foreign market intervention must the central bank of Colombia conduct to keep the exchange rate at a level where the currency is not under- or overvalued in terms of PPP?
Baichuan2-7B	To answer the above question, we need to know the current exchange rate of the Colombian peso relative to the U.S. dollar. Assuming the current exchange rate is not under- or overvalued in terms of PPP, the central bank of Colombia should conduct sterilized intervention to maintain this exchange rate. Sterilized intervention involves selling foreign assets (e.g., dollars) and purchasing domestic assets (e.g., pesos) to offset the effect of the intervention on the money supply. If the central bank sells dollars and buys pesos, the money supply in pesos will increase while the money supply in dollars will decrease. This will cause the value of the peso to appreciate relative to the dollar, maintaining the current exchange rate.
GPT-3.5-turbo	Based on the content provided, the information necessary to answer the question is not present. The content does not mention the central bank of Colombia or provide any information about the exchange rate of the Colombian currency. Therefore, we cannot determine which type of foreign market intervention the central bank of Colombia must conduct to keep the exchange rate at a level where the currency is not under- or overvalued in terms of PPP.
Ground Truth	To eliminate the overvaluation in terms of PPP, the exchange rate for the Columbian peso needs to decline. The central bank of Colombia should undertake an unsterilized foreign market intervention in which it will sell the domestic currency (COP) and buy foreign reserves, in order to increase reserves of its banking system, decrease the domestic interest rate, and shift the expected return on domestic currency denominated assets curve to the left.

Table 24: An Example of Unanswered Questions. We compare answers generated by GPT-3.5-turbo and Baichuan2-7B with the same embedder, retriever, and Reranker.