Synthetic Data Generation Using Large Language Models for Financial Question Answering

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha,
Sai Akhil Puranam, Shashishekar Ramakrishna
EY Global Delivery Services India LLP
{Chetan.Harsha,Karmvir.Phogat,Sridhar.Dasaratha}@gds.ey.com,
{Sai.Puranam,Shashishekar.R}@gds.ey.com

Abstract

Recent research has shown excellent performance of large language models (LLMs) for answering questions requiring multi-step financial reasoning. While the larger models have been used with zero-shot or few-shot prompting, the smaller variants need fine-tuning on training data containing questions and the corresponding answers that includes detailed reasoning demonstrations. To alleviate the significant cost of creating a data set with complex questions and corresponding answers, we explore the use of synthetic data for financial question answering using a multi-step LLM based approach to generate question as well as the answers with reasoning steps. We consider standard as well as conversational financial question answering scenarios.

We experiment with synthetic data generation for three different real financial reasoning problems that already have manually collected data sets created with the help of financial experts. Using the same document sources, we use the proposed LLM based approach to generate synthetic questions and answers. To measure the effectiveness, we train multiple small language models (SLMs) on these synthetic data and compare the performance with that of the same SLMs trained on the real data. We further perform extensive experimental analysis generating important evidence on the potential of using synthetic data in financial reasoning tasks.

1 Introduction

Developing machine learning systems for answering questions in the financial domain is a challenging problem. These systems must be capable of complex multi-step reasoning using real-world financial data. In recent years, the creation of large-scale financial question-answering datasets has led to significant improvements in this specialized domain (Chen et al., 2021). Nonetheless, assembling these datasets is a complicated, labor-intensive, and costly process, requiring the expertise of skilled annotators (Zhao et al., 2022).

As LLMs continue to advance, researchers have explored their potential to address these financial reasoning problems. Using methods that rely on an LLM to encode the reasoning steps into python programs which are then executed by external Python interpreters, state-of-the-art results have been obtained (Chen et al., 2023). While these extremely large models offer the benefit of easy use through prompting and eliminate the need for large-scale manual data set curation, their deployment at scale is expensive due to significant computational and inference costs.

To alleviate the reliance on extremely large models, recent research has focused on fine-tuning SLMs using data containing reasoning demonstrations that are generated using a large model (Magister et al., 2023). Promising results on various tasks including arithmetic, symbolic, common-sense reasoning (Ho et al., 2023) and financial reasoning (Phogat et al., 2024) have been achieved. However, for tasks without any existing data, these methods still rely on the time-consuming and expensive manual collection of data.

Synthetic data generation via zero-shot or few-shot LLM prompting provides an appealing alternative to manual data creation, as demonstrated in recent studies (Wang et al., 2023; Peng et al., 2023; Ye et al., 2022; Wang et al., 2021; Tang et al., 2023; Gou et al., 2021). While conceptually simple, achieving both high correctness and diversity in synthetic data sets is challenging (Gandhi et al., 2024), with current methods showing variable success (Ding et al., 2023).

In the realm of question answering (QA), the generation of synthetic QA data from text has been previously investigated (Li and Tajbakhsh, 2023; Wu et al., 2024; Schmidt et al., 2024) with promising results. These studies have focused on question generation requiring deep semantic comprehension,

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 76–95
January 19–20, 2025.

as opposed to questions that demand numerical analysis. Currently, there is a scarcity of studies examining the use of LLMs to create high-quality datasets specifically tailored for financial reasoning tasks.

In this work, we undertake a detailed and methodical inquiry into the effectiveness of LLMs driven synthetic financial reasoning data generation from financial documents. We focus on studying zero-shot prompting both with and without example questions. In addition to a standard scenario that requires the creation of a single question from a provided financial text passage, we consider the creation of a set of questions representative of a conversation over a financial document. The conversational scenario challenges the LLM to create a series of inter-related questions that are coherent, require context tracking and reference resolution across the questions.

We conduct thorough experiments in both scenarios to evaluate the ability of LLM-based techniques for creating questions demanding multi-hop numerical reasoning and their detailed answers with reasoning steps. In the standard scenario, we design a zero-shot prompt with constraints to direct the type of question generation, sometimes adding actual examples. Answers, formatted as Python code encoding the required calculations, are produced separately and then screened to remove any incorrect pairs. The filtering process excludes pairs with codes that are non-executable or yield outputs in incorrect formats, without evaluating the data's domain-specific correctness. For conversational data synthesis, we include an additional instruction that directs the LLM to formulate a sequence of questions conversationally.

Our key contributions are outlined below:

- We evaluate synthetic data generation by comparing the performance of three SLMs fine-tuned on synthetic data with those fine-tuned on three real-world financial QA datasets: FinQA (Chen et al., 2021), TATQA (Zhu et al., 2021), and ConvFinQA (Chen et al., 2022).

- We explore two approaches for generating conversational financial QA data and assess their effectiveness for different conversational flow types.

- We examine the influence of synthetic data volume on model performance and generalization abilities, as well as the SLMs' sensitivity to the synthetic data's similarity to the actual datasets.

Our results indicate that models trained on synthetic data nearly match the performance of those trained on real data for standard financial QA. Synthetic data sets yield acceptable results for conversational financial QA, though they fall short of real data's effectiveness. Additionally, two key results hint at better generalization of models fine-tuned with synthetic data (1) SLM fine-tuned on synthetic data outperformed the same model trained on real data when evaluated on a similar but independent test data set (2) SLM trained on a dataset deliberately crafted to have low similarity to the real data performed on par with the same model trained on data with higher similarity. These findings highlight interesting characteristics of synthetic financial reasoning data that merit further investigation.

2 Related Work

LLM generated synthetic data has been shown to be effective in multiple domains (Liu et al., 2024). (Li et al., 2023) study synthetic text classification data generation by zero-shot and few-shot prompting of an LLM, finding the effectiveness to be task dependent. (Chan et al., 2024) classify synthetic data generation into three types: answer augmentation, question rephrasing, and new question creation from real samples, noting that their performance varies with the problem. For mathematical reasoning tasks, data augmentation has been shown to be effective (Luo et al., 2023; Yu et al., 2024). Further, on mathematical tasks models have been shown to benefit from scaling the training data using synthetic data (Li et al., 2024). In (Takahashi et al., 2023) an instruct tuned model is used to generate synthetic QA pairs from Japanese wiki articles, news and contexts from JSquad. These prior studies do not focus on generating synthetic data for numerical multi-hop reasoning over financial reports.

For financial question answering, (Chen et al., 2021, 2022; Zhu et al., 2021) create data sets that support the development of multi-step financial reasoning systems. (Phogat et al., 2024) enhance these data with reasoning demonstrations generated by an LLM and fine-tune SLMs using these data sets, demonstrating significant improvement in SLM performance. We use the same real datasets primarily as a baseline for evaluating the synthetic variants of these datasets, which we generate us-

ing only LLMs. More recently, FinLLMs (Yuan et al., 2024) provide a method to generate synthetic data starting with a compilation of a list of common financial formulas, while (Hwang et al., 2023) generate new contexts for questions in an existing financial dataset to augment the training data. In contrast, we use LLMs to directly generate question answer pairs for both standard and conversational settings, from financial reports without providing any additional financial knowledge.

3 Methodology

We now present the procedure for generating synthetic multi-hop financial reasoning question answer pairs from financial document excerpts utilizing LLMs. For these problems, we choose to generate the answer in the form of python code that encodes the reasoning to solve the generated question. The python code is executed using an external Python interpreter to generate the actual answer to the financial question. As shown in previous work (Gao et al., 2023; Chen et al., 2023), for numerical reasoning, a code generation and external execution strategy is more effective than methods requiring the language model to perform the computation.

Our approach encompasses two distinct data generation strategies tailored to different settings: First, the *Financial QA* setting in which the LLM is prompted to generate financial questions from document excerpts. Second, the *Conversational Financial QA* setting in which a sequence of interconnected sub-questions are generated that collectively lead to the resolution of a complex financial query. In both cases, the python code (answers) for the synthetic question is generated separately using zero-shot program of thought (PoT) approach (Chen et al., 2023).

3.1 Financial QA

A high-level workflow of the synthetic question-python code pairs generation from financial excerpts is outlined in Figure 1. We use a four-step approach: selection of pages from financial documents, synthetic question generation, answer (python code) generation and data filtering. While LLMs can be used to identify candidate pages, for the scope of this paper, we assume candidate pages have already been selected and focus on the problem of synthetic question-answer pair generation. For question generation, an LLM is prompted to generate multiple financial questions using the provided image or text of a financial extract. As in previous synthetic data work for math problems, we use a temperature of 0.7 to encourage diversity in questions. We consider two options for the question generation prompt.

Financial QA using zero-shot: The zero-shot question generation prompt includes instructions about the question generation task, constraints regarding the type of arithmetic operations that can be used in solving the problem, the type of answer and additional instructions to ensure consistency and diversity in the financial question generation.

Financial QA using zero-shot with examples: When a few real example questions are available, the zero-shot with examples[1] prompt includes those examples in the zero-shot prompt.

In both cases, we pass the image/text along with the generated question to an LLM and prompt the LLM to write Python code to answer the synthetic question. For the code generation step we use a temperature of zero, and we utilize the zero-shot PoT approach with the context provided either as an image or text. In the final step, the generated samples with non-executable python codes or codes generating answers which indicate non-conformance with provided guidelines, are filtered out, see Appendix D for further details.

3.2 Conversational QA

In this setting, we explore a more general class of question-answering scenario in which a sequence of interconnected sub-questions is used to arrive at the answer for a complex financial reasoning question. We provide two methods to generate this sequence of interconnected questions:

Derived Conversational QA: In the first approach, we derive a sequence of interconnected sub-questions from a question-python code pair generated in *Financial QA* style along with the corresponding financial excerpt, see Figure 2 for details. As for *Financial QA* we consider two options: (1) zero-shot where we instruct the LLM to generate conversational style sub-questions and (2) we use the zero-shot prompt with examples that demonstrate a series of sub-questions that is equivalent to an original question.

The code generation step is not required here

[1]We refer to the approach as zero-shot with examples as we only provide example questions without any associated context. A few-shot approach would involve providing one or more examples with a context and a question generated using that context.

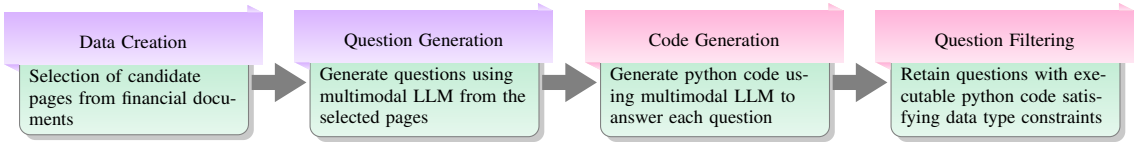

Figure 1: Workflow for generating synthetic data for financial question-answering.

Figure 2: Workflow for generating synthetic data for conversational financial question-answering.

as the final answer to the sequence of questions remains the same as that of the original single question.

Direct Conversational QA: In this approach we directly instruct the LLM to generate sequences of interconnected sub-questions using zero-shot prompting which is similar to the workflow described in Figure 1.

As for *Financial QA* data generation, we use zero-shot PoT prompting to generate python codes to answer the sub-questions, followed by filtering the generated samples.

4 Experiments

4.1 Datasets

We assess synthetic data generation by replicating three manually curated English financial QA datasets — FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022) and TATQA (Zhu et al., 2021). Our synthetic versions aim to mimic the original datasets where FinQA and TATQA involve answering questions from the provided financial text, and ConvFinQA focuses on answering the final question in a conversation chain, based on similar content.

The financial datasets, with their respective train and test splits, are listed in Table 7, Appendix A. For each data set, the total number of synthetic samples we generate is equal to the number of train samples in their respective data sets.

We begin synthetic data creation by using the same financial documents as the original studies. We outline the data creation steps for each dataset, clearly define the starting point for synthetic generation, detail the methodology, and describe the evaluation process.

4.1.1 Synthetic FinQA Data Generation

The FinQA dataset was constructed by selecting 12719 pages from the S&P 500 companies' earnings reports from 1999 to 2019, sourced from FinTabNet (Zheng et al., 2021). The selected pages containing simple tables meeting specific criteria, were annotated by expert annotators to create questions and reasoning programs. We converted these same pages into images to start the synthetic data generation.

For synthetic question generation, we input each image into GPT-4o with a custom prompt that guides it to produce questions aimed at boolean or float answers, requiring multi-hop reasoning and arithmetic, based on the image's table and text content (see Figure 3 in Appendix C). The Python codes for these questions are generated with a zero-shot prompt described in Figure 7 in Appendix C. We generate multiple distinct questions per page by including previous questions in the prompts, instructing the LLM to generate a question different from the prior questions.

Despite instructions in the prompt to generate questions that have boolean/float scalar answer, some questions yield answers in composite data structures like list/dictionaries (multiple values) or lead to non-executable code. We employ a filtering algorithm to remove such question-code pairs.

Additionally, we adopt a zero-shot with examples approach, incorporating examples into the prompt, as detailed in Figure 4 in Appendix C.

4.1.2 Synthetic TATQA Data Generation

The TATQA dataset, was sourced from around 500 financial reports, includes tables and accompanying text (Zhu et al., 2021). Only tables with 3 $\sim$

30 rows and 3 $\sim$ 6 columns and their related reports were considered. The question-answer pairs were created by annotators with financial expertise, using valid hybrid contexts, defined as consisting of a table and at least two related paragraphs. We initiate our synthetic data generation from these hybrid contexts.

We replicate the synthetic FinQA methodology, differing only in feeding the multimodal LLM with textual hybrid contexts for question and code generation. The zero-shot and zero-shot with examples approaches for synthetic TATQA data generation are detailed in Figure 5, Figure 6 and Figure 7 in Appendix C.

4.1.3 Synthetic ConvFinQA Data Generation

(Chen et al., 2022) provide the ConvFinQA dataset comprising conversational questions on financial reports, constructed from the FinQA (Chen et al., 2021) dataset's multi-step reasoning solutions. They provide annotators conversational skeletons and corresponding FinQA report data to craft sub-questions.The conversation skeletons are of two types: simple, derived from one multi-hop question, and hybrid, created from two multi-hop questions on the same report page.

For synthetic ConvFinQA data, we employ two methods. The *Derived Conversational QA* approach prompts GPT-4O with FinQA's synthetic question, solution code, report image, and instructions for sub-question generation, aiming for a conversational style that requires interpretability of a sub-question from previous sub-questions. This is done in zero-shot and zero-shot with examples settings, detailed in Figure 8 and Figure 9 in Appendix C.

The *Direct Conversational QA* approach uses FinQA page images, directing GPT-4O to create 2-5 conversational sub-questions, as per Figure 10 in Appendix C. The page image and sub-questions are passed to GPT-4O for generating the python code to answer the last sub-question, using the prompt shown in Figure 11 in Appendix C. We apply the same filtering as in FinQA synthetic generation.

4.2 Evaluation Approach

Using the generated synthetic data, we fine-tune three models: PHI-3-MINI, PHI-3-MEDIUM, and MISTRAL 7B (see Table 8 in Appendix B). We then compare the accuracy of these models with that of the same SLMs trained on the real data. The fine-tuning uses the same method and hyper-parameters as in (Phogat et al., 2024). We ran the fine-tuning for 4 epochs and evaluated the model at the end of the fourth epoch on the test split. We employ the vLLM4[2] framework for conducting inference on fine-tuned models. The experiments are performed on a compute instance with 24 cores, 220GB RAM and a A100 GPU (80GB).

The Python codes generated by the fine-tuned models are executed using the Python exec function to determine the resulting answer, which is then compared against the ground truth. We use the performance of the fine-tuned model on real data as the baseline for comparison.

5 Results

5.1 Evaluation of Synthetic Financial QA Data

Table 1 summarizes the comparative performance of SLMs trained with different data: synthetic data using zero-shot prompt, synthetic data with zero-shot with examples prompt and real data. The accuracy is measured on the real FinQA and TATQA test data sets.

Our findings show that for both data sets, models trained on real data perform better than those trained on synthetic data, whether using zero-shot or zero-shot with example question approaches. Nevertheless, synthetic data-trained models are competitive, especially for TATQA, where the performance gap between synthetic and real data-trained PHI-3 models is a mere 1% to 3%, and for MISTRAL 7B, the outcomes are nearly identical. The models fine-tuned on synthetic FinQA data exhibit accuracy within 5% of those fine-tuned on real data for the PHI-3 models and approximately 9% for the MISTRAL 7B model.

The inclusion of examples in the prompt for generating the synthetic data minimally impacted the fine-tuned models' accuracy, indicating that the LLM's inherent domain knowledge suffices for creating pertinent questions without needing illustrative examples.

We conducted a detailed analysis of models trained on synthetic FinQA data, comparing their performance based on (a) the source of entity values required to answer the question —Table only, Text & Table, Text only (Table 9 in Appendix E), and (b) the type of answer—numerical or Boolean (Table 10 in Appendix E). The discrepancy between real and synthetic data is notably higher

[2]https://docs.vllm.ai/en/latest/

Fine-tuning datasets	PHI-3-MINI	PHI-3-MEDIUM	MISTRAL 7B
Accuracy on real FinQA test data			
Synthetic FinQA data: 0-shot⋆	68.43	73.49	67.21
Synthetic FinQA data: 0-shot + EQs⋆	68.09	73.58	68.09
Real FinQA data	73.49	77.59	76.63
Accuracy on real TATQA test data			
Synthetic TATQA data: 0-shot⋆	88.99	90.80	88.44
Synthetic TATQA data: 0-shot + EQs⋆	87.74	90.66	88.85
Real TATQA data	90.94	93.03	88.71

⋆ The synthetic data is generated using *Financial QA* setting for both FinQA and TATQA datasets. The prompts 0-shot and 0-shot + EQs represent *zero-shot prompt* and *zero-shot prompt with example questions* respectively.

Table 1: Comparison of models trained on synthetic and real data for financial question answering.

for Text only questions, particularly with PHI-3 models, as shown in Table 9. Boolean questions reveal a marked underperformance by smaller models PHI-3-MINI and MISTRAL 7B, as seen in Table 10. Through an audit of 50 synthetic FinQA questions, we found less than 5% were Text only or Boolean, suggesting a bias in the synthetic data generation. Enhancing the prompt could yield a more varied question set and improve model performance.

Overall, the results indicate that synthetic data generated with the proposed prompt and methodology can closely match the performance of the models achieved by training on the real data.

Accuracy on FinQA test data				
Training dataset	**750**	**1500**	**3000**	**Full⋆**
Synthetic FinQA	63.99	64.95	68.61	68.43
Real FinQA	69.83	71.31	71.49	73.49

Accuracy on TATQA test data				
Training dataset	**750**	**1500**	**3000**	**Full⋆**
Synthetic FinQA	82.17	83.56	82.31	84.26
Real FinQA	81.19	79.66	81.75	81.19

⋆ Full denotes the full Synthetic/Real FinQA dataset.

Table 2: Performance of PHI-3-MINI model trained on synthetic and real FinQA data for various sample sizes.

5.2 Effect of Sample Size

We conduct experiments to assess the impact of training data volume on model performance and its generalization capabilities. We fine-tune the PHI-3-MINI model with six distinct training sets comprising 750, 1500, and 3000 samples each derived exclusively from either synthetic or real FinQA data. The efficacy of the fine-tuned models was measured using FinQA test data, while their capacity to generalize was assessed through testing on the independently collected TATQA test data, see Table 2 for details.

Results in Table 2 show that both fine-tuned models demonstrate performance improvement with larger training data sizes when tested on FinQA test data. In contrast, when tested on the TATQA test data, the model trained on real FinQA data does not benefit from increasing data volume while the model trained on synthetic FinQA data shows slight improvement. Moreover, the model trained with full synthetic FinQA data achieves a 3% higher accuracy on the test split of TATQA data than the one trained on full real FinQA data.

We perform a similar experiment, training models on synthetic and real TATQA data and evaluating their performance on both the FinQA and TATQA test data sets, see Table 3 for details. With increasing training data size, the model trained with synthetic TATQA data showed performance gains on both the FinQA and TATQA test datasets. In contrast, the model trained on real TATQA data showed performance improvements only on the TATQA test set, with a slight decline on the FinQA test set.

These findings suggest synthetic data may offer generalizability benefits due to its broader question variety, as opposed to real data which may underperform on similar but independent datasets due to differences in question style and nature.

5.3 Synthetic Data Analysis

To gain further insights, we conduct an analysis to assess synthetic FinQA and TATQA data quality. We first vectorize questions of the real and synthetic

Accuracy on FinQA test data				
Training dataset	**750**	**1500**	**3000**	**Full***
Synthetic TATQA	65.47	65.47	67.56	67.82
Real TATQA	64.86	64.16	63.46	63.81

Accuracy on TATQA test data				
Training dataset	**750**	**1500**	**3000**	**Full***
Synthetic TATQA	85.51	87.88	88.02	88.99
Real TATQA	87.04	86.09	88.3	90.94

* Full denotes the full Synthetic/Real TATQA dataset.

Table 3: Performance of PHI-3-MINI model trained on synthetic and real TATQA data for various sample sizes.

Accuracy on FinQA test data			
Training dataset	***Closest***	***Farthest***	***Random***
Synthetic FinQA	66.43	64.16	63.99
Synthetic TATQA	63.20	65.91	65.47

Accuracy on TATQA test data			
Training dataset	***Closest***	***Farthest***	***Random***
Synthetic FinQA	83.42	82.17	82.17
Synthetic TATQA	84.67	85.65	85.51

Table 4: Performance of PHI-3-MINI model trained on 750 samples drawn from the synthetic data.

samples using text embeddings[3]. We then calculate the nearest neighbor distance (NN-distance) from the vectorized question of the synthetic sample q_i to the corresponding real dataset, as follows:

$$d(q_i, \mathcal{D}_{\text{real}}) = \underset{\tilde{q} \in \mathcal{D}_{\text{real}}}{\text{maximize}}\ 1 - \langle q_i, \tilde{q} \rangle$$

where d represents the cosine distance from q_i to the vectorized question $\tilde{q}$ in the real dataset $\mathcal{D}_{\text{real}}$. The histogram plots of NN-distances for synthetic FinQA and TATQA datasets are presented in Figure 12.

We perform a detailed examination of 500 random synthetic FinQA and TATQA questions. For both data sets, synthetic questions exhibiting NN-distances less than 0.1 to their nearest real dataset counterpart are mostly identical with minor variations. The synthetic questions with NN-distances between 0.1 and 0.3 demonstrate significant overlap in financial entities compared with their real counterparts. However, they start to differ when it comes to the specific calculations required. Finally, synthetic questions that have NN-distances above 0.3 bear little or no relation to the corresponding real questions.

To evaluate the sensitivity of the SLM to training samples, we select 750 synthetic questions that are the nearest matches to the real questions (denoted as *Closest*), as well as 750 that are the farthest (denoted as *Farthest*), from both the TATQA and FinQA datasets. A selection of these samples from TATQA is presented in Table 11 and Table 12, and from FinQA in Table 13 and Table 14 in Appendix F.

We fine-tune PHI-3-MINI model on the *Closest* and *Farthest* data for both FinQA and TATQA. We evaluate the test accuracy of all models on their respective test sets and compare with the results from the corresponding random sample of 750 (see Table 4. Despite the difference in the two data sets, the accuracy of the fine-tuned PHI-3-MINI models on the *Closest* and *Farthest* training samples falls within 2% of the accuracy of the PHI-3-MINI model trained on a random selection of 750 synthetic samples (denoted as *Random*). These results suggest that the models trained with QA pairs generated by a LLM may generalize to a test dataset with dissimilar questions.

5.4 Evaluation of Synthetic Conversational Financial QA Data

Table 5 presents a comparison of accuracies on ConvFinQA test data for models fine-tuned on real and synthetic conversational financial QA data in zero-shot and zero-shot with examples scenarios. In addition to overall accuracy, we assess the performance on simple and hybrid conversations. Models trained on synthetic data generated using the *Derived Conversational QA* show notably lower accuracy than those fine-tuned on real data, with up to a 15% discrepancy for simple conversations and a 28% to 48% gap for hybrid conversations. These results could be due to the approach targeting the generating of simple conversations which may impact the performance on hybrid conversations.

For synthetic data generated using the *Direct Conversational QA* approach, the accuracy on simple questions across the models is comparable to the *Derived Conversational QA* approach. However, we observe a large improvement on hybrid questions over the Derived approach, with less than 17% performance gap from the model fine-tuned on real data. These results indicate that directly prompting the LLM does better at generating conversational data that is better aligned with the hy-

[3]The embeddings are generated using `text-embedding-3-small` model from OpenAI.

ConvFinQA datasets for Supervised Fine-tuning	PHI-3-MINI			PHI-3-MEDIUM			MISTRAL 7B		
	Simple	Hybrid	Overall	Simple	Hybrid	Overall	Simple	Hybrid	Overall
Accuracy on real ConvFinQA test data									
Syn: Derived 0-shot⋆	66.66	28.91	55.81	73.66	45.45	65.58	65	22.13	52.73
Syn: Derived 0-shot + EQs⋆	64	25.61	52.96	71.66	46.28	64.37	69.33	27.27	57.24
Syn: Direct 0-shot⋆	67	49.58	62	75.33	61.98	71.49	69.66	54.54	65.32
Syn: ConvFinQA + FinQA†	68.66	62.80	67	77.33	65.28	73.81	67.33	61.98	65.79
Real ConvFinQA dataset	80	66.11	76	85.33	73.55	81.94	79.33	70.24	76.72

⋆ The synthetic data is generated using *Conversational QA* setting for the ConvFinQA dataset. The synthetic datasets Syn: Derived 0-shot, Syn: Derived 0-shot + EQs, Syn: Direct 0-shot are generated using *derived zero-shot prompt*, *derived zero-shot prompt with example questions* and *direct zero-shot prompt* respectively.

† The Syn: ConvfinQA + FinQA dataset is combined from ConvFinQA dataset generated using *derived zero-shot prompt* and FinQA dataset generated using *zero-shot prompt*.

Table 5: Comparison of models trained on synthetic and real ConvFinQA data for financial question answering.

brid questions in the ConvFinQA data set.

We further experiment with augmenting the directly generated synthetic ConvFinQA data with the synthetic FinQA data. The results shown in Table 5 indicate significantly improved performance on the hybrid questions for PHI-3-MINI (13%) and MISTRAL 7B (8%) with a modest improvement for PHI-3-MEDIUM (3%). These improvements translate to a 5% increase in overall accuracy for the PHI-3-MINI model and a 2% increase for PHI-3-MEDIUM. For MISTRAL 7B, there is little change in overall accuracy as the improvement on hybrid conversations, is accompanied by a small degradation on the simple conversations. These results demonstrate the LLMs capability to generate conversational financial QA data with the PHI-3 models fine-tuned entirely on synthetic data achieving an accuracy within 9% of that using real data.

5.5 Performance on Mixture of Synthetic and Real Data

While synthetic conversational data yields promising results, models trained on it underperform compared to those trained on real data. We explore the necessary proportion of real data in the training set to close this performance gap, utilizing synthetic data generated via the second approach. Table 6 compares the accuracy of the PHI-3-MINI model fine-tuned with a mix of real and synthetic data in a zero-shot setting to the PHI-3-MINI fine-tuned solely on real ConvFinQA data, maintaining equal sample sizes. The findings reveal a notable accuracy boost with just 10% real data, and with 20% real data, performance nears that of the fully real data fine-tuned model. This suggests that LLM generated synthetic data can greatly reduce the need for extensive real-world data collection in conversational financial QA tasks.

Accuracy on ConvFinQA test data			
x% of synthetic + y% of real data	**Simple**	**Hybrid**	**Overall**
x=90%, y=10%	72	54.54	69.98
x=80%, y=20%	74.33	59.50	73.07
x=60%, y=40%	77.66	64.46	74.87
y=100%	80	66.11	76

Table 6: Performance of PHI-3-MINI trained on mixtures of synthetic and real ConvFinQA data.

6 Conclusion

We explored synthetic data creation for financial reasoning in both standard and conversational settings through a multi-step process. To assess the generation methods, synthetic datasets were produced from the same sources used for creating three existing manually annotated financial reasoning datasets. By comparing SLMs trained on both synthetic and real data, we demonstrated the viability of synthetic data for both standard and conversational financial QA. Our findings provide valuable insights into the strengths and limitations of large language models in generating synthetic datasets for financial reasoning tasks.

Disclaimer

The views reflected in this article are the views of the authors and do not necessarily reflect the views of the global EY organization or its member firms.

References

Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer, and Samuel Marc Denton. 2024. Balancing Cost and Effectiveness of Synthetic Data Generation Strategies for LLMs. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711. 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*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better Synthetic Data by Retrieving and Transforming Existing Datasets. *arXiv preprint arXiv:2404.14361*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large Language Models Are Reasoning Teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Yechan Hwang, Jinsu Lim, Young-Jun Lee, and Ho-Jin Choi. 2023. Augmentation for Context in Financial Numerical Reasoning over Textual and Tabular Data with Large-Scale Language Model. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024. Common 7B Language Models Already Possess Strong Math Capabilities. *arXiv preprint arXiv:2403.04706*.

Shengzhi Li and Nima Tajbakhsh. 2023. SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs. *arXiv preprint arXiv:2308.03349*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best Practices and Lessons Learned on Synthetic Data. *arXiv preprint arXiv:2404.07503*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. *arXiv preprint arXiv:2308.09583*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching Small Language Models to Reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.

Karmvir Singh Phogat, Sai Akhil Puranam, Sridhar Dasaratha, Chetan Harsha, and Shashishekar Ramakrishna. 2024. Fine-tuning Smaller Language Models for Question Answering over Financial Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida. Association for Computational Linguistics.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. Prompting-based Synthetic Data Generation for Few-Shot Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13168–13178, Torino, Italia. ELRA and ICCL.

Kosuke Takahashi, Takahiro Omi, Kosuke Arima, and Tatsuya Ishigaki. 2023. Training Generative Question-Answering on Synthetic Data Obtained from an Instruct-tuned Model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 786–791, Hong Kong, China. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? *arXiv preprint arXiv:2303.04360*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards Zero-Label Language Learning. *arXiv preprint arXiv:2109.09193*.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. 2024. Synthetic Multimodal Question Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Ziqiang Yuan, Kaiyuan Wang, Shoutai Zhu, Ye Yuan, Jingya Zhou, Yanlin Zhu, and Wenqi Wei. 2024. FinLLMs: A Framework for Financial Reasoning Dataset Generation with Large Language Models. *arXiv preprint arXiv:2401.10744*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 697–706.

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 (Volume 1: Long Papers)*, pages 3277–3287. Association for Computational Linguistics.

A Financial Question Answering Datasets

In our fine-tuning experiments on SLMs, we used training (train) and testing (test) splits of the financial datasets. The FinQA dataset comes with predefined splits and their corresponding ground truths. For the ConvFinQA and TATQA datasets, which lack ground truths in their predefined test splits, we used the predefined dev splits as test sets. The dataset splits are detailed in Table 7, where for TATQA, only arithmetic questions are considered.

Financial datasets	Train	Test
FinQA	6251	1147
ConvFinQA	2737	421
TATQA*	4992	718

* Only arithmetic questions are considered.

Table 7: Dataset splits used in our experiments.

B SLMs for Supervised Fine-tuning

We perform fine-tuning experiments on MISTRAL and PHI-3 model families. The additional details on SLMs such as model size, license and HuggingFace API are provided in Table 8.

C Synthetic Data Generation Prompts

In this section, we list all the prompts used for synthetic data generation. We experimented with three datasets: FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), and TATQA (Zhu et al., 2021). For the FinQA and TATQA datasets, we generated synthetic data using the *Financial QA* setting, while for the ConvFinQA dataset, we employed the *Conversational QA* setting for data generation.

Model Name	Parameters	HuggingFace API	License
MISTRAL 7B	7B	`mistralai/Mistral-7B-Instruct-v0.2`	apache-2.0
PHI-3-MINI	3.8B	`microsoft/Phi-3-mini-128k-instruct`	mit
PHI-3-MEDIUM	14B	`microsoft/Phi-3-medium-128k-instruct`	mit

Table 8: Description of SLMs used for supervised fine-tuning

C.1 FinQA & TATQA Datasets

Under the *Financial QA* setting, synthetic questions are generated from excerpts of financial documents using GPT-4O with zero-shot and zero-shot with examples prompting for the FinQA dataset, as described in Figures 3 and Figure 4, respectively. Similarly, for the TATQA dataset, the zero-shot and zero-shot with examples prompting is described in Figure 5 and Figure 6 respectively. The answers to the generated questions, in the form of python code, are produced using GPT-4O with the python code generation prompt outlined in Figure 7.

C.2 ConvFinQA Dataset

Under the *Conversational QA* setting, synthetic questions are generated from excerpts of financial documents using GPT-4O with derived conversational QA prompting and direct conversational QA prompting. In derived conversational QA setting, the conversational financial questions are generated from the questions generated using *Financial QA* setting using zero-shot and zero-shot with examples prompting as described in Figure 8 and Figure 9 respectively. In direct conversational prompting, the financial questions are generated directly from financial documents using GPT-4O with the zero-shot prompt described in Figure 10. The answers to the generated conversational questions, in the form of python codes, are produced using GPT-4O with the python code generation prompt outlined in Figure 11.

D Filtering Technique for Synthetic Samples

The FinQA and ConvFinQA datasets exclusively feature questions with numerical or boolean answers. In our current study, from the TATQA dataset, we selectively consider only those questions yielding numerical or boolean responses. In a few cases the synthetically generated data creates questions that leads to answers that are neither numerical nor boolean. The filtering algorithm checks the data type of the answer generated by executing the python code. If it is not numerical or boolean, the sample is eliminated from the training set. In addition, we also eliminate samples where the generated python code results in code that is non-executable.

E FinQA Performance Breakdown

The performance metrics of the models, which were fine-tuned on both the real and synthetic FinQA datasets, are further breakdown based on the different question types within the test split. The accuracy of the FinQA test questions are categorized along two dimensions (a) Table only, Text & Table, Text only where the different categories refer to the location of the entity values required to answer the questions (see Table 9) and (b) Questions that have a numerical vs Boolean answer (see Table 10).

F Synthetic Data Analysis

For analyzing the synthetic data, we first compute nearest neighbor distance for a synthetic sample to the real dataset using cosine distance metric as discussed in Sec. 5.3. The density plots of these nearest neighbor distances for the synthetic TATQA and FinQA datasets are given in Figure 12.

We now present a selection of illustrative synthetic questions and their corresponding nearest neighbors from the real dataset. For the TATQA dataset, we showcase a series of 10 evenly distributed questions from both the *Closest* and *Farthest* splits in Table 11 and Table 12, respectively. Similarly, for the FinQA dataset, the same arrangement of questions from the *Closest* and *Farthest* split can be found in Table 13 and Table 14.

System Prompt

You are an expert financial analyst skilled in generating questions that are meaningful for financial analysis or generating insights from the company financial reports. Your task involves analyzing an image of a single page from a financial report to generate a relevant question. While performing this task, you must adhere to a set of specified constraints.

User Prompt

Given the image of a page from a financial report, generate a financial reasoning question. Please adhere to the following constraints meticulously when formulating the question.

Constraints:

1. The question must be generated such that it always leads to a single numerical or boolean answer.
2. The question should preferably use concepts different than the concepts used in any question you see in the following list: [`first_generated_question`, `second_generated_question`, `third_generated_question`].
3. The question must be strictly derived from the content present in the image.
4. Answering the question should require the use of values from the table and/or values present in the text around the table.
5. Calculating the answer should involve multi-hop reasoning with the following arithmetic operations:
   -Basic operations: Addition, Subtraction, Multiplication, Division, Exponential, Greater Than
   -Table aggregation operations: Sum, Average, Minimum, Maximum

The final response should be formatted as a JSON object with only question and no other objects should be included:
"Question": "<Generated Question>".

Figure 3: FinQA question generation prompt: Zero-shot

FinQA Questions	**PHI-3-MINI***		**PHI-3-MEDIUM***		**MISTRAL 7B***	
categorization	Real	Synthetic	Real	Synthetic	Real	Synthetic
Table Only	78.61	74.36	83	79.60	81.44	73.37
Text Only	69.25	60.07	71.37	63.95	69.25	59.25
Text & Table	58.22	56.96	64.55	63.29	59.49	53.79

* These models are trained on the real FinQA dataset and synthetic FinQA dataset generated using *zero-shot prompt* in setting Real and Synthetic respectively.

Table 9: Performance breakdown of the FinQA test accuracy for the models trained on synthetic/real FinQA dataset.

System Prompt

You are an expert financial analyst skilled in generating questions that are meaningful for financial analysis or generating insights from the company financial reports. Your task involves analyzing an image of a single page from a financial report to generate a relevant question. While performing this task, you must adhere to a set of specified constraints.
Below are a few examples of the types of questions you can generate:

- what was the percentage cumulative total return for the five year period ended 31-dec-2017 of citi common stock?
- what percentage of the total oil and gas mmboe comes from canada?
- what are the deferred fuel cost revisions as a percentage of the increase in fuel cost recovery revenues?
- at the end of 2014 , the notional value of derivatives designated as hedging instruments under gaap was what percent of the fair value?
- in 2010 what was the percentage change of the carrying amount of loan receivable net of the allowance?
- what is the average amortization amount , in millions , from 2015-2019?
- what amount of long-term debt due in the next 36 months as of december 31 , 2003 , in millions?
- what is the average future minimum annual rental payment for the next five years?
- what was the difference in percentage cumulative 5-year total stockholder return for cadence design systems inc . compared to the nasdaq composite for the five years ended 12/29/2012?
- what is the growth of the additions in comparison with the growth of the deductions during 2003 and 2004?

User Prompt

Given the image of a page from a financial report, generate a financial reasoning question. Please adhere to the following constraints meticulously when formulating the question.

Constraints:

1. The question must be generated such that it always leads to a single numerical or boolean answer.
2. The question should preferably use concepts different than the concepts used in any question you see in the following list: [`first_generated_question`, `second_generated_question`, `third_generated_question`].
3. The question must be strictly derived from the content present in the image.
4. Answering the question should require the use of values from the table and/or values present in the text around the table.
5. Calculating the answer should involve multi-hop reasoning with the following arithmetic operations:
   -Basic operations: Addition, Subtraction, Multiplication, Division, Exponential, Greater Than
   -Table aggregation operations: Sum, Average, Minimum, Maximum

The final response should be formatted as a JSON object with only question and no other objects should be included:
"Question": "<Generated Question>".

Figure 4: FinQA question generation prompt: Zero-shot with examples

System Prompt

You are an expert financial analyst skilled in generating questions that are meaningful for financial analysis or generating insights from the company financial reports. Your task involves analyzing text of a single page from a financial report to generate a relevant question. While performing this task, you must adhere to a set of specified constraints.

User Prompt

Given the text of a page from a financial report, generate a financial reasoning question. Please adhere to the following constraints meticulously when formulating the question.
Constraints:

1. The question must be generated such that it always leads to a single numerical or boolean answer.
2. The question should preferably use concepts different than the concepts used in any question you see in the following list: [`first_generated_question`, `second_generated_question`, `third_generated_question`].
3. The question must be strictly derived from the content present in the text.
4. Answering the question should require the use of values from the table and/or values present in the text around the table.
5. Calculating the answer should involve multi-hop reasoning with the following arithmetic operations:
   -Basic operations: Addition, Subtraction, Multiplication, Division, Exponential, Greater Than
   -Table aggregation operations: Sum, Average, Minimum, Maximum

The final response should be formatted as a JSON object with only question and no other objects should be included:
"Question": "<Generated Question>".

Figure 5: TATQA question generation prompt: Zero-shot

System Prompt

You are an expert financial analyst skilled in generating questions that are meaningful for financial analysis or generating insights from the company financial reports. Your task involves analyzing an image of a single page from a financial report to generate a relevant question. While performing this task, you must adhere to a set of specified constraints.
Below are a few examples of the types of questions you can generate:

- What is the Value Realized on Vesting of Mark J. Barrenechea expressed as a percentage of total Value Realized on Vesting?
- What was the average trading profit for 2017/18 and 2018/19?
- What is the average net restructuring and exit costs over the 3 year period?
- What is the ratio of net cash used in investing activities from 2018 to 2019?
- What is the average of Financing under Global Financing?
- What is the percentage of non-UK activities in loss before income taxes and equity in net loss of affiliates for the year ended December 31, 2019?
- How much did the company pay upon the signing of the toxicology studies agreement?
- What percentage of total contractual obligations were due less than a year?
- What is the Total contractual cash obligations for years 2020-2024 inclusive?
- What is the amount of net sales derived in 2018?

User Prompt

Given the image of a page from a financial report, generate a financial reasoning question. Please adhere to the following constraints meticulously when formulating the question.
Constraints:

1. The question must be generated such that it always leads to a single numerical or boolean answer.
2. The question should preferably use concepts different than the concepts used in any question you see in the following list: [`first_generated_question`, `second_generated_question`, `third_generated_question`].
3. The question must be strictly derived from the content present in the image.
4. Answering the question should require the use of values from the table and/or values present in the text around the table.
5. Calculating the answer should involve multi-hop reasoning with the following arithmetic operations:
   -Basic operations: Addition, Subtraction, Multiplication, Division, Exponential, Greater Than
   -Table aggregation operations: Sum, Average, Minimum, Maximum

The final response should be formatted as a JSON object with only question and no other objects should be included:
"Question": "<Generated Question>".

Figure 6: TATQA question generation prompt: Zero-shot with examples

System Prompt

You are an expert financial analyst skilled in generating python code to answer financial reasoning questions.

User Prompt

Given the image of a page from a financial report and the financial question, write Python code to answer the question.
###Question: `Generated Question`
###Instructions:

1. First, identify entities required to answer the question. Extract the identified entities and store in python variables.
2. Then perform calculations with the entities and strictly store the answer to the python variable "ans".
3. Python code must end after the variable "ans" is defined. Comments must begin with character "#".

The final response should be formatted as a JSON object with the following fields and no others:
"Question": "<Generated Question>",
"Explanation": "Explanation of the steps to generate the answer",
"Python_code": "###Python <Python code to calculate the answer> ###End Python".

Figure 7: FinQA/TATQA code generation prompt: Zero-shot

System Prompt

You are an expert in generating financial subquestions in a conversational style for a given question. A conversational style means that a given subquestion needs to look at the previous subquestions to be interpreted and cannot be interpreted by itself.

User Prompt

Given an image of a page from a financial statement, a question to be answered from the provided page and the python code which encodes the calculations required to answer the question, generate a sequence of conversational style subquestions for the given original question.
###Original_Question: `Question`
###Python_Code_to_Answer: `Python code`
Constraints:

1. Ensure that the answers to the subquestions involve financial entities or calculations.
2. The sequence of subquestions must be strictly equivalent to the original question with the answer to the last question being the same as the answer to the given original questions.
3. These subquestions must be significantly different from each other.
4. Verify that the generated python code contains the correct logic and calculations to answer the generated sequence of subquestions.
5. If you can't generate meaningful subquestions or the python code does not correctly answer the generated subquestions , return an empty list.

The final response should be formatted as a JSON object with the following fields and no others:
"Convfinqa_Subquestions": "<[subquestion1, subquestion2, subquestion3, . . .]>"

Figure 8: ConvFinQA question generation prompt: Derived zero-shot

FinQA Questions	**PHI-3-MINI***		**PHI-3-MEDIUM***		**MISTRAL 7B***	
categorization	Real	Synthetic	Real	Synthetic	Real	Synthetic
Binary	95	60	90	95	90	65
Numerical	73.11	68.58	77.37	73.11	75.15	67.25

* These models are trained on the real FinQA dataset and synthetic FinQA dataset generated using *zero-shot prompt* in setting Real and Synthetic respectively.

Table 10: Performance breakdown of the FinQA test accuracy for the models trained on synthetic/real FinQA dataset.

System Prompt

You are an expert in generating financial subquestions in a conversational style for a given question.A conversational style means that a given subquestion needs to look at the previous subquestions to be interpreted and cannot be interpreted by itself.
Below are the set questions and subquestions which can be used as reference for generating the confinqa subquestions.

Example 1:
Original Question: by how much did the weighted average exercise price per share increase from 2005 to 2007?
Convfinqa_Subquestions: [‘what was the weighted average exercise price per share in 2007?’, ‘and what was it in 2005?’, ‘what was, then, the change over the years?’, ‘what was the weighted average exercise price per share in 2005?’, ‘and how much does that change represent in relation to this 2005 weighted average exercise price?’]

Example 2:
Original Question: what percentage of amounts expensed in 2009 came from discretionary company contributions?
Convfinqa_Subquestions: [‘what is the ratio of discretionary company contributions to total expensed amounts for savings plans in 2009?’, ‘what is that times 100?’]

Example 3:
Original Question: what is the total return is $ 100000 are invested in s&p500 on january 1st , 2015 and sold at the end of 2016?
Convfinqa_Subquestions: [‘what is the change in price of the s&p 500 from 2015 to 2016?’, ‘what is 100000 divided by 100?’, ‘what is the product of the change by the quotient?’]

Example 4:
Original Question: what is the growth rate in total shipment volume from 2010 to 2011?
Convfinqa_Subquestions: [‘what was the difference in total shipment volume between 2010 and 2011?’, ‘and the specific value for 2010?’, ‘so what was the growth rate over this time?’]

Example 5:
Original Question: what portion of total long-term borrowings is due in the next 36 months?
Convfinqa_Subquestions: [‘what was the amount of notes maturing in june 2022?’, ‘and the maturity amount due in 2017?’, ‘combined, what is the total of these two values?’, ‘and the total long-term borrowings?’, ‘and the total portion due in the next 36 months?’]

Example 6:
Original Question: what was the percentage cumulative total return for the five year period ended 31-dec-2017 of citi common stock?
Convfinqa_Subquestions: [‘what is the value of citi common stock in 2017 less an initial $100 investment?’, ‘what is that divided by 100?’]

Example 7:
Original Question: what is the total amount of cash outflow used for shares repurchased during november 2007 , in millions?
Convfinqa_Subquestions: [‘what was the total amount of cash outflow used for shares repurchased during november 2007, in millions of dollars?’, ‘and how much is that in dollars?’]

Example 8:
Original Question: considering the year 2012 , how bigger is the capital expenditures on a non-gaap basis than the one on a gaap basis?
Convfinqa_Subquestions: [‘what were the capital expenditures on a non-gaap basis in 2012?’, ‘and what were the capital expenditures on a gaap basis in that same year?’, ‘how much, then, do the capital expenditures on a non-gaap basis represent in relation to the ones on a gaap basis, in 2012?’, ‘and what is the difference between this value and the number one?’]

Example 9:
Original Question: what was the percentage growth in the operating profit as reported from 2017 to 2018?
Convfinqa_Subquestions: [‘what was reporting operating profit in 2018?’, ‘what was it in 2017?’, ‘what is the net change?’, ‘what is the percent change?’]

Example 10:
Original Question: what was the cost per car for the buyout of locomotives in 2012?
Convfinqa_Subquestions: [‘what was the value included in the capital investments for buyout of locomotives in 2012, in dollars?’, ‘and how many locomotives were bought with that value?’, ‘what was, then, the average cost of each one of those locomotives?’]

User Prompt

Given an image of a page from a financial statement, a question to be answered from the provided page and the python code which encodes the calculations required to answer the question, generate a sequence of conversational style subquestions for the given original question.
###Original_Question: `Question`
###Python_Code_to_Answer: `Python code`
Constraints:
1. Ensure that the answers to the subquestions involve financial entities or calculations.
2. The sequence of subquestions must be strictly equivalent to the original question with the answer to the last question being the same as the answer to the given original questions.
3. These subquestions must be significantly different from each other.
4. Verify that the generated python code contains the correct logic and calculations to answer the generated sequence of subquestions.
5. If you can’t generate meaningful subquestions or the python code does not correctly answer the generated subquestions , return an empty list.

The final response should be formatted as a JSON object with the following fields and no others:
"Convfinqa_Subquestions": "<[subquestion1, subquestion2, subquestion3, . . .]>"

Figure 9: ConvFinQA question generation prompt: Derived zero-shot with examples

System Prompt

You are an expert in generating financial sub-questions in a conversational style for a given image. A conversational style means that a given sub-question needs to look at the previous sub-questions to be interpreted and cannot be interpreted by itself.

User Prompt

Given an image of a page from a financial statement, generate a sequence of conversational style subquestions involving multihop numerical reasoning.
Constraints:

1. Ensure that the answers to the subquestions must be float and are derived from numerical financial entities.
2. Calculating the answers should involve multi-hop reasoning with the following arithmetic operations:
   -Basic operations: Addition, Subtraction, Multiplication, Division, Exponential, Greater Than
   -Table aggregation operations: Sum, Average, Minimum, Maximum
3. Each subquestion must be a follow up to the previous subquestions.
4. Ensure the number of subquestions are between 2 and 5

The final response should be formatted as a JSON object with the following fields and no others:
"Convfinqa_Subquestions": "<[subquestion1, subquestion2, subquestion3, . . .]>"

Figure 10: ConvFinQA question generation prompt: Direct zero-shot

System Prompt

You are an expert financial analyst skilled in generating python code to answer financial reasoning questions.

User Prompt

Given the image of a page from a financial report write Python code to answer the last question from the series of questions.
###Questions: `A series of sub-questions`
###Last Question: `Last sub-question from the series`
###Instructions:

1. First, identify entities required to answer the question. Extract the identified entities and store in python variables.
2. Then perform calculations with the entities and strictly store the answer to the python variable "ans".
3. Python code must end after the variable "ans" is defined. Comments must begin with character "#".

The final response should be formatted as a JSON object with the following fields and no others:
"Question": "<Generated Question>",
"Explanation": "Explanation of the steps to generate the answer",
"Python_code": "###Python <Python code to calculate the answer> ###End Python".

Figure 11: ConvFinQA code generation prompt: Zero-shot

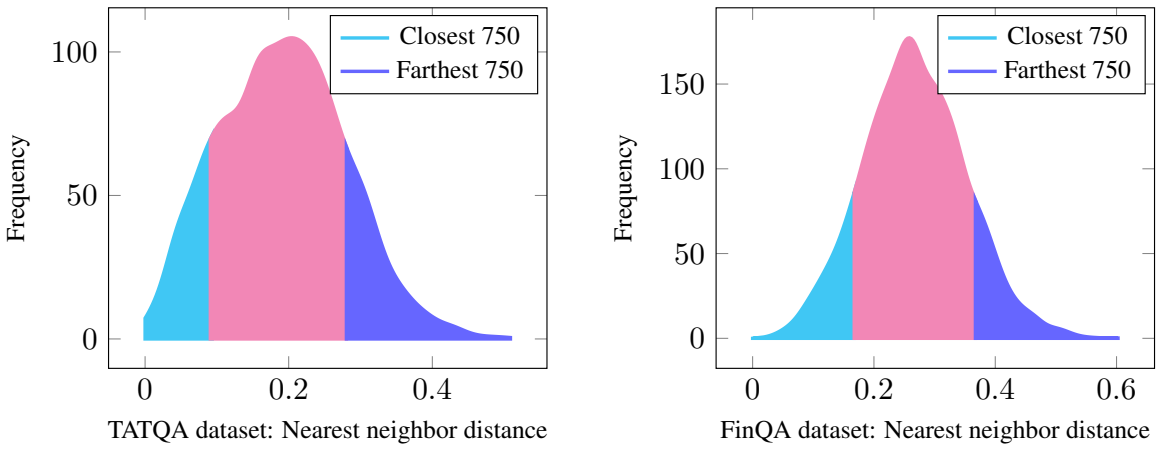

Figure 12: Distribution of the nearest neighbor distance for a sample from the synthetic dataset to the real dataset.

S. No.	Question from synthetic TATQA dataset	Nearest neighbor question from real TATQA dataset	Cosine distance
1	What is the percentage change in total deferred tax assets from 2018 to 2019?	What is the percentage change in total deferred tax assets from 2018 to 2019?	0
2	What was the percentage change in Gross profit as a percentage of revenue from 2018 to 2019?	What was the percentage change in gross profit between 2018 and 2019?	0.0313
3	What is the percentage increase in Total Assets from 2018 to 2019?	What was the percentage increase / (decrease) in the total assets from 2018 to 2019?	0.0313
4	What was the average net cash provided by (used for) operating activities over the 3-year period 2017-2019?	What was the average net cash provided by operating activities from 2017-2019?	0.0482
5	What is the percentage increase in the total of other non-current assets from 2018 to 2019?	What was the percentage change in total other non-current assets from 2018 to 2019?	0.0571
6	What is the percentage decrease in total stock-based compensation expense from 2017 to 2019?	What is the percentage change in the total stock-based compensation expense from 2018 to 2019?	0.0667
7	What is the average risk-free interest rate over the years 2017, 2018, and 2019?	What is the average risk-free interest rate for 2018 and 2019?	0.0734
8	What is the percentage change in total financial resources from 2017 to 2019?	What is the percentage increase / (decrease) in the Total financial resources from 2018 to 2019?	0.0791
9	What is the percentage change in Net Operating (Loss) Income from 2018 to 2019?	What is the percentage change in net loss from 2018 to 2019?	0.0848
10	What is the percentage change in the balance of allowances for doubtful accounts from December 31, 2018 to December 31, 2019?	What is the percentage change in the ending balance of allowance for doubtful accounts from 2018 to 2019?	0.0901

Table 11: Samples from synthetic TATQA which are closest to the real TATQA dataset.

S. No.	Question from synthetic TATQA dataset	Nearest neighbor question from real TATQA dataset	Cosine distance
1	What is the total amount charged to costs and expenses for Allowance for Deferred Tax Assets over the three fiscal years?	What is the percentage change in the allowance for deferred tax assets at the end of period between 2018 and 2019?	0.281
2	What is the total amount added to the net book value from additions and transfers between classes for Software under development during the year ended 30 June 2019?	What was the change in net book amount for software under development between 2018 and 2019?	0.289
3	What is the ratio of the current portion to the noncurrent portion of total financing receivables, net at December 31, 2019?	What was the difference in the reported total between current and noncurrent financing receivables?	0.295
4	What was the total revenue change attributable to the foreign exchange impact for the American broadband services segment for the three months ended August 31, 2019?	What is the average Revenue between Canadian and American broadband services for year ended August 31, 2019?	0.301
5	What is the total cost for Staff costs, Contractor costs, Depreciation of property, plant and equipment, and Amortisation of intangible assets for the year 2019?	What is the average Depreciation and amortisation for 2017-2019?	0.307
6	What is the percentage contribution of Mobile and ancillary net revenues to the Total consolidated net revenues for the year 2019?	What is the percentage of total consolidated net revenues in 2019 that consists of net revenue from PC?	0.316
7	What is the net effect on total assets due to the adoption of the New Revenue Standard as of March 31, 2019?	What is the change in total assets from 2018 to 2019?	0.326
8	What is the total amount of rent expense incurred by the Group during the fiscal years 2017 to 2019, and what is the average annual rent expense over these three years?	What is the average total operating expense from 2017 to 2019?	0.340
9	What is the total amount of additions for allowances for sales returns and price protection and other allowances over the three-year period?	What is the average allowance for impairment losses across the 3 years?	0.359
10	What is the total fair value of foreign debt and U.S. debt as of December 31, 2019?	What is the percentage of Total long-term debt, less current portion to Total debt as of December 31, 2019?	0.390

Table 12: Samples from synthetic TATQA which are farthest to the real TATQA dataset.

S. No.	Question from synthetic FinQA dataset	Nearest neighbor question from real FinQA dataset	Cosine distance
1	What is the percentage change in total wholesale credit-related assets from 2012 to 2013?	what was the percentage change in total wholesale credit-related assets from 2012 to 2013?	0.017
2	What is the percentage increase in general and administrative expenses from 2011 to 2012?	what was the percentage change in the general and administrative expenses in 2012	0.086
3	What was the percentage increase in net sales for North American Industrial Packaging from 2010 to 2012?	what was the percentage change in the north american industrial packaging net sales in 2012	0.104
4	What is the average cumulative total return of United Parcel Service, Inc. over the five years from 12/31/06 to 12/31/10?	what was the percentage five year cumulative total return for united parcel service inc . for the period ended 12/31/07?	0.116
5	What was the average weighted-average grant date fair value of Nonvested Incentive/Performance Units in 2015 and 2016?	what was the average weighted-average grant-date fair value of incentive/ performance unit share awards and restricted stock/unit awards granted in 2012 and 2011?	0.127
6	What is the difference between the weighted average grant date fair value per share for the years ended December 31, 2010 and 2009?	what was the difference in the weighted average grant-date fair value per share between 2012 and 2013?	0.135
7	What is the total occupied square footage of the properties with lease expiration dates in 2020 and 2028?	considering the properties with lease expiration dates in 2020 , what is the average occupied square footage?	0.144
8	What is the percentage change in the total net of all collateral from 2015 to 2016?	what was the percentage change in collateral posted between 2013 and 2014?	0.151
9	By how much did the operating income margin increase from 2009 to 2011?	what was the percent of the increase in the operating income from 2010 to 2011	0.158
10	What was the percentage change in net sales from 2011 to 2013 for Space Systems?	what were average net sales for space systems from 2011 to 2013 in millions?	0.163

Table 13: Samples from synthetic FinQA which are closest to the real FinQA dataset.

S. No.	Question from synthetic FinQA dataset	Nearest neighbor question from real FinQA dataset	Cosine distance
1	What is the difference between the non-cash operating activities and the sum of pension and postretirement plan contributions and changes in working capital and other noncurrent assets and liabilities for the year 2012?	what percentage of net cash from operating activities was derived from non-cash operating activities in 2012?	0.367
2	What is the total number of rooms in hotels that are either owned or have land leases expiring after 2030?	what is the total square feet of buildings whose lease will expire in 2020?	0.373
3	What is the ratio of the total value of acquired in-place leases to the total assets acquired from the 2007 acquisition of Reckson?	what is the ratio of total assets acquired to total liabilities assumed?	0.378
4	What was the total amount of pension settlement losses recognized in 2018 and 2019 combined, before tax?	what would the ending amount of unrecognized tax benefits for 2015 be (in millions) without settlements?	0.384
5	What is the difference between the preliminary estimated fair values of customer-related intangible assets and acquired technology as of May 31, 2016?	for acquisitions in 2017 what percentage of recorded a total acquired intangible assets was in-process technology?	0.391
6	What is the difference between the sum of remaining net rentals and estimated unguaranteed residual value in 2010 and the sum of non-recourse mortgage debt and unearned and deferred income in 2009?	from 2005-2006 , what was the total amount of remaining net rentals , in millions?	0.399
7	What is the difference between the total assets and the sum of Global Core Liquid Assets (GCLA) and Secured Client Financing for the year 2016?	by what amount is the total gains/ (losses) on financial assets and financial liabilities at fair value at 2017 different from 2016?	0.408
8	What is the ratio of the total number of transactions to the number of cards in circulation for MasterCard, and is this ratio greater than 0.017?	what was the percent of the growth of the mastercard from 2013 to 2014	0.422
9	What is the difference between the fair value of developed technology and the total liabilities assumed as of the Implex acquisition date?	what was the change in the fair value of the debt acquisition date fair value of the borrowings	0.439
10	What is the sum of 'Capital stock', 'Paid-in surplus', 'Retained earnings', and 'Treasury stock'?	how is the treasury stock affected after the stock repurchases in the last three months of 2016 , (in millions) ?	0.468

Table 14: Samples from synthetic FinQA which are farthest to the real FinQA dataset.