Resubmission Overview

This document provides an overview of all changes made for the resubmission. We closely followed the reviewers' feedback and implemented all requested changes to the best of our understanding. We added a colored version of all changes made to the paper to the end of this PDF document.

Removal of the VQAonBD Subset

The most important change is the removal of the VQAonBD subset from the benchmark, which was included in the original submission. We made this decision based on two main reasons:

1. As highlighted by reviewers, VQAonBD contains **only tables** without any accompanying text. Since our benchmark is designed to evaluate performance on mixed text-and-table inputs, this subset did not align well with our primary goal.
2. During the requested error analysis, we also **reviewed** the reformulated questions for this subset. Many of the original questions were simple lookups or basic arithmetic with little domain context. After reformulation, the added metadata often resulted in lower data quality, making the reformulated questions noticeably worse than those in other subsets.

Given these issues, we chose to remove VQAonBD entirely. We carefully considered the pros and cons of this step. We ultimately decided to prioritize a more minor, but higher-quality benchmark over one that includes lower-quality, LLM-generated questions.

Manual Error Analysis

In addition to the dataset update, we conducted a detailed manual error analysis to better understand common failure cases in the Oracle-Context setting. This helps assess whether the benchmark presents an upper-bound challenge and where further improvements could be made. We manually annotated over 1,500 samples (approximately 25% of all identified error cases) and added a corresponding section to the paper (Lines 501-530). We hope this analysis provides deeper insights and helps guide future development in this area.

To facilitate understanding of the remaining changes, we have organized them in chapter order. We corrected numerous minor typos and revised key sections. The changes are visible in our **color-based** version attached to this PDF.

Motivation & Problem Statement

- We updated the Abstract to add that we did human validation (L019-L023).
- Reviewers noted that our definitions of context-dependence and context-independence are not sufficiently clear and are not adequately cited in the introduction. We added a clear definition in L72-L85 and hope that it is now clearer.

Related Work

- One Reviewer pointed out that we did not discuss potential extensions to other domains. We have Table 2, which explains other datasets from different domains. However, we have completely restructured Section 2 to clarify the current gaps in other domains and to highlight why this dataset is the only one suitable to evaluate RAG methods on text-and-table data (L113-147).

Dataset Construction & Validation

- The reviewer pointed out that the dataset lacks examples of original versus reformulated questions. Therefore, we highlighted in Figure 1 that this is a sample dataset and also pointed out that we have many examples in the Appendix (L237).
- Cohen's Kappa: We reevaluated the calculation of Cohen's Kappa and were able to fix a bug in the calculation. Additionally, we excluded VQAonBD, which had the lowest score. Therefore, the score increased from 0.57 to 0.87. We also added it to Figure 3 to make it more straightforward, where it belongs, and rewrote the entire section, highlighting more of what was going on (L295-L308). To make it more transparent what the annotators choose, we added examples of the original and reformulated questions to the Appendix and added all raw data into our repository (L306-L308).
- Quality control of the dataset: One reviewer noted that the quality of our dataset is not high enough. Therefore, we conducted additional error analysis (L501-L530) and demonstrated that only 6% of the error cases originate from falsely reformulated questions. That is even less important because the benchmark is primarily used to evaluate RAG, where the primary task is to identify the correct document in the first place. Therefore, we remain convinced that the benchmark is beneficial to the community.

Evaluation

- We changed the title of the section (L320)
- Computational Costs: We made clear that we used quantized models and explained why (L351-353)
- We moved R@3 to the main table, because we gained space through the drop of VQAonBD (Table 3).
- We did the Error Analysis of the Oracle-Context error Cases (L501-530)

Ethical Statement

- We added an ethical statement to the paper (L623-638)

T^2-RAGBench: Text-and-Table Benchmark for Evaluating Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Since many real-world documents combine textual and tabular data, robust Retrieval Augmented Generation (RAG) systems are essential for effectively accessing and analyzing such content to support complex reasoning tasks. Therefore, this paper introduces **T^2-RAGBench**, a benchmark comprising **23,088** question-context-answer triples, designed to evaluate RAG methods on real-world text-and-table data. Unlike typical QA datasets that operate under Oracle-Context settings, T^2-RAGBench challenges models to first retrieve the correct context before conducting numerical reasoning. Existing QA datasets containing text-and-table data typically contain context-dependent questions, which may yield multiple correct answers depending on the provided context. To address this, we transform SOTA datasets into a context-independent format, validated by experts as 91.3% context-independent questions, enabling reliable RAG evaluation. Our comprehensive evaluation identifies *Hybrid BM25*, a technique that combines dense and sparse vectors, as the most effective approach for text-and-table data. However, results demonstrate that T^2-RAGBench remains challenging even for SOTA LLMs and RAG methods. Further ablation studies examine the impact of embedding models and corpus size on retrieval performance. T^2-RAGBench provides a realistic and rigorous benchmark for existing RAG methods on text-and-table data. Code and dataset are available online[1].

1 Introduction

Documents containing a mixture of text and tables are widely utilized in various fields, such as financial reporting (Baviskar et al., 2021), scientific research (Pramanick et al., 2024), and organizational documentation (Rebman Jr et al., 2023).

Recent advancements in Large Language Models (LLMs) have demonstrated solid SOTA performance answering numerical and free-form question-answering (QA) tasks when appropriate documents are provided (Nan et al., 2021; Chen et al., 2021, 2022; Zhu et al., 2021, 2022). Despite increasing context window sizes for LLMs, using the entire corpus remains impractical due to computational constraints and programmatic latency (Wang et al., 2024b; Li et al., 2024). Therefore, retrieving relevant documents is essential in real-world applications to answer questions correctly.

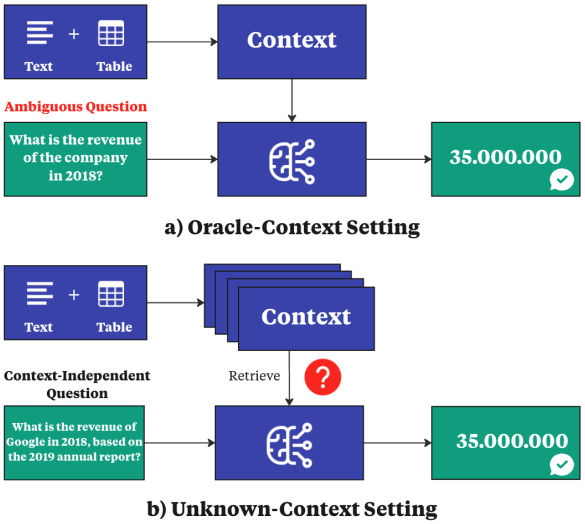

Figure 1: Overview of current SOTA approaches and dataset example. a) Most benchmarks test models in an oracle-context setting, (Chen et al., 2021, 2022). While our task (b) targets the unknown-context setting, requiring retrieval from mixed text-tables before answering.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a promising solution for single-hop QA on numerical tasks, providing appropriate context and has led to an explosion of methods in this area (Gao et al., 2023b; Nikishina et al., 2025). While most RAG methods are effective at retrieving semantically similar text, embedding tabular data remains challenging due to its structural complexity and the predominance of numerical values, which lack semantic context (Khattab et al., 2022).

[1] Anonymous GitHub Repository

In addition, RAG methods are typically trained and evaluated on text-only datasets (Jiang et al., 2023; Lan et al., 2023; Wang et al., 2024b), Wikipedia-derived QA benchmarks (Pasupat and Liang, 2015; Yang et al., 2018) heavily used during LLM pre-training (Grattafiori et al., 2024), or narrow domain-specific datasets (Sarthi et al., 2024; Yan et al., 2024), making it difficult to estimate the performance on text-and-table data. Moreover, as illustrated in Figure 1, existing datasets with text-and-table data operate exclusively under the oracle-context setting, where questions are tightly coupled with the given context. These questions are inherently ambiguous and may yield multiple correct answers depending on the context, we refer to them as **context-dependent**. In contrast, **context-independent** questions have a single correct answer without having access to the context, which is essential for evaluating RAG methods, as they require identifying one ground truth document containing the answer. To our knowledge, no text-and-table dataset meets this requirement.

To fill this gap, we present the **T**ext-**T**able **R**etrieval-**A**ugmented **G**eneration **Bench**mark (T^2-RAGBench), a benchmark designed to evaluate existing RAG methods on text-table retrieval and numerical reasoning tasks. Our benchmark comprises three subsets extracted from existing datasets, totaling 23,088 question-context-answer (QCA) triples and 7,318 real-world financial documents. Each triplet includes a reformulated, context-independent question, a verified answer, and the associated context containing all information to answer the question.

Our contributions are as follows:

- We introduce **T^2-RAGBench**, a benchmark containing **23,088** QCA triples from financial reports designed to evaluate RAG methods on text-and-table and numerical reasoning.
- We systematically evaluate popular RAG methods on **T^2-RAGBench**, demonstrating that it remains a challenging and relevant benchmark for current methods.
- We compare SOTA closed and open-source embedding models and analyze the effect of corpus size on promising RAG methods.

2 Related Work

Text-and-Table QA Datasets. Table 1 gives an overview of existing Q&A datasets containing text and/or tables. While datasets in common knowledge (Joshi et al., 2017; Chen et al., 2020; Nan et al., 2021), scientific documents (Pramanick et al., 2024; Dasigi et al., 2021), or medicine (Fan et al., 2025) focusing exclusively on tables (Katsis et al., 2022), combining text with tables becomes essential for effectively parsing whole PDF documents. Another challenge is data contamination, as common knowledge and scientific datasets often rely on Wikipedia or open-access papers, which are heavily used during LLM pretraining (Grattafiori et al., 2024). This makes it difficult to separate retriever and generator performance in RAG evaluation.

In other domains, such as finance, VQAonBD (Raja et al., 2023) focuses also only on tables, but FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), and TAT-DQA (Zhu et al., 2022) incorporate both text-and-tabular data from financial reports. Nonetheless, all financial datasets contain mainly context-dependent questions.

Moreover, several datasets are not publicly available, such as FinDER (Choi et al., 2025) and BioTABQA (Luo et al., 2022), or represent tables as images rather than structured text in markdown format (Tito et al., 2021; Pramanick et al., 2024). Other datasets are cross-domain, such as TableBench (Wu et al., 2025), which provides multi-domain table QA for Oracle-Context evaluation, while the UDA benchmark (Hui et al., 2024) aggregates multiple datasets. However, both remain limited by context-dependent questions. T^2-RAGBench closes this gap by providing a benchmark that focuses on text-and table-data, has no data contamination, and contains only context-independent questions.

RAG on Text-and-Table. RAG shows promise on text (Lewis et al., 2020), but text-and-table evaluation is limited. THoRR (Kim et al., 2024) simplifies tables via header-based retrieval, complementing ERATTA (Roychowdhury et al., 2024), which uses modular prompts and SQL for enterprise data. FinTextQA (Chen et al., 2024) evaluates full RAG pipelines. FinTMMBench (Zhu et al., 2025) adds multi-modal and temporal RAG via dense/graph retrieval. Robust RAG (Joshi et al., 2024) links text, tables, visuals via image-based VLLMs, though less flexible than text methods. Despite progress, most works (Asai et al., 2024; Gao et al., 2023a,b) test only a few RAG baselines, limiting generalizability.

Dataset	Domain	Text	Table	Visual Independence	Context-Independent	Available	QA Pairs
TriviaQA (Joshi et al., 2017)	Wikipedia	✓	✗	✓	✓	✓	650K
HybridQA (Chen et al., 2020)	Wikipedia	✗	✓	✓	✓	✓	70K
FeTaQA (Nan et al., 2021)	Wikipedia	✗	✓	✓	✓	✓	10K
Qasper (Dasigi et al., 2021)	NLP Papers	✗	✓	✓	✗	✓	5K
SPIQA (Pramanick et al., 2024)	NLP Papers	✗	✓	✗	✗	✓	270K
FinQA (Chen et al., 2021)	Finance	✓	✓	✓	✗	✓	8K
ConvFinQA (Chen et al., 2022)	Finance	✓	✓	✓	✗	✓	14K
TAT-DQA (Zhu et al., 2022)	Finance	✓	✓	✓	✗	✓	16k
VQAonBD (Raja et al., 2023)	Finance	✗	✓	✗	✗	✓	1,531K
FinDER (Choi et al., 2025)	Finance	✓	✓	✓	✓	✗	50K
DocVQA (Tito et al., 2021)	Multiple	✗	✓	✗	✗	✓	50K
TableBench (Wu et al., 2025)	Multiple	✓	✓	✗	✗	✓	∼1K
UDA (Hui et al., 2024)	Multiple	✓	✓	✓	✗	✓	30K
T^2-RAGBench (Ours)	Finance	✓	✓	✓	✓	✓	23K

Table 1: Summary and comparison of Q&A datasets. Visual Independence: The contexts are presented as text and are not only images. Context-Independent: Without a context, questions still only have one unambiguous answer.

3 Task Definition

To clarify the task addressed by our benchmark, we define the following problem to be solved.

Problem Formulation. The benchmark evaluates both the retrieval function f and the reasoning model M to optimize answer accuracy and efficiency in the unknown-context text-and-table QA setting. We denote the user's question by Q and the corresponding ground truth answer by A. The evidence comes from two modalities: a segment of text content and a structured table, which we consider together as a single context entity denoted by C. Thus, our entire context corpus is defined as $\mathcal{C} = \{C_i\}$. The task is divided into two stages:
Retrieval: A function

$$f : \mathcal{C} \times Q \mapsto [C_k^*]_{k=1}^n \tag{1}$$

selects the top-n most relevant context entities from the corpus $\mathcal{C}$ for a given question Q.
Answer Extraction: A language model

$$M : \left([C_k^*]_{k=1}^n,\, Q\right) \mapsto A^* \tag{2}$$

generates an answer A^* by reasoning over the retrieved text and tables.
Number Match: Numerical reasoning is evaluated using a new metric. It allows for minor deviations and unit scale shifts. Let A^* and A be the predicted and ground truth answers, and denote their absolute values as $a^* = |A^*|$ and $a = |A|$.
Given a tolerance threshold $\varepsilon > 0$, the prediction is considered correct if either $a^* < \varepsilon$ and $a < \varepsilon$, or $|q - 1| < \varepsilon$ where

$$q = \frac{a^*}{a} \cdot 10^{-\text{round}(\log_{10}(a^*/a))}.$$

Here, round denotes rounding to the nearest integer. This metric ensures robustness to rounding errors and magnitude scaling.

Retrieval Metrics. Let

$$\mathcal{D} = \{(Q_i, A_i, C_i)\}_{i=1}^N$$

represent our dataset, where each tuple (Q_i, A_i, C_i) consists of a question Q_i, its unique ground-truth answer A_i, and the corresponding unique ground-truth context C_i. Define the retrieval output:

$$R_i = f(\mathcal{C}, Q_i) = [C_{i,1}^*, C_{i,2}^*, \ldots, C_{i,n}^*]. \tag{3}$$

The true rank is given by

$$r_i = \min\{k \mid C_{i,k}^* = C_i\}. \tag{4}$$

We consider the Mean Reciprocal Rank at k (MRR@k), which focuses on the relevance of the top k retrieved contexts. It is defined as

$$\text{MRR@}k = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_i} \cdot \mathbb{I}(r_i \le k), \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function, valued at 1 if the condition is met (i.e., $r_i \le k$), and 0 otherwise.

Subset	Domain	PDF Source	#Documents			#QA Pairs		Avg. Question Tokens	
			Original	Extracted	Avg. Token	Original	Generated	Original	Generated
FinQA	Finance	FinTabNet	2,789	2,789	950.4	8,281	8,281	21.1	39.2
ConvFinQA	Finance	FinTabNet	2,066	1,806	890.9	14,115	3,458	17.8	30.9
TAT-DQA	Finance	TAT-DQA	2,758	2,723	915.3	16,558	11,349	17.8	31.7
Total	**Finance**	**Multiple**	7,613	**7,318**	924.2	38,954	**23,088**	19.0	34.3

Table 2: Comparison of original and generated QA pairs, documents, and average question and context lengths across T^2-RAGBench subsets. FinQA (Chen et al., 2021) and ConvFinQA (Chen et al., 2022) use FinTabNet (Zheng et al., 2020) as their PDF source, while TAT-DQA (Zhu et al., 2022) uses its own dataset. Avg. token count based on Llama 3.3 tokenizer.

4 T^2-RAGBench

To construct our benchmark for text-table data suitable for RAG evaluation, we first surveyed existing datasets, as summarized in Table 1. As none fully met our criteria, we selected FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), and TAT-DQA (Zhu et al., 2022) and restructured them to context-independent questions.

A question is considered context-independent if it has exactly one correct answer, even without access to C. For all selected datasets, we applied custom preprocessing steps and reformulated questions using Llama 3.3-70B[2] to ensure context-independence. Each benchmark sample is a triple (Q, A, C), where Q is a question, A the answer, and C the context composed of both text and table. Since all triples originate from oracle-context settings, we assume that all required information to answer Q is fully contained within C, and only within C. Table 2 provides a detailed breakdown of the three subsets of T^2-RAGBench. While FinQA and ConvFinQA are based on FinTabNet, TAT-DQA is based on its own financial documents. The subsets consist of 1,806 to 2,789 documents, with each containing between 3,458 and 11,349 QA pairs. We included samples for each subset in Appendix A.

4.1 Data Preparation

All subsets required tailored preprocessing to align with the requirements of our benchmark. FinQA is a numerical QA dataset based on financial reports from FinTabNet. We used it with company metadata and standardized all answer formats. ConvFinQA extends FinQA by adding multi-turn questions. We filtered only to include first-turn questions and normalized the answers for consistency. TAT-DQA is an independent dataset with diverse answer types. We filtered it to keep only numerical questions and normalized answer formats. Full details can be found in Appendix B.

[2] kosbu/Llama-3.3-70B-Instruct-AWQ

4.2 Data Creation

Following the preprocessing, the context-independent questions were generated. First, the questions were reformulated using an LLM. Subsequently, both quantitative and qualitative analyses were performed to verify that (1) the data quality remained consistent with the original, and (2) the reformulation process produced genuinely context-independent questions.

Question Reformulation. To generate context-independent questions, the original questions were reformulated, but the answers remained unchanged to preserve human-annotated quality. For each of the 23,088 samples, a new question was generated using Llama 3.3-70B[2] with temperature = 0.7. The generation process was conducted by incorporating meta-information, such as company name, sector, and report year, which were not included in the original document. The exact prompting template is detailed in Appendix C.

Quantitative Analysis. To verify that the rephrased questions remain consistent with the original answer, we conducted a quantitative comparison of the original and reformulated questions across all subsets using Llama 3.3-70B[2] and Oracle-Context, as presented in Figure 2. Since the context is given, only Number Match was used to evaluate the QA pairs. The accuracy between original and generated questions shows minimal deviation, with differences maximal 2% per subset and in average $< 0.05\%$. The ability of the LLM to answer the reformulated questions indicates that they retain the essential information required for numerical reasoning.

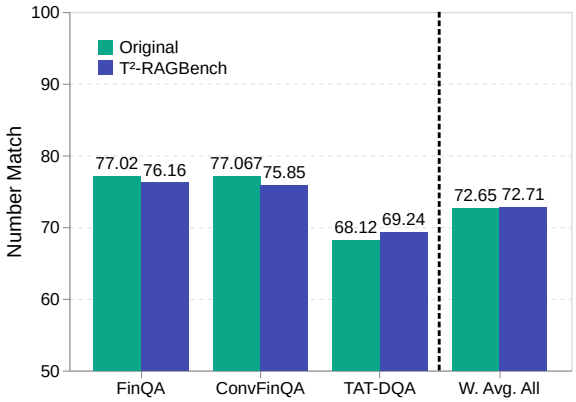

Figure 2: Number Match comparison per subset and weighted average all between original and reformulated questions from our new benchmark.

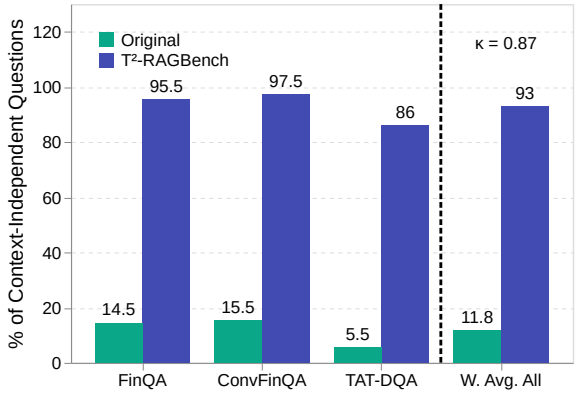

Figure 3: Percentage of context-independent questions (100 per subset, weighted avg overall). κ indicates inter-annotator agreement.

Human Validation. To further investigate whether the questions are now context-independent after reformulation, we conducted a human evaluation after the quantitative analysis. Therefore, a random sample of 100 original and generated QA pairs per subset was manually labeled via a custom annotation tool (Appendix D). Each of the four financial experts annotated 200 samples from two different subsets, assessing whether the original questions were context-independent or context-dependent. The analysis reveals that only 11.8% of questions in the original dataset were context-independent, compared to 93% in the reformulated version (see Figure 3). This ensures that nearly all of the newly created QCA triples are suitable for RAG evaluation. Cohen's Kappa was calculated to assess inter-annotator agreement, yielding an overall value of 0.87, indicating almost perfect agreement. Notably, only 1/3 of the uncertain cases involved reformulated questions, suggesting that most ambiguity stemmed from original question formulations. For better transparency, we include representative disagreement examples in Appendix E and in our repository [1].

4.3 Data Statistics

Table 2 presents an overview of the dataset. It comprises 7,318 real-world documents with an average length of 924.2 tokens. In total, T^2-RAGBench consists of 23,088 QCA triples extracted from roughly 40k questions. Questions increased by ~15 tokens with added semantic details (e.g., company names, years), making them context-independent and suitable for RAG evaluation. All other parameters (Metadata, IDs, etc.) of the dataset remained the same.

5 Experiments

To evaluate the suitability of our benchmark for RAG methods, we report results across all subsets using various models and RAG approaches. This section details the experimental setup (Section 5.1), compares the methods (Section 5.2), outlines the evaluation metrics (Section 5.3), and presents the main results (Section 5.4), which reveal a substantial gap between Oracle and current advanced RAG performance. To investigate this gap, we conduct two ablation studies (Section 5.5) followed by a manual error analysis (Section 5.6) of errors in the Oracle-Context setting to investigate error patterns.

5.1 Experimental Setup

For the evaluation of the benchmark, each subset was evaluated independently. First, all contexts were transformed into markdown format and uniquely stored into a Chroma vector db using the embeddings created with the multilingual e5-large instruct model (Wang et al., 2024a), having an embedding size of 1024. That was done for all RAG methods except for Summarization and SumContext, where the summarized context was embedded. A retrieval query was used to retrieve from the embedding model (See Appendix F). The Top-3 documents were selected and passed to the generator in the main evaluation. As generators, we employed quantized LLaMA 3.3 70B[2], a decoder-only transformer, and QwQ-32B[3], to evaluate performance across multiple model architectures on two NVIDIA H100. Due to resource limitations, we utilize quantized models, which exhibit negligible performance loss (Jin et al., 2024). The prompt template is provided in Appendix G.

[3] Qwen/QwQ-32B-AWQ

5.2 RAG Methods

The following section briefly describes all evaluated RAG methods to show the SOTA performance on T^2-RAGBench, categorized by the retrieval complexity and augmentation strategy.

Pretrained-Only and Oracle Context. In the *Pretrained-Only* setup, no retriever is employed, and models must answer questions solely based on their pretraining knowledge. Conversely, the *Oracle Context* setting assumes that the relevant context is directly passed to the generator.

Basic RAG Methods. This category includes approaches that retrieve documents using standard embedding-based methods. The *Base RAG* implementation follows the original RAG approach (Lewis et al., 2020), where only the question is embedded to retrieve the top-k documents, which are then passed unchanged to the generator. *Hybrid BM25* (Gao et al., 2021) combines sparse lexical retrieval using BM25 with dense vector retrieval, leveraging both methods to improve recall and relevance. Additionally, the *Reranker* method (Tito et al., 2021) applies a cross-encoder model[4] after initial retrieval to reorder documents based on their relevance in a shared embedding space.

Advanced RAG Methods. This category consists of methods that modify the query, transform retrieved contexts, or employ iterative retrieval strategies. The *HyDE* method (Gao et al., 2023a) generates hypothetical answers for each question, using them as refined queries to retrieve more relevant documents (For prompt see Appendix H). *Summarization* reduces noise by summarizing each retrieved context using an LLM, focusing on essential information. *SumContext* applies the similar summarization step but retains the original full documents for generation, aiming to reduce distractions while preserving content fidelity (See Appendix I).

5.3 Evaluation Metrics

We use Number Match and MRR@k as our main metrics as defined in Section 3, but also report Recall@k for better comparability and transparency. **Number Match** evaluates if a numerical prediction closely matches the gold numerical answer. It compares predicted and ground truth values using relative tolerance ($\epsilon = 1\text{e}{-2}$), accounting for scale invariance. Non-numeric predictions or mismatches are considered incorrect. For **MRR** and **Recall**, we choose $k = 3$, which measures whether the first relevant document appears in the top-3 retrieved results, rewarding higher ranks for MRR. We limit evaluation to 3 documents, as the average length is 924.2 tokens. Increasing the number of documents increases input size, slows inference, and hinders LLM performance, making it impractical (Li et al., 2024).

[4] Cross-encoder/ms-marco-MiniLM-L-6-v2

5.4 Main Results

This section discusses our main results presented in Table 3 for all three evaluation subcategories.

Pretrained-Only and Oracle Context. The results from the *Pretrained-Only* setting show that across all subsets, the questions cannot be answered directly from the models' pretraining data. This highlights the importance of RAG and the need for a dedicated benchmark. While reformulated questions may resemble seen content, especially since most S&P 500 reports predate 2023, this applies to both foundation and reasoning models. In contrast, the Oracle Context setting shows consistently high performance on Number Match across all subsets and both models, highlighting both the strong numerical reasoning abilities of the models and the feasibility of the task for modern LLMs in this setting. Notably, there is no significant performance difference between Llama and QwQ ($< 0.3\%$).

Base RAG Methods. For base RAG methods, the benchmark shows that all SOTA models still struggle to match the performance achieved in Oracle-Context. Nevertheless, this benchmark offers the possibility to compare the different methods precisely. For *Base-RAG*, MRR@3 and R@3 averaging below 40%, meaning relevant documents are often missing in the top-3, which leads to a significant drop in Number Match. This effect is particularly evident in TAT-DQA, where, despite having a similar number of documents as FinQA, relevant information is harder to retrieve for all tested methods. *Hybrid BM25* consistently outperforms base RAG in Number Match, MRR@3, and R@3 on average. Interestingly, the Reranker performs worse than *Base* and *Hybrid BM25* RAG methods, suggesting that the reranking model struggles with text-and-table data.

Advanced RAG Methods. One way to improve the performance of RAG methods is to improve the linking of the query with the context. However, *HyDE* shows even a drop in performance in

Model	RAG Method	FinQA			ConvFinQA			TAT-DQA			W. Avg Total		
		NM	MRR@3	R@3	NM	MRR@3	R@3	NM	MRR@3	R@3	NM	MRR@3	R@3
Llama 3.3-70B + Multilingual E5-Large Instruct	+ *Pretrained-Only*	7.9	–	–	2.8	–	–	3.7	–	–	5.1	–	–
	+ *Oracle Context*	76.2	100	100	75.8	100	100	69.2	100	100	72.7	100	–
	+ *Base-RAG*	39.5	38.7	49.7	47.4	42.2	53.8	29.6	25.2	28.4	35.8	32.6	39.8
	+ *Hybrid BM25*	41.7	40.0	53.0	50.3	43.5	57.2	**37.4**	**29.2**	**44.4**	**40.9**	35.2	**49.4**
	+ *Reranker*	32.4	29.0	36.2	37.3	32.3	40.5	27.0	22.8	28.4	30.5	26.4	33.0
	+ *HyDE*	38.4	35.4	45.7	44.8	39.8	50.9	26.7	20.8	26.7	33.6	28.9	37.1
	+ *Summarization*	27.3	47.3	**59.5**	35.2	52.1	**63.8**	14.6	24.7	31.5	22.2	**36.9**	46.4
	+ *SumContext*	**47.2**	**47.3**	**59.4**	**55.5**	**52.1**	**63.8**	29.1	24.8	31.4	39.5	**37.0**	46.3
QwQ-32B + Multilingual E5-Large Instruct	+ *Pretrained-Only*	7.5	–	–	2.4	–	–	4.4	–	–	5.2	–	–
	+ *Oracle Context*	72.4	100	–	85.4	100	–	71.1	100	–	73.7	100	–
	+ *Base-RAG*	39.6	38.7	49.7	48.7	42.4	53.8	27.9	25.2	28.4	35.2	32.6	39.8
	+ *Hybrid BM25*	41.8	39.8	53.0	51.6	43.6	57.2	**37.2**	**29.3**	**44.4**	**41.0**	35.2	**49.4**
	+ *Reranker*	30.8	29.0	36.2	37.5	32.7	40.5	25.6	22.9	28.4	29.2	26.6	33.0
	+ *HyDE*	36.8	35.4	45.7	45.7	39.9	50.9	24.7	20.7	26.7	32.2	28.8	37.1
	+ *Summarization*	26.9	47.2	**59.5**	35.6	52.2	**63.8**	13.9	24.7	31.5	21.8	**36.9**	46.4
	+ *SumContext*	**45.6**	**47.3**	**59.4**	**56.9**	**52.2**	**63.8**	27.3	24.7	31.4	38.3	**36.9**	46.3

Table 3: Overall performance (Number Match (NM), MRR@3, and R@3) of both models on T^2-RAGBench. Number Match represents the percentage of correctly answered questions based on their numerical representation. R@3 and MRR@3 evaluate retrieval effectiveness. Cells in **Bold** indicate the highest value over all RAG methods, and underlined indicate the best value across RAG method categories.

MRR@3 and R@3 across all subsets in comparison to the *Base-RAG*. This may be due to the models' difficulty in generating well-structured content matching the format of the documents, which often include both text and tables.

The *Summarization* approach performed well on MRR@3 for FinQA and ConvFinQA by condensing relevant information and removing noise. However, it underperforms on TAT-DQA, warranting further investigation. In general, this often led to a drop in NM, as essential information needed to answer the questions was also lost during summarization. *SumContext* retrieves from a summarized context but provides the full original context. This approach improved MRR@3 while maintaining stable NM, achieving an average NM of 37.4% resp. 36.7%. Nevertheless, the performance does not improve across all subsets, indicating strong sensitivity to prompts and datasets. Interestingly, MRR@3 is 1.8% higher than *Hybrid BM25*, despite lower R@3, suggesting retrieved documents are ranked higher in Summarization and SumContext.

5.5 Ablation Studies

Embedding Models. We evaluate various embedding models with the *Base-RAG* approach to assess their impact on retrieval performance. As shown in Table 4, among the open-source models, *Multilingual E5-Instruct* performs best, achieving 29.4% R@1 and 38.6 MRR@5. The closed-source models perform slightly better, with the OpenAI model reaching the highest R@1 of 33.8% and MRR@5 of 43.6. However, none of the models, regardless of model size, achieve satisfactory performance on the challenging text-and-table setting at R@1, indicating that retrieving the correct document remains a core challenge in T^2-RAGBench, because text-and-table documents seem to be challenging for SOTA embedding models.

Embedding Model	R@1	R@5	MRR@5
Stella-EN-1.5B	2.2	5.2	3.3
GTE-Qwen2 1.5B Instruct	12.5	27.6	18.0
Multilingual E5-Instruct	**26.4**	**49.7**	**35.1**
Gemini: Text-Embedding-004	32.3	53.6	41.7
OpenAI: Text-Embedding-3 Large	**34.6**	**57.4**	**44.7**

Table 4: Retrieval performance of embedding models on T^2-RAGBench using *Base-RAG* with $k = 5$ retrieved documents, evaluated on R@1, R@5 and MRR@5. Scores are weighted avg. over all subsets. Model descriptions are in Appendix J.

Number of Documents. Figure 4 shows how retrieval performance changes with the number of documents for *Base-RAG* and *Summarization*, using 5 random percentage ascending subsets per dataset. Two main findings emerge: (1) MRR@3 drops below 50% with 3K documents, meaning the correct document appears in the top 3 only half the time; (2) Summarization improves results for FinQA and ConvFinQA, performs similarly on TAT-DQA, where summarizing tabular content is more challenging.

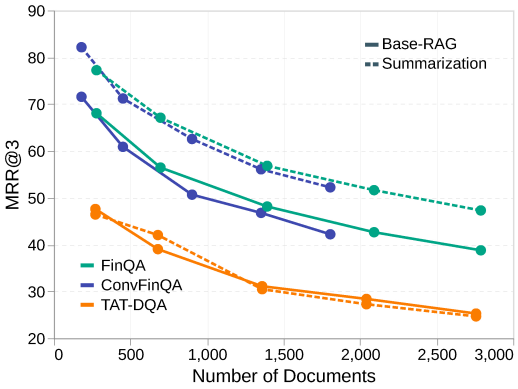

Figure 4: MRR@3 comparison for FinQA, ConvFinQA, and TAT-DQA across five evenly split document subsets.

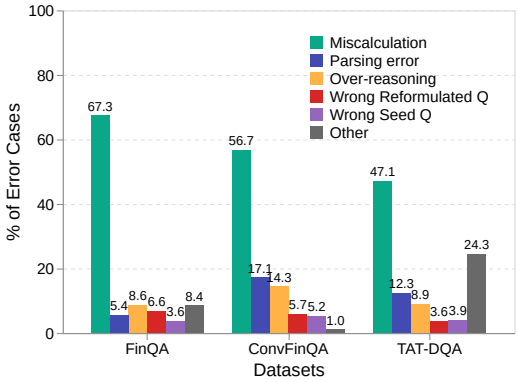

Figure 5: Results of the manual error analysis. Percentage of each error category per subset.

5.6 Manual Error Analysis

We performed a manual qualitative error analysis on 25% of the Oracle-Context errors from our main results, comprising 1,583 annotated cases across all subsets (see Figure 5). Each error was categorized into one of six categories: Miscalculation, Parsing error, Over-reasoning, Wrong reformulated question, wrong seed question, and Other (see Appendix K for more information).

The majority of error cases arise from arithmetic mistakes, parsing errors, or instances of unnecessary reasoning, indicating that models continue to struggle with reliably answering certain types of questions. A common failure involves inserting incorrect values into tables or producing arithmetic results that deviate slightly from the correct answer. This pattern is consistent across all three subsets, suggesting that such challenges persist irrespective of the underlying data source. Additionally, approximately 6% of errors in each subset are attributed to reformulation failures. In nearly 90% of these, the metric changed from 'value' to 'percentage value,' which confuses the generator. Approximately 5% of errors originate from unclear or ambiguous seed questions. Other errors include parsing issues and outputs with only NA values (especially for TAT-DQA), making diagnosis difficult. Overall, the benchmark remains challenging, with room for improvement in generation, but most questions remain suitable for evaluating RAG.

5.7 Main Takeaways

Overall, our results show that even the strongest RAG method examined (*Hybrid BM25*) falls short of Oracle-Context performance in NM by almost 30%. This performance gap underscores the benchmark's ability to quantify retrieval effectiveness and highlights the remaining challenges in achieving Oracle-level performance with RAG. Even when using other RAG methods like Hybrid BM25, the performance can only be improved by 2.5% on average on MRR and 5% in comparison to *Base-RAG*. We further analyzed the impact of other factors and find that even SOTA retrieval models achieve less than 50% MRR@5, highlighting that RAG on text-and-table data remains challenging; additionally, retrieval performance with 3K documents reveals that this task still offers significant room for improvement.

6 Conclusion

In this paper, we introduced our newly created benchmark, **T^2-RAGBench**, which contains 23,088 question-answer-context triples. It includes questions derived from over 7,318 documents and is designed to evaluate RAG methods for numerical reasoning over text-table data in the Unknown-Context Setting. While other datasets are defined in an Oracle-Context, our benchmark uses context-independent question making it possible to evaluate RAG methods. We demonstrate that our benchmark meets its intended goals through quantitative analysis and human validation. We test multiple RAG methods on the benchmark and find that *Hybrid BM25*, which combines dense and sparse retrieval, performs best. Additionally, we conducted ablation studies showing that current SOTA embedding models achieve low R@5 and MRR@5 scores on text-and table contexts. With T^2-RAGBench, we aim to facilitate the development of more RAG methods suitable for text-and-table documents, supporting the creation of real-world systems that can automatically analyze complex documents.

Limitations

This section outlines the key limitations related to the methodology and dataset that may affect the validity and generalizability of the presented results.

Lack of Human Verification and Authenticity. The questions used in the benchmark were generated synthetically, which can lead to distortions, as models do not inevitably generate the type of questions that real-world users would ask. Therefore, transferability to real systems may be affected. Although humans annotated the original question-answer pairs, there is no definitive guarantee that the generated questions will be formulated in a way that allows other models to answer them equivalently.

Another point is that a comprehensive verification process was only partly conducted on the benchmark questions. While we verified 100 samples per subset with four annotators in the benchmark, that the benchmark fulfills the requirements to be an evaluation dataset for our proposed task. Nevertheless, they can still be some questions that are not suitable to find the right context.

Domain-Specific Application. The presented work aims to present a benchmark that can test text-table datasets from different document types with different knowledge. Nevertheless, the dataset consists only of financial documents that have the same standardized structure, consistent terminology, and domain-specific content. As a result, the model's performance is tailored to this domain and can only be partly assumed to generalize to other types of document layouts or content types, such as medical reports, scientific publications, or administrative forms, where table-text relationships can vary significantly. Still, given the wide-ranging application of financial reporting standards, our work contributes to this specific domain.

Use of Quantized Models. Due to limited resources, all evaluations were conducted using quantized versions of the models, which enabled faster inference times and the execution of large open-source models. While quantization offers clear advantages in terms of computational efficiency, it often comes at the cost of reduced numerical precision and model accuracy. Therefore, the performance may be lower than that of full-precision SOTA models. However, since the focus of this paper is on comparing suitable RAG methods, we consider this negligible.

Ethical Considerations

This work introduces a benchmark dataset constructed from publicly available financial documents. All data used originates from previously published datasets (FinQA, ConvFinQA, TAT-DQA), which are either publicly accessible or sourced from publicly available company reports. No private, confidential, or personally identifiable information is included. The reformulated questions were synthetically generated using LLMs and subsequently validated by experts to ensure quality and context-independence. Human evaluation was conducted with informed consent and anonymized input. We acknowledge that while synthetic reformulation enhances benchmarking utility, it may not fully capture the natural distribution of user queries.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria. ICLR.

Dipali Baviskar, Swati Ahirrao, Vidyasagar Potdar, and Ketan V. Kotecha. 2021. Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions. *IEEE Access*, 9:72894–72936.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A Dataset for Long-form Financial Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047, Bangkok, Thailand. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics*, volume EMNLP 2020 of *Findings of ACL*, pages 1026–1036, Online Event. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. 2025. FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. *arXiv preprint*. ArXiv:2504.15800.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2025. MedOdyssey: A Medical Domain Benchmark for Long Context Evaluation Up to 200K Tokens. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 32–56, Albuquerque, New Mexico. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complementing Lexical Retrieval with Semantic Residual Embedding. *arXiv preprint*. ArXiv:2004.13969.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783.

Yulong Hui, YAO LU, and Huanchen Zhang. 2024. UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-World Document Analysis. In *Advances in Neural Information Processing Systems*, volume 37, pages 67200–67217. Curran Associates, Inc.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A Comprehensive Evaluation of Quantization Strategies for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. 2024. Robust Multi Model RAG Pipeline For Documents Containing Text, Table & Images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999, Salem, India. IEEE.

Yannis Katsis, Saneem A. Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2022. AIT-QA: Question Answering Dataset over Complex Tables in the Airline Industry. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 305–314, Hybrid: Seattle, Washington, USA + Online. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint*. ArXiv:2212.14024.

Kihun Kim, Mintae Kim, Hokyung Lee, Seong Ik Park, Youngsub Han, and Byoung-Ki Jeon. 2024. THoRR: Complex Table Retrieval and Refinement for RAG. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024) co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 3784 of *CEUR Workshop Proceedings*, pages 50–55, Washington DC, USA.

Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. 2023. Copy is All You Need. In

The Eleventh International Conference on Learning Representations, Kigali, Rwanda.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, virtual.

Xinze Li, Yixin Cao, Yubo Ma, and Aixin Sun. 2024. Long Context vs. RAG for LLMs: An Evaluation and Revisits. *arXiv preprint*. ArXiv:2501.01880.

Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. BioTABQA: Instruction Learning for Biomedical Table Question Answering. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 291–304, Bologna, Italy. CEUR-WS.org.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Irina Nikishina, Özge Sevgili, Mahei Manhai Li, Chris Biemann, and Martin Semmann. 2025. Creating a Taxonomy for Retrieval Augmented Generation Applications. *arXiv preprint*. ArXiv:2408.02854.

Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 1470–1480, Beijing, China. The Association for Computer Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*, Vancouver, BC, Canada.

Sachin Raja, Ajoy Mondal, and C. V. Jawahar. 2023. ICDAR 2023 Competition on Visual Question Answering on Business Document Images. In *Document Analysis and Recognition*, pages 454–470, Cham, Germany. Springer Nature Switzerland.

Carl M Rebman Jr, Queen E Booker, Hayden Wimmer, Steve Levkoff, Mark McMurtrey, and Loreen Marie Powell. 2023. An Industry Survey of Analytics Spreadsheet Tools Adoption: Microsoft Excel vs Google Sheets. *Information Systems Education Journal*, 21(5):29–42. Publisher: ERIC.

Sohini Roychowdhury, Marko Krema, Anvar Mahammad, Brian Moore, Arijit Mukherjee, and Punit Prakashchandra. 2024. ERATTA: Extreme RAG for Table To Answers with Large Language Models. *arXiv preprint*. ArXiv:2405.03963.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria. The Association for Computational Linguistics.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document Collection Visual Question Answering. In *16th International Conference on Document Analysis and Recognition*, volume 12822 of *Lecture Notes in Computer Science*, pages 778–792, Lausanne, Switzerland. Springer.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*. ArXiv: 2402.05672.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024b. Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025. TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. In *Association for the Advancement of Artificial Intelligence*, pages 25497–25506, Philadelphia, PA, USA. AAAI Press.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. *arXiv preprint*. ArXiv:2401.15884.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2020. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. *arXiv preprint*. ArXiv:2005.00589.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards Complex Document Understanding By Discrete Reasoning. In *MM '22: The 30th ACM International Conference on Multimedia*, pages 4857–4866, Lisboa, Portugal. ACM.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3277–3287, Virtual Event. Association for Computational Linguistics.

Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. 2025. FinTMMBench: Benchmarking Temporal-Aware Multi-Modal RAG in Finance. *arXiv preprint*. ArXiv:2503.05185.

A Dataset Samples

In the following, we give two examples for each dataset subset, including the original question, the reformulated question, and the corresponding context. Due to the limited page width, we had to wrap the text of the context.

Dataset / ID:
train_finqa2516

Question:
what is the growth rate in net revenue from 2010 to 2011?

Reformulated:
What was the percentage change in Entergy's net revenue from 2010 to 2011, considering the impact of the mark-to-market tax settlement sharing, retail electric price adjustments, and other factors as outlined in the 2011 financial discussion and analysis?

Context:

```
entergy louisiana , llc and subsidiaries management 2019s financial discussion and
analysis plan to spin off the utility 2019s transmission business see the 201cplan to spin
off the utility 2019s transmission business 201d section of entergy corporation and
subsidiaries management 2019s financial discussion and analysis for a discussion of this
matter , including the planned retirement of debt and preferred securities .results of
operations net income 2011 compared to 2010 net income increased $ 242.5 million primarily
due to a settlement with the irs related to the mark-to-market income tax treatment of
power purchase contracts , which resulted in a $ 422 million income tax benefit .the net
income effect was partially offset by a $ 199 million regulatory charge , which reduced
net revenue , because a portion of the benefit will be shared with customers .see note 3
to the financial statements for additional discussion of the settlement and benefit
sharing .2010 compared to 2009 net income decreased slightly by $ 1.4 million primarily
due to higher other operation and maintenance expenses , a higher effective income tax
rate , and higher interest expense , almost entirely offset by higher net revenue .net
revenue 2011 compared to 2010 net revenue consists of operating revenues net of : 1 ) fuel
, fuel-related expenses , and gas purchased for resale , 2 ) purchased power expenses ,
and 3 ) other regulatory charges ( credits ) .following is an analysis of the change in
net revenue comparing 2011 to 2010 .amount ( in millions ) ._|    |
| amount ( in millions )
||---:|:-------------------------------------|:-------------------------||  0 | 2010 net
revenue                      | $ 1043.7                ||  1 | mark-to-market tax
settlement sharing | -195.9 ( 195.9 )         ||  2 | retail electric price
| 32.5                     ||  3 | volume/weather                         | 11.6
||  4 | other                               | -5.7 ( 5.7 )             ||  5 | 2011 net
revenue                      | $ 886.2                 |_the mark-to-market tax
settlement sharing variance results from a regulatory charge because a portion of the
benefits of a settlement with the irs related to the mark-to-market income tax treatment
of power purchase contracts will be shared with customers , slightly offset by the
amortization of a portion of that charge beginning in october 2011 .see notes 3 and 8 to
the financial statements for additional discussion of the settlement and benefit sharing
.the retail electric price variance is primarily due to a formula rate plan increase
effective may 2011 .see note 2 to the financial statements for discussion of the formula
rate plan increase. .
```

Dataset / ID:
train_finqa518

Question:
at december 312008 what was the total liabilities acquired for this plan in millions

Reformulated:
As of December 31, 2008, what was the total amount of liabilities acquired by Republic Services for the BFI post-retirement healthcare plan, as disclosed in their 2008 consolidated financial statements?

Context:

```
estimated future pension benefit payments for the next ten years under the plan ( in
millions ) are as follows : estimated future payments: ._|    | 2009             |   $
14.9 ||---:|:-----------------|---------:||  0 | 2010             |     15.9 ||  1 |
2011             |     16.2 ||  2 | 2012             |     19.2 ||  3 | 2013
|     21.9 ||  4 | 2014 through 2018 |    142.2 |_bfi post retirement healthcare plan we
acquired obligations under the bfi post retirement healthcare plan as part of our
acquisition of allied .this plan provides continued medical coverage for certain former
employees following their retirement , including some employees subject to collective
bargaining agreements .eligibility for this plan is limited to certain of those employees
who had ten or more years of service and were age 55 or older as of december 31 , 1998 ,
and certain employees in california who were hired on or before december 31 , 2005 and who
retire on or after age 55 with at least thirty years of service .liabilities acquired for
this plan were $ 1.2 million and $ 1.3 million , respectively , at the acquisition date
and at december 31 , 2008 .multi-employer pension plans we contribute to 25 multi-employer
pension plans under collective bargaining agreements covering union- represented employees
.we acquired responsibility for contributions for a portion of these plans as part of our
acquisition of allied .approximately 22% ( 22 % ) of our total current employees are
participants in such multi- employer plans .these plans generally provide retirement
benefits to participants based on their service to contributing employers .we do not
administer these multi-employer plans .in general , these plans are managed by a board of
trustees with the unions appointing certain trustees and other contributing employers of
the plan appointing certain members .we generally are not represented on the board of
trustees .we do not have current plan financial information from the plans 2019
administrators , but based on the information available to us , it is possible that some
of the multi-employer plans to which we contribute may be underfunded .the pension
protection act , enacted in august 2006 , requires underfunded pension plans to improve
their funding ratios within prescribed intervals based on the level of their underfunding
.until the plan trustees develop the funding improvement plans or rehabilitation plans as
required by the pension protection act , we are unable to determine the amount of
assessments we may be subject to , if any .accordingly , we cannot determine at this time
the impact that the pension protection act may have on our consolidated financial position
, results of operations or cash flows .furthermore , under current law regarding multi-
employer benefit plans , a plan 2019s termination , our voluntary withdrawal , or the mass
withdrawal of all contributing employers from any under-funded , multi-employer pension
plan would require us to make payments to the plan for our proportionate share of the
multi- employer plan 2019s unfunded vested liabilities .it is possible that there may be a
mass withdrawal of employers contributing to these plans or plans may terminate in the
near future .we could have adjustments to our estimates for these matters in the near term
that could have a material effect on our consolidated financial condition , results of
operations or cash flows .our pension expense for multi-employer plans was $ 21.8 million
, $ 18.9 million and $ 17.3 million for the years ended december 31 , 2008 , 2007 and 2006
, respectively .republic services , inc .and subsidiaries notes to consolidated financial
statements %%transmsg*** transmitting job : p14076 pcn : 133000000 ***%%pcmsg|131
|00027|yes|no|02/28/2009 21:12|0|0|page is valid , no graphics -- color : d| .
```

Dataset / ID:
TatQA 8e642bdce983286cbaffa9661d24157a

Question:
What was the total intrinsic value of RSUs which vested during 2019?

Reformulated:
What was the total intrinsic value of RSUs that vested during the year ended March 31, 2019, for Microchip Technology Inc.?

Context:

Microsemi Acquisition-related Equity AwardsIn connection with its acquisition of Microsemi on May 29, 2018, the Company assumed certain restricted stock units (RSUs), stock appreciation rights (SARs), and stock options granted by Microsemi. The assumed awards were measured at the acquisition date based on the estimated fair value, which was a total of $175.4 million. A portion of that fair value, $53.9 million, which represented the pre-acquisition vested service provided by employees to Microsemi, was included in the total consideration transferred as part of the acquisition. As of the acquisition date, the remaining portion of the fair value of those awards was $121.5 million, representing post-acquisition share-based compensation expense that will be recognized as these employees provide service over the remaining vesting periods. During the year ended March 31, 2019, the Company recognized $65.2 million of share-based compensation expense in connection with the acquisition of Microsemi, of which $3.5 million was capitalized into inventory and $17.2 million was due to the accelerated vesting of outstanding equity awards upon termination of certain Microsemi employees.Atmel Acquisition-related Equity AwardsIn connection with its acquisition of Atmel on April 4, 2016, the Company assumed certain RSUs granted by Atmel. The assumed awards were measured at the acquisition date based on the estimated fair value, which was a total of $95.9 million. A portion of that fair value, $7.5 million, which represented the pre-acquisition vested service provided by employees to Atmel, was included in the total consideration transferred as part of the acquisition. As of the acquisition date, the remaining portion of the fair value of those awards was $88.4 million, representing post-acquisition share-based compensation expense that will be recognized as these employees provide service over the remaining vesting periods.Combined Incentive Plan InformationRSU share activity under the 2004 Plan is set forth below:| | Number of Shares | Weighted Average Grant Date Fair Value ||----------|----------|----------|| Nonvested shares at March 31, 2016 | 6,307,742 | $36.76 || Granted | 1,635,655 | 51.46 || Assumed upon acquisition | 2,059,524 | 46.57 || Forfeited | (722,212) | 43.58 || Vested | (2,861,253) | 38.60 || Nonvested shares at March 31, 2017 | 6,419,456 | 42.06 || Granted | 1,267,536 | 77.26 || Forfeited | (279,051) | 49.65 || Vested | (1,735,501) | 38.00 || Nonvested shares at March 31, 2018 | 5,672,440 | 50.79 || Granted | 1,951,408 | 77.83 || Assumed upon acquisition | 1,805,680 | 91.70 || Forfeited | (408,242) | 73.36 || Vested | (2,729,324) | 61.51 || Nonvested shares at March 31, 2019 | 6,291,962 | $64.81 |The total intrinsic value of RSUs which vested during the years ended March 31, 2019, 2018 and 2017 was $229.3 million, $146.0 million and $166.1 million, respectively. The aggregate intrinsic value of RSUs outstanding at March 31, 2019 was $522.0 million, calculated based on the closing price of the Company’s common stock of $82.96 per share on March 29, 2019. At March 31, 2019, the weighted average remaining expense recognition period was 1.91 years.

Dataset / ID:
TatQA a210c0538af4df5f8881dcb8f1bf00ff

Question:
What was the Accrued compensation and employee benefits in 2018?

Reformulated:
What was the accrued compensation and employee benefits for Jabil Circuit Inc. as of August 31, 2018?

Context:

```
Intangible asset amortization for fiscal years 2019, 2018 and 2017 was approximately $31.9
million, $38.5 million and $35.5 million, respectively. The estimated future amortization
expense is as follows (in thousands):| Fiscal Year Ended August 31,
|     ||-----------------------------------------------------|-----|| 2020
..................................................................... $ 54,165 ||
2021 ..................................................................... 43,780
|| 2022 .....................................................................
28,291 || 2023
..................................................................... 25,877 ||
2024 ..................................................................... 10,976
|| Thereafter .....................................................................
43,174 || **Total
.....................................................................** $206,263 |7.
Accrued ExpensesAccrued expenses consist of the following (in thousands):|
| August 31, 2019 | August 31, 2018
||-------------------------|-----------------|-----------------|| Contract liabilities
| $ 511,329       | −               || Deferred income         | −               | 691,365
|| Accrued compensation    | 600,907         | 570,400         || and employee benefits  |
|                 || Obligation             | 475,251         | −               ||
associated with        |                 |                 || securitization         |
|                 || programs               |                 |                 || Other
accrued expenses | 1,402,657       | 1,000,979       || **Accrued expenses**    |
$2,990,144      | $2,262,744      |8. Notes Payable and Long-Term DebtNotes payable and
long-term debt outstanding as of August 31, 2019 and 2018 are summarized below (in
thousands):|                         | August 31, 2019 | August 31, 2018
||------------------------|-----------------|-----------------|| 5.625% Senior Notes    |
398,886         | 397,995         || (1)(2)                  | Dec 15, 2020    |
|| 4.700% Senior Notes    | 498,004         | 497,350         || (1)(2)                 |
Sep 15, 2022    |                 || 4.900% Senior Notes    | 299,057         | 298,814
|| (1)                    | Jul 14, 2023    |                 || 3.950% Senior Notes    |
494,825         | 494,208         || (1)(2)(3)               | Jan 12, 2028    |
|| Borrowings under      |                 |                 || credit facilities(4)   |
|                 || (5)(6)                 | Nov 8, 2022 and|                 ||
Borrowings under      |                 |                 || loans(4)(5)            |
|                 || (4)                    |                 |                 || Total
notes payable    | 2,496,465       | 2,518,699       || and long-term debt     |
|                 || (1)                    |                 |                 || Less
current          | 375,181         | 25,197          || installments of notes  |
|                 || payable and long-term  |                 |                 || debt
|                 |                 || (2)                    |                 |
|| Total notes payable    | $2,121,284      | $2,493,502      || and long-term debt,    |
|                 || less current install-  |                 |                 || ments
|                 |                 |(1) The notes are carried at the principal amount of
each note, less any unamortized discount and unamortized debt issuance costs.(2) The
Senior Notes are the Company's senior unsecured obligations and rank equally with all
other existing and future senior unsecured debt obligations.(3) During the fiscal year
ended August 31, 2018, the Company issued $500.0 million of publicly registered 3.950%
Senior Notes due 2028 (the "3.950% Senior Notes"). The net proceeds from the offering were
used.
```

Dataset / ID:
convfinqa_1119

Question:
what was the change in percentage points of data center cost between the years of 2014-13 and 2013-12?

Reformulated:
What was the percentage point decrease in data center cost growth between fiscal 2013-2012 and fiscal 2014-2013 for Adobe Inc.?

Context:

subscription cost of subscription revenue consists of third-party royalties and expenses related to operating our network infrastructure , including depreciation expenses and operating lease payments associated with computer equipment , data center costs , salaries and related expenses of network operations , implementation , account management and technical support personnel , amortization of intangible assets and allocated overhead . we enter into contracts with third-parties for the use of their data center facilities and our data center costs largely consist of the amounts we pay to these third parties for rack space , power and similar items . cost of subscription revenue increased due to the following : % (%) change 2014-2013 % (%) change 2013-2012 .| | % (%) change2014-2013 | % (%) change2013-2012 || --- | --- | --- || data center cost | 10% (10 %) | 11% (11 %) || compensation cost and related benefits associated with headcount | 4 | 5 || depreciation expense | 3 | 3 || royalty cost | 3 | 4 || amortization of purchased intangibles | 2014 | 4 || various individually insignificant items | 1 | 2014 || total change | 21% (21 %) | 27% (27 %) |cost of subscription revenue increased during fiscal 2014 as compared to fiscal 2013 primarily due to data center costs , compensation cost and related benefits , deprecation expense , and royalty cost . data center costs increased as compared with the year-ago period primarily due to higher transaction volumes in our adobe marketing cloud and creative cloud services . compensation cost and related benefits increased as compared to the year-ago period primarily due to additional headcount in fiscal 2014 , including from our acquisition of neolane in the third quarter of fiscal 2013 . depreciation expense increased as compared to the year-ago period primarily due to higher capital expenditures in recent periods as we continue to invest in our network and data center infrastructure to support the growth of our business . royalty cost increased primarily due to increases in subscriptions and downloads of our saas offerings . cost of subscription revenue increased during fiscal 2013 as compared to fiscal 2012 primarily due to increased hosted server costs and amortization of purchased intangibles . hosted server costs increased primarily due to increases in data center costs related to higher transaction volumes in our adobe marketing cloud and creative cloud services , depreciation expense from higher capital expenditures in prior years and compensation and related benefits driven by additional headcount . amortization of purchased intangibles increased primarily due to increased amortization of intangible assets purchased associated with our acquisitions of behance and neolane in fiscal 2013 . services and support cost of services and support revenue is primarily comprised of employee-related costs and associated costs incurred to provide consulting services , training and product support . cost of services and support revenue increased during fiscal 2014 as compared to fiscal 2013 primarily due to increases in compensation and related benefits driven by additional headcount and third-party fees related to training and consulting services provided to our customers . cost of services and support revenue increased during fiscal 2013 as compared to fiscal 2012 primarily due to increases in third-party fees related to training and consulting services provided to our customers and compensation and related benefits driven by additional headcount , including headcount from our acquisition of neolane in fiscal 2013. .

926

Dataset / ID:
convfinqa_2966

Question:
what was the value of free cash flow in 2009?

Reformulated:
What was the free cash flow of Union Pacific Corporation in 2009, as calculated from cash provided by operating activities, less cash used in investing activities and dividends paid?

Context:

2022 asset utilization 2013 in response to economic conditions and lower revenue in 2009 , we implemented productivity initiatives to improve efficiency and reduce costs , in addition to adjusting our resources to reflect lower demand . although varying throughout the year , our resource reductions included removing from service approximately 26% (26 %) of our road locomotives and 18% (18 %) of our freight car inventory by year end . we also reduced shift levels at most rail facilities and closed or significantly reduced operations in 30 of our 114 principal rail yards . these demand-driven resource adjustments and our productivity initiatives combined to reduce our workforce by 10% (10 %) . 2022 fuel prices 2013 as the economy worsened during the third and fourth quarters of 2008 , fuel prices dropped dramatically , reaching $ 33.87 per barrel in december 2008 , a near five-year low . throughout 2009 , crude oil prices generally increased , ending the year around $ 80 per barrel . overall , our average fuel price decreased by 44% (44 %) in 2009 , reducing operating expenses by $ 1.3 billion compared to 2008 . we also reduced our consumption rate by 4% (4 %) during the year , saving approximately 40 million gallons of fuel . the use of newer , more fuel efficient locomotives ; increased use of distributed locomotive power ; fuel conservation programs ; and improved network operations and asset utilization all contributed to this improvement . 2022 free cash flow 2013 cash generated by operating activities totaled $ 3.2 billion , yielding free cash flow of $ 515 million in 2009 . free cash flow is defined as cash provided by operating activities , less cash used in investing activities and dividends paid . free cash flow is not considered a financial measure under accounting principles generally accepted in the united states (gaap) by sec regulation g and item 10 of sec regulation s-k . we believe free cash flow is important in evaluating our financial performance and measures our ability to generate cash without additional external financings . free cash flow should be considered in addition to , rather than as a substitute for , cash provided by operating activities . the following table reconciles cash provided by operating activities (gaap measure) to free cash flow (non-gaap measure) : millions of dollars 2009 2008 2007 .| millions of dollars | 2009 | 2008 | 2007 || --- | --- | --- | --- || cash provided by operating activities | $ 3234 | $ 4070 | $ 3277 || cash used in investing activities | -2175 (2175) | -2764 (2764) | -2426 (2426) || dividends paid | -544 (544) | -481 (481) | -364 (364) || free cash flow | $ 515 | $ 825 | $ 487 |2010 outlook 2022 safety 2013 operating a safe railroad benefits our employees , our customers , our shareholders , and the public . we will continue using a multi-faceted approach to safety , utilizing technology , risk assessment , quality control , and training , and by engaging our employees . we will continue implementing total safety culture (tsc) throughout our operations . tsc is designed to establish , maintain , reinforce , and promote safe practices among co-workers . this process allows us to identify and implement best practices for employee and operational safety . reducing grade-crossing incidents is a critical aspect of our safety programs , and we will continue our efforts to maintain , upgrade , and close crossings ; install video cameras on locomotives ; and educate the public about crossing safety through our own programs , various industry programs , and other activities . 2022 transportation plan 2013 to build upon our success in recent years , we will continue evaluating traffic flows and network logistic patterns , which can be quite dynamic from year-to-year , to identify additional opportunities to simplify operations , remove network variability and improve network efficiency and asset utilization . we plan to adjust manpower and our locomotive and rail car fleets to .

B Data Preparation

FinQA. The FinQA dataset is based on human-annotated questions about documents from FinTabNet, a large corpus of PDF files containing annual reports of S&P 500 companies. In addition to existing data, company-specific information such as founding year, sector, and report year was added. Since the answers consisted either of formulas or numerical values, all formulas were parsed and converted into numerical values, as discrepancies between formulas and their numerical solutions were observed. Moreover, approximately 150 yes/no questions were normalized by converting their answers to 0 and 1, respectively.

ConvFinQA. The ConvFinQA dataset is also based on FinTabNet and was enriched with additional metadata. Similar to FinQA, answers were standardized by converting formulas and numeric responses into a uniform format. To reduce task complexity and eliminate potential confounding factors, only the first question from each conversation was included. This reduced the dataset size from 14,115 to 3,458 QA pairs.

TAT-DQA. TAT-DQA is an independent dataset based on publicly available financial reports. The original dataset included four answer types: Span, Multi-span, Arithmetic, and Count. To ensure consistency with other datasets focused solely on numerical reasoning and to maintain uniform evaluation prompts, Multi-span questions were removed. Additionally, Span answers were normalized by removing symbols such as $ and %, and converting words like "million" or "billion" into their numeric equivalents. Dates were also reformatted to the US standard. After these filtering steps, the dataset size was reduced from 16,558 to 11,349 QA pairs.

C Reformat Prompt

The prompt for reformulating the questions to be context-independent is given in Figure 6

```
## System Prompt
You are a financial education assistant. Your task is to **rephrase a question** based on a specific
table from a financial document. The goal is to ensure that the question:
- Refers to details that **only make sense in this specific context**
- **Does not use generic phrases** like "based on the data above" or "according to the table"
- Is **not answerable** with any other financial document or context
- Keeps the **original answer correct**
- Sounds natural, precise, and unambiguous
- Try to cut of unnecessary words and phrases
You will also be provided with **metadata** from the document (e.g., company name, report
title, year, section).
Use this metadata to ground the question further in context.
The explanation must:
- Describe the **reasoning steps** required to reach the answer
- Refer to **specific values, labels, rows, or relationships** in the table
- Show that the answer is uniquely valid for this table and **tied to the metadata/context**
### Output Format:
Question:
Answer:
Explanation:
```

Figure 6: System prompt to reformulate the questions.

D Annotation Tool

The annotations by financial experts were performed with a simple web tool shown in Figure 7. For each question, the annotator can see the original question, the reformulated question, and the context as given in the dataset. The annotators were guided by the following explanations. Annotation Guide: Label the question as 'Context-depending' if the answer depends on the context and can be answered in another context with another true answer, otherwise, label it as 'Unambiguous', when there is only one true answer.

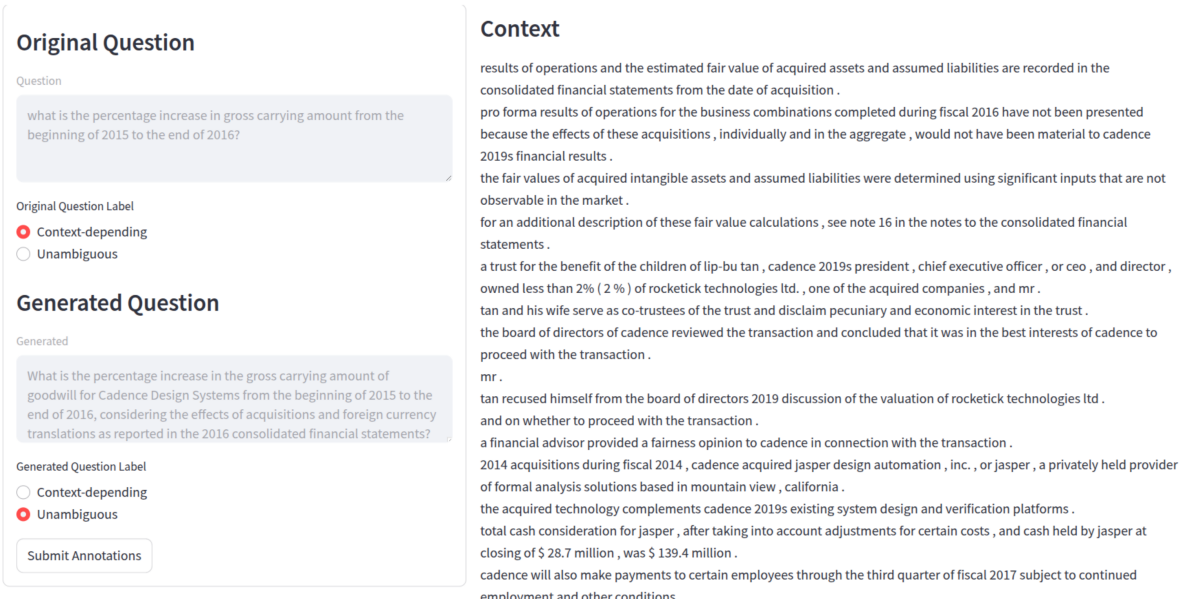

Figure 7: Annotation tool for labeling reformulated questions.

E Annotation Samples for Disagreement

The following six examples illustrate the cases where the commentators disagreed and show where they disagreed. In addition, less than 10% of the examples were commented on differently.

ConvFinQA

Original (convfinqa_10477):
what was the investment on the alcoainc. in 2014?

Reformulated (convfinqa_5653):
What was the goodwill balance for Cadence Design Systems as of December 30, 2017, following the business combinations and foreign currency translations during fiscal 2017?

FinQA

Original (train_finqa1426):
as of december 312016 what was the ratio of the approximate number of residential vehicles to the large-container industrial?

Reformulated (train_finqa1183):
What was the percent change in Entergy's net revenue from 2013 to 2014, as reported in the 2015 financial discussion and analysis for Entergy Corporation and Subsidiaries?

TAT-DQA

Original (788a22ceb71d2db8786f136e6dd1eed0):
What was the total value of the changes in principal on the issuance of 2024 Notes, 2026 Notes, 2027 Notes, 2029 Notes, and 2030 Notes?

Reformulated (9636d16b010a57a424ab8c02d0f9e46b):
What percentage of the Australian Prime Storage Fund did National Storage REIT own as at 30 June 2018?

F Retrieval Template

The prompt used to encode the question in the retrieval step is given in Figure 8

```
Given a question about a company, retrieve relevant passages that answer the query.
Question:{question}
```

Figure 8: System prompt for the retrieval step.

G System Prompt for Generation

We use the same prompt for generating answers (the Generation step in RAG) for all methods we compared. The generation prompt is given in Figure 9-11.

```
YOU ARE A FINANCIAL REASONING EXPERT TRAINED TO ANALYZE A QUESTION AND ITS ASSOCIATED CONTEXT
IN A SINGLE PASS.

  YOUR TASK IS TO:
  - INTERNALLY: READ the question and accompanying financial table/text
    1. UNDERSTAND what the question is asking
    2. IDENTIFY numeric values from the context
    3. CONSTRUCT a valid mathematical FORMULA using a strict symbolic syntax
    4. EVALUATE the formula if it contains only constants
  - FINALLY: OUTPUT one JSON object that includes reasoning, the formula, and the computed result

  THERE IS ONLY ONE INPUT AND ONE OUTPUT. DO ALL THINKING INTERNALLY.
  ---
  FORMULA SYNTAX RULES:

  A formula is either:
  - A number (e.g., 7, 3.14)
  - One of the following symbolic operations, each with exactly two arguments:
    - add(f1, f2)
    - subtract(f1, f2)
    - multiply(f1, f2)
    - divide(f1, f2)
    - exp(f1, f2)
    - greater(f1, f2)

  Nesting is allowed. All values must come from the provided context.
  ---
  PERCENTAGE HANDLING RULES:

  - IF the question asks for a **percentage**, you MUST:
    - REPRESENT the result in the `final_formula` as a **decimal between 0 and 1**
    - COMPUTE the actual percentage internally using divide(part, total)
    - DO NOT multiply by 100 — keep `computed_formula` also between 0 and 1
  - IF a percentage is given in the context (e.g., "12.5%"):
    - CONVERT it to a decimal using divide(12.5, 100) **before using it in a formula**
  - EVEN IF the question says "how much percentage...", your output stays in **0 to 1 scale**
    - Example: A 12.5% result = "computed_formula": "0.125"
  ---
  OUTPUT FORMAT:
  {
    "reasoning_steps": ["<short bullet 1>", "<short bullet 2>", "..."],
    "final_formula": "<valid formula or 'None'>",
    "computed_formula": "<decimal result as string or 'N/A'>"
  }
  ---
  EXAMPLES:
  EXAMPLE 1 (compute percentage from raw values):

  Input Question:
  What percentage of restricted shares is set to vest after 2021?

  Input Context:
  | Year         | Vesting Count |
  |--------------|---------------|
  | 2021         | 199850        |
  | thereafter   | 110494        |
  | total        | 9038137       |
```

Figure 9: System prompt to answer the questions (1/3).

```
Output:
{
  "reasoning_steps": [
    "Located total outstanding restricted shares = 9038137",
    "Found restricted shares vesting after 2021 = 110494",
    "Computed percentage = divide(110494, 9038137)"
  ],
  "final_formula": "divide(110494, 9038137)",
  "computed_formula": "0.01222458878059346"
}

---

EXAMPLE 2 (compute profit margin — also a percentage):

Input Question:
What was the profit margin for 2022?

Input Context:
| Year | Revenue   | Net Income |
|------|-----------|------------|
| 2022 | 5000000   | 750000     |

Output:
{
  "reasoning_steps": [
    "Identified revenue for 2022 = 5000000",
    "Identified net income for 2022 = 750000",
    "Computed profit margin = divide(750000, 5000000)"
  ],
  "final_formula": "divide(750000, 5000000)",
  "computed_formula": "0.15"
}

---

EXAMPLE 3 (must compute % even if context contains a % value):

Input Question:
How much percentage of revenue was allocated to R&D in 2022?

Input Context:
| Category      | Amount ($)  |
|---------------|-------------|
| Revenue       | 5000000     |
| R&D Expense   | 625000      |

Output:
{
  "reasoning_steps": [
    "Found R&D expense = 625000 and revenue = 5000000",
    "Computed R&D percentage as decimal = divide(625000, 5000000)"
  ],
  "final_formula": "divide(625000, 5000000)",
  "computed_formula": "0.125"
}

---
```

Figure 10: System prompt to answer the questions (2/3).

```
UNCLEAR DATA EXAMPLE:

Input Question:
What is the average interest coverage ratio?

Input Context:
No interest expense or earnings values provided.

Output:
{
  "reasoning_steps": [],
  "final_formula": "None",
  "computed_formula": "N/A"
}

---

STRICT RULES (DO NOT VIOLATE):

- DO NOT include %, $, €, "million", or any other unit
- DO NOT guess values or invent data
- DO NOT return text, markdown, or extra formatting
- DO NOT multiply by 100 — all percentages must remain in 0–1 decimal form
- DO NOT use invalid function names or wrong number of arguments
- DO NOT return "answer": keys — use only final_formula and computed_formula
- DO NOT include any formulas or operators in the computed_formula
- IF a % is provided in the context, convert it to a decimal with divide(X, 100) if needed
```

Figure 11: System prompt to answer the questions (3/3).

H HyDE Prompt

The prompt used to generate hypothetical documents for the HyDE method is given in Figure 12

```
You are a financial analyst. Given a financial question, generate a detailed and realistic
hypothetical financial document using typical language and structure found in financial reports and
documents.
Your answer may include plausible numerical values, trends, and terminology, as if it came from an
actual financial report.
The goal is to produce a text that matches the type of content found in financial documents containing
both text and tables, to aid dense retrieval.
```

Figure 12: Prompt for the HyDE method.

I Summarizing Prompt

The prompt used to generate summarizations for the *Summarization* and *SumContext* methods is given in Figure 13.

```
You are a helpful assistant. Your task is to summarize the context text that the user provides
for better performance in a RAG system.
Pay special attention to all the numerical information, especially those contained in tables.
The summary does not necessarily have to contain all the numerical information, but from
reading the summary, one should be able to tell what information are contained in the text.
When you receive the context text from the user, ONLY output the summarized text WITHOUT any
extra reasoning or prefix / postfix text.
```

Figure 13: Summarization prompt.

J Retrieval Models Source

Model	Size	Source
Stella-EN-1.5B	1B	NovaSearch/stella_en_1.5B_v5
GTE-Qwen2 1.5B Instruct	1B	Alibaba-NLP/gte-Qwen2-1.5B-instruct
Multilingual E5-Instruct	560M	intfloat/multilingual-e5-large-instruct
Gemini: Text-Embedding-004	unknown	Google Gemini API
OpenAI: Text-Embedding-3 Large	unknown	OpenAI API Documentation

Table 5: Model sizes and sources of evaluated embedding models.

K Error Analysis

To better understand the model's failure cases, we conducted a manual error analysis on the Oracle-Context setting where the LLaMA 3.3 70B model was used. On average, the model answered 72.7% of the questions correctly across all subsets. We define the remaining 27.3% of questions as *error cases*. From these, we randomly sampled 25% to reduce annotation effort, resulting in a total of 1,583 examples for manual inspection. To derive meaningful error categories, we began by annotating a small subset of 20 examples from each data split freely. This exploratory step allowed us to identify recurring patterns in the model's failure modes. Based on this qualitative analysis, we established a set of consistent error categories, which are summarized in Table 6. Many of the observed errors were systematic and repeated across examples, indicating that our sampled subset provides a representative estimate of the broader error distribution.

Category	Description and Example
Miscalculation	Basic arithmetic mistake (e.g., sum, difference, average). *Example:* subtract(196545, 176675) = 19870, but model returned 19670.
Parsing error	Incorrect extraction of values from table (wrong row/column). *Example:* Summed wrong entries or picked incorrect column values.
Over-reasoning	Performed unnecessary computation instead of direct lookup. *Example:* Answer in plain text, but model tried to compute.
Wrong Reformulated Question	Reformulation subtly changed the metric. *Example:* Original asks for *sum*, reformulation asks for *average*.
Wrong Seed Question	Original query in seed dataset is unanswerable. *Example:* Asked for 2016/17 data when table ends at 2015.
Other	Cases where the answer was *NA*, JSON was parsed incorrectly, or other unclear issues. *Example:* Empty answer, malformed input, or ambiguous logic.

Table 6: Error categories for model failures with updated labels.