FM^{3}QA: Financial Multimodal Multi-Hop Multiple-Choice QA over Text, Tables, and Charts

Anonymous ACL submission

Abstract

With recent advancements in large language models (LLMs), their areas of application are expanding, including the financial sector. A robust system to support financial analysts should be capable of analyzing and reasoning over information from multiple modalities, such as text, tables, and charts. However, related studies predominantly focus on processing single modalities or partial integration, failing to handle all modalities simultaneously. To mitigate this limitation, we propose a new framework called M^{3}QAGen, designed for automatically creating Multimodal Multi-Hop Multiple-Choice QA datasets. Utilizing this framework, we introduce a dataset called FM^{3}QA (Financial Multimodal Multi-Hop Multiple-Choice QA). Based on corporate Annual Simplified Financial Statements and 10-K reports, this benchmark presents queries that demand multi-hop reasoning across various modalities. Each question in the dataset is carefully designed to simulate real-world corporate analysis workflows, requiring skills such as numerical reasoning, ranking, and fact-checking. Through extensive evaluation with GPT-4o, we demonstrate that FM^{3}QA is a challenging benchmark even for state-of-the-art LLMs, pushing the boundaries of improving LLMs' capability to effectively engage with multimodal financial data. The proposed M^{3}QAGen framework also proves to be efficient and scalable, potentially applicable to other domains beyond finance.

1 Introduction

Despite recent advancements in artificial intelligence, many challenges remain in developing systems capable of achieving human-level understanding and reasoning. Humans perceive and understand the world through multiple senses, including vision, hearing, and touch. While research on models that process multi-modal data is gaining significant attention (Baltrušaitis et al., 2018; Radford et al., 2021; *inter alia*), it still lags behind progress in designing AI that can integrate information from multiple modalities, akin to human cognition.

Multimodal question answering (QA) has been regarded as a crucial and promising research topic that can contribute to mitigating the aforementioned issue (Antol et al., 2015; Kafle and Kanan, 2017). The task requires combining data from various modalities—such as text, tables, and images—to answer questions. However, most existing studies in the field focus on handling queries that can be answered using only one or two modalities, without fully addressing real-world scenarios where information from multiple data sources must be interconnected (Talmor et al., 2021; Chen et al., 2020, 2021a). For instance, a financial analyst may need to carry out the following tasks simultaneously: review a company's Annual Simplified Financial Statements presented in tables to make investment decisions, interpret trends from charts (images), and verify findings through textual information in financial reports. This example highlights the importance of benchmarking the ability of LLMs for cross-modal QA, which demands exploring and combining information from different modalities.

To construct a dataset that meets this desideratum, securing data sources with several interrelated modalities is essential. Existing research on multimodal QA (Talmor et al., 2021; Chen et al., 2020, 2021a) has largely relied on general-purpose knowledge bases such as Wikipedia. While Wikipedia provides a reasonable foundation for dataset construction, its coverage is often insufficient for creating domain-specific examples.

In this work, we aim to evaluate the capability of LLMs to tackle multimodal multi-hop QA in specialized domains, particularly in the financial domain. The main challenge lies in constructing QA instances that require a comprehensive understanding of all modalities. This is relatively straightforward if one hinges on Wikipedia, where text, tables, and images within a single document

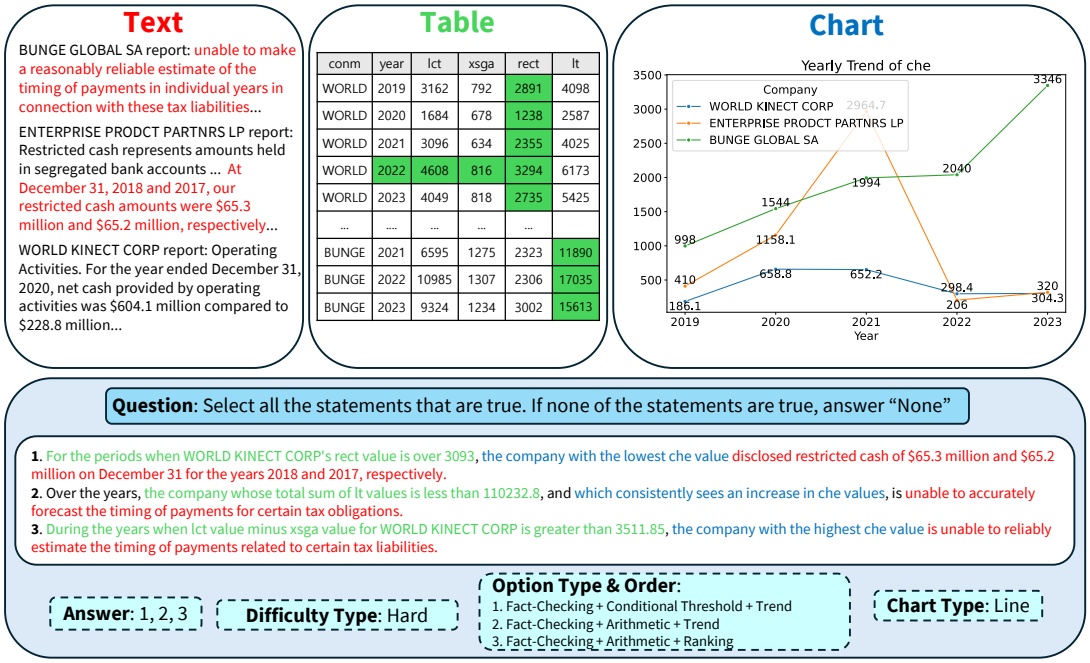


Figure 1: An example from FM^{3}QA. For the Hard difficulty level, all options are designed to require reasoning across three modalities. The information indicated by dashed lines is not included in the actual QA data instance.

are inherently interconnected. However, in cases where we need to utilize data from different modalities and domains without explicit connections, we should identify a solution to automatically establish links between these distinct modalities.

To this end, we first propose M^{3}QAGen, a framework designed to automatically **gen**erate coherent **M**ultimodal **M**ulti-Hop **M**ultiple-Choice **QA** datasets. This robust framework enables the generation of these types of datasets as long as at least one common entity is shared between the text source and the table source. Therefore, it becomes possible to efficiently create challenging cross-modal QA datasets specialized in various domains, including finance and material science.

Furthermore, as an ongoing demonstration of M^{3}QAGen, we introduce a novel benchmark in the financial domain, titled FM^{3}QA (**F**inancial **M**ultimodal **M**ulti-Hop **M**ultiple-Choice **QA**; see Figure 1). It consists of cross-modal 3-hop QA examples using text, tables, and charts.[1] In experiments, we show that state-of-the-art LLMs, such as GPT-4, struggle to effectively solve the problems posed by our benchmark. This suggests that FM^{3}QA presents highly challenging tasks and that current LLMs are still limited in their ability to integrate information across multiple modalities.

Finally, we explore the potential of applying M^{3}QAGen to other domains, particularly material science, demonstrating that the proposed framework could offer significant benefits for multimodal QA in various specialized fields.

2 Related Work

Existing multimodal multi-hop QA datasets can be classified into Text + Image, Text + Table, and Text + Table + Image datasets, as shown in Table 1.

2.1 Text + Image Multi-Hop QA

WebQA (Chang et al., 2022) introduces multimodal multi-hop QA by combining multiple text and image sources, drawing text from Wikipedia and images from Wikimedia Commons. However, its limitation lies in the fact that the text is not closely tied to the images, as they only share broad categorical connections rather than specific relevance. On the other hand, MuMuQA (Reddy et al., 2022) leverages news articles as a data source, generating questions that describe individuals in the text and replacing names with visual information derived from images. In this approach, the text and images are obtained from the same news article, ensuring a close relationship between the two modalities. Nonetheless, neither approach incorporates the table modality, which is crucial for financial applications. To address this limitation, FM^{3}QA

[1]The dataset is derived from Annual Simplified Financial Statements and 10-K reports, as detailed in Appendix A.1.

Dataset	Modality				Domain			Evaluation	
	Text	Table	Image	Image Type	Text	Table	Image	Answer Type	Metric
WebQA	✓	✗	✓	Photographic	Wikipedia	-	Wikipedia	Descriptive	F1, BARTScore
MuMuQA	✓	✗	✓	Photographic	News	-	News	Short span	F1
FinQA	✓	✓	✗	-	Finance	Finance	-	Short span	Execution Acc, Program Acc
TAT-QA	✓	✓	✗	-	Finance	Finance	-	Short span	F1, Exact Match
HybridQA	✓	✓	✗	-	Wikipedia	Wikipedia	-	Short span	F1, Exact Match
OTT-QA	✓	✓	✗	-	Wikipedia	Wikipedia	-	Short span	F1, Exact Match
TANQ	✓	✓	✗	-	Wikipedia	Wikipedia	-	Table Format	F1
MMQA	✓	✓	✓	Photographic	Wikipedia	Wikipedia	Wikipedia	Short span	F1, Exact Match
MMCoQA	✓	✓	✓	Photographic	Wikipedia	Wikipedia	Wikipedia	Short span	F1, Exact Match
FM^{3}QA (Ours)	✓	✓	✓	Chart	Finance	Finance	Finance	Multi-Choice	Accuracy

Table 1: Comparison of existing multimodal multi-hop datasets with FM^{3}QA. **Modality** indicates which modalities are utilized in each dataset. **Domain** describes the domain of each modality. **Evaluation** explains the format of answers and evaluation metrics. FM^{3}QA is unique in that it incorporates all three modalities, focuses on the finance domain (including charts), and requires the verification of every answer candidate related to each query.

is designed to include tables alongside text and images, while maintaining strong connections between all modalities.

2.2 Text + Table Multi-Hop QA

OTT-QA (Chen et al., 2021a) and HybridQA (Chen et al., 2020) are multi-hop QA datasets built using tables from Wikipedia and the related text linked to those tables. TANQ (Akhtar et al., 2024) is another dataset where answers must be provided in table format, derived from the information found on Wikipedia. These datasets share the common limitation of being restricted to the Wikipedia domain, which is highly likely to have been extensively exposed to recent LLMs during their pre-training.

For domain-specific applications of multi-modal QA systems, particularly in the financial domain, FinQA (Chen et al., 2021b) and TAT-QA (Zhu et al., 2021) are constructed using texts and tables from financial annual reports. However, they still have shortcomings due to the lack of images in their datasets. In addition, these datasets emphasize basic numerical tasks such as arithmetic operations and value comparisons, without adequately capturing the complex and unique characteristics of the financial domain. In contrast, FM^{3}QA includes the image modality aligned with text and tables, and introduces diverse option types that capture the distinct complexities of financial problems.

2.3 Text + Table + Image Multi-Hop QA

MMQA (Talmor et al., 2021) is one of the few and representative datasets constructed by aligning text, tables, and images simultaneously.[2] However, the dataset is again based on a general knowledge base, i.e., Wikipedia, making it unsuitable for evaluating the performance of domain-specific QA systems.[3] Additionally, MMQA is constrained by the fact that only 0.8% of the entire dataset involves cross-modal 3-hop QA, which requires comprehensive reasoning over text, tables, and images at once. It also relies on human annotators, leading to a high cost of \$0.33 per question, while our framework reduces costs to \$0.004 per question by employing LLMs. We refer readers to Table 7 for a detailed comparison between MMQA and FM^{3}QA.

3 M^{3}QAGen: A Framework for Automatic M^{3}QA Generation

M^{3}QAGen, the framework proposed in this study, automatically generates **M**ultimodal **M**ulti-Hop **M**ultiple-Choice (=**M**^{3}) QA by reasoning over text, tables, and charts. It only requires table and text sources that share a common entity as the foundation. This corresponds to the Modular Part, and by simply replacing this Part, data from other domains can also be easily generated using the automated M^{3}QAGen framework. In this section, we outline the procedure for constructing the Financial Multimodal Multi-Hop Multiple-Choice QA dataset, referred to as FM^{3}QA, based on the proposed framework.

[2] MMCoQA (Li et al., 2022) is a variation of MMQA, with a focus on conversational interactions.

[3] One reason most cross-modal QA datasets rely on Wikipedia is that its structure naturally provides well-organized connections between text, tables, and images.

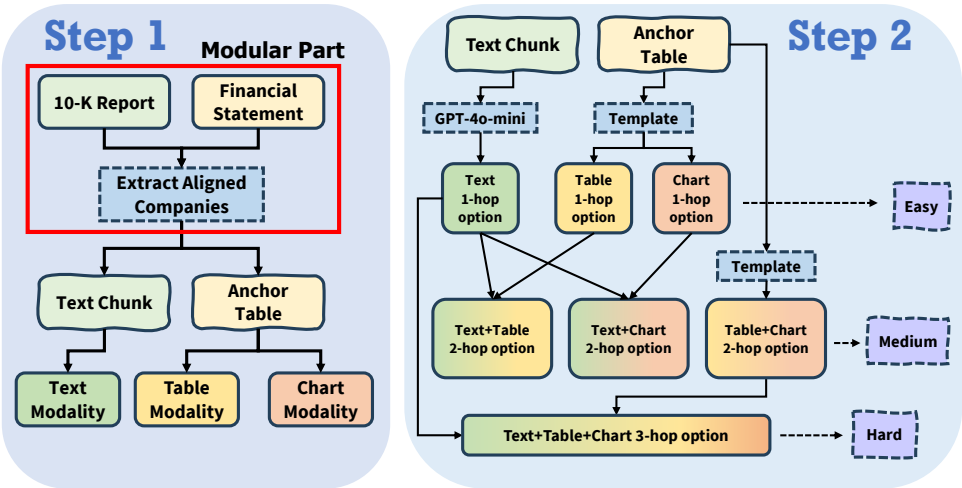


Figure 2: Dataset generation pipeline of the **M^{3}QAGen** framework. If only the Modular Part is replaced when applying to different domain data, the M^{3}QAGen pipeline allows for the easy creation of multimodal multi-hop multiple-choice QA. In this section, we introduce the process of creating FM^{3}QA using 10-K reports and Annual Simplified Financial Statements. **Step 1:** After selecting common companies between the 10-K report and the Annual Simplified Financial Statements, text chunks are extracted from the 10-K report, and anchor tables are obtained from the Annual Simplified Financial Statements. These are then converted into text, table, and chart modality data. **Step 2:** This step involves generating multiple-choice options using text chunks and anchor tables. First, unimodal 1-hop options consisting of a single modality are created, followed by the construction of cross-modal 2-hop options by combining them based on corporate entities. Similarly, cross-modal 3-hop options are built through the combination of 1-hop and cross-modal 2-hop options.

3.1 Source of Data

3.1.1 WRDS Compustat: Annual Simplified Financial Statements

Considering the practicality of multimodal multi-hop QA, we utilize *annual financial statements*, an essential element in real corporate analysis, as the table data source. Wharton Research Data Services (WRDS) Compustat[4] provides various financial data of publicly traded companies in North America. Among them, we use the **Annual Simplified Financial Statement**, which includes key financial columns such as Company Name, Ticker Code, Year, Net Sales, and Total Assets, spanning multiple years for each company. Each column of data can be found in Figure 8. And detailed description of this Annual Simplified Financial Statement is provided in Appendix A.1. This Annual Simplified Financial Statement will later be used to construct the table and chart modalities.

Preprocessing The raw dataset contains a total of 80 columns. We standardize the unique symbol IDs to ticker codes and perform preprocessing to remove two columns that are not in millions of units to unify the units by column, leaving a total of 70 columns. We also use data from the most recent five years, 2019 to 2023.

3.1.2 SEC EDGAR: 10-K Report

To construct texts that are closely related to the Annual Simplified Financial Statement of company entities, we focus on corporate financial reports. Companies listed on the U.S. stock market are required to periodically provide financial reports to the U.S. Securities and Exchange Commission (SEC), and these reports can be publicly accessed through the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR).[5] We use the annual disclosure report, the 10-K report, of companies to match the Annual Simplified Financial Statement. This 10-K report differs from the summary-style annual reports typically used in datasets such as FinQA (Chen et al., 2021b) and TAT-QA (Zhu et al., 2021), as it provides more in-depth financial data and disclosures. All companies' 10-K reports are composed of a common table of contents format. Among the various items in the 10-K report, we use five key items—ITEM 1, 2, 7, 7A, and 8—that are frequently utilized in actual corporate analysis. The detailed descriptions of the items in the 10-K report are provided in Ap-

[4] https://wrds-www.wharton.upenn.edu/pages/grid-items/compustat-global-wrds-basics/

[5] https://www.sec.gov/search-filings

pendix A.1. This 10-K report data will later be used to construct the text modality.

Preprocessing To align with the Annual Simplified Financial Statement data, we filter companies where both the Annual Simplified Financial Statements and 10-K annual reports exist, ensuring all formats of reports for the most recent five years are fully present. Among them, we select the top 101 companies based on Net Sales in 2023.

3.2 Anchor Table Construction

An anchor table serves as an intermediate bridge connecting the text, table, and chart modalities and is used as a base anchor for creating multi-hop options in multiple-choice QA. After sampling three companies from the Annual Simplified Financial Statement data of the 101 companies, we construct the anchor table by randomly sampling seven financial columns excluding company name, ticker code, and year. One column, used for chart generation, is chosen to avoid NaN values. The final generated anchor table consists of the Annual Simplified Financial Statement data of three companies, each having five years, and is composed of ten columns.

3.3 Text Chunk and Fact Creation

The 10-K reports obtained through SEC EDGAR are too lengthy to use entirely at once as text modality. Therefore, we divide the 10-K reports of each company into chunks of three consecutive paragraphs and use GPT-4o-mini to create three facts for each chunk. Later, these chunks will be used as the text modality, and the facts will be used to generate options for the multiple-choice questions.

3.4 Text, Table, Chart Data Construction

The data refers to each dataset that the language model needs to see and reason upon when the QA is provided, and a single data instance contains a total of three companies. From the anchor table with three companies, we select one column without NaN values as the chart column and convert the anchor table excluding the chart column into the table modality and the chart column into the chart modality. To preserve the structural information of the table, the table modality is constructed in JSON format, and to ensure data diversity, the chart modality uses three different libraries and four chart types (line, bar, scatter, pie) commonly used in financial domains. The text modality corresponds to the text chunks of the three companies.

Difficulty	Shot	Total	0 correct	1 correct	2 correct	3 correct
Easy	0-shot	82.0	89.58	91.67	75.00	71.88
	1-shot	80.2	83.33	89.58	78.12	69.79
Medium	0-shot	58.9	69.79	57.29	59.38	48.96
	1-shot	65.8	84.38	65.62	61.46	51.04
Hard	0-shot	34.9	40.62	36.46	29.17	33.33
	1-shot	37.9	47.92	36.46	30.21	36.46

Table 2: Accuracy analysis by difficulty level on the benchmark dataset, based on GPT-4. All values are expressed as percentages. 'n correct' indicates that there are n correct options among the choices

All of these processes are automated through a Python script.

3.5 Text, Table, Chart Multi-Hop Multiple-Choice QA Generation

To construct Multi-hop Multiple-Choice QAs, we develop three options per instance. Although each option is 1-hop, the presence of three options that require understanding different modalities collectively forms a multimodal 3-hop question. Inspired by this, we design datasets of three difficulty levels: Easy, Medium, and Hard, where each option is 1-hop, cross-modal 2-hop, and cross-modal 3-hop, respectively. In creating multi-hop options, we avoid the simple method of concatenating single-hop options and instead adopt a combination approach based on corporate entities to produce more natural options. We design five fundamental base options: Fact-Checking, Conditional Threshold, Arithmetic, Trend, and Ranking. Based on different combinations across difficulty levels, we create various option type templates. The option types according to modality are presented in Table 6, and the examples and templates for each option type can be found in Table 4 and Table 5.

As shown in Table 1, most previous studies focus on short-span or explanatory answer types. However, as argued in recent research (Pang et al., 2024), metrics such as F1 Score and Exact Match are insufficient for properly evaluating QA datasets. Therefore, we adopt the multiple-choice QA format proposed in that study. Additionally, to enable more nuanced evaluations, we construct questions in the Type K multiple-choice format, where all correct options must be selected, going beyond the conventional multiple-choice questions (Type A) with a single correct answer. We consider cases where none of the given options are correct, designing the number of correct answers to be 0, 1, 2, or 3. However, multiple-choice QA also has lim-

itations, as it can introduce bias in the options. To minimize biases that could arise in multiple-choice QA and to enhance dataset diversity, we construct each difficulty level's dataset by ensuring that the proportions of correct answer types, option types, option order types, and chart types are all equal.

Difficulty (Easy) Each multiple-choice option consists of a 1-hop option that requires reasoning on one of the modalities: Text, Table, or Chart. In this process, we aim to construct practical and natural option types by considering the characteristics of financial data and each modality. Specifically, numerical reasoning for tables and visual structure reasoning for charts are essential in the financial domain. Thus, we subdivide the 1-hop options for tables into *Conditional Threshold* and *Arithmetic* types, and those for charts into *Trend* and *Ranking* types. The 1-hop option for text is a *Fact-Checking* type based on a chunk from the 10-K report.

Difficulty (Medium) We generate cross-modal 2-hop options, namely Text + Table, Text + Chart, and Table + Chart, by combining the 1-hop options obtained above based on corporate entities. These options require a higher level of reasoning ability compared to the Easy level and are well-aligned with real human scenarios.

Difficulty (Hard) Each multiple-choice option consists of a cross-modal 3-hop option that requires reasoning across all modalities: Text + Table + Chart, demanding the highest level of reasoning ability. We construct these cross-modal 3-hop options by combining a Text 1-hop option and a Table + Chart cross-modal 2-hop option based on corporate entities.

Distractor Generation Instead of simply adjusting numerical values to generate incorrect options, we reflect realistic scenarios in the financial domain, where analysis of multiple companies is common, by generating distractors based on corporate entities. Since each multiple-choice option in Easy, Medium, and Hard levels is combined with corporate entities, we generate distractors by replacing them with other companies in the same instance.

3.6 Two-Stage Paraphrasing

Based on the base templates for each option type, we apply a two-stage paraphrasing technique to enhance sentence diversity in each instance. First, we perform a lexical paraphrasing iteration to increase vocabulary diversity, followed by syntax paraphrasing on all instances to diversify syntactic structures. For the two-stage paraphrasing, we use GPT-4o for the benchmark dataset. The dataset undergoes meticulous review by human annotators with NLP expertise to correct any awkward or ambiguous expressions.

3.7 Efficiency and Generality of M^{3}QAGen

M^{3}QAGen, the dataset generation framework introduced in this paper is based on an automated pipeline, enabling the construction of large-scale cross-modal 3-hop QA datasets at low cost, demonstrating superior efficiency compared to previous studies. Additionally, the base templates in this framework are created using Python scripts, ensuring that no computational errors occur during the generation process. Moreover, our proposed pipeline allows for the creation of high-quality cross-modal 3-hop QA datasets across any domain, due to the ease of replacing or detaching the Modular Part. This contributes significantly to the general applicability of the framework, extending beyond the financial domain to fields such as medicine, law, and science. To demonstrate this, we apply this framework to the material science domain to easily construct the M^{3}QA dataset, which will be further discussed in Section 6.

4 Analysis of the FM^{3}QA Dataset

4.1 Dataset Diversity

We construct a dataset, consisting of 1,152 samples: 384 easy, 384 medium, and 384 hard. Although we do not provide a dataset of large data instances, if needed, data augmentation is trivial thanks to the M^{3}QAGen framework. We use a total of 101 companies to create the anchor tables and text chunks, utilizing the most recent five years of data for each company. Given that the average number of facts per annual report is 380, the total number of facts across 101 companies over five years amounts to approximately 191,900. Each fact comprises five ITEMs from the 10-K reports, allowing us to construct a dataset with substantial volume and diversity through various combinations. Additionally, the dataset's diversity is significantly enhanced by the combinatorial approach of selecting 7 columns from 67 financial columns and sampling 3 companies from 101, leading to numerous potential combinations. We use three libraries—matplotlib, seaborn, and plotly—for charts, including four types: bar, line, scatter, and pie charts.

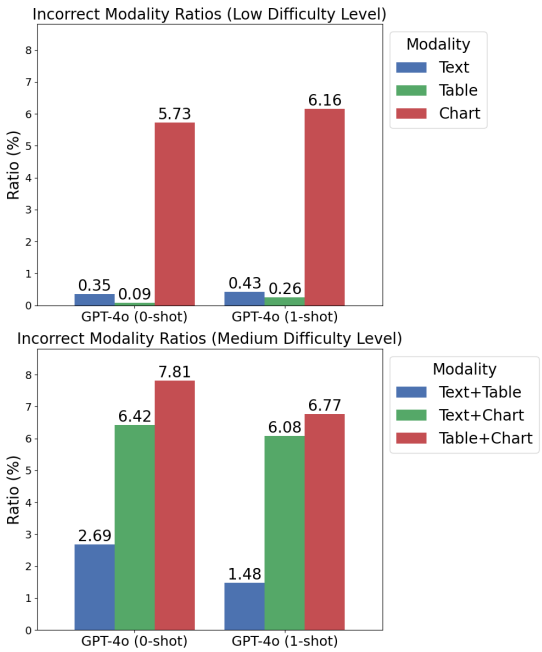


Figure 3: The ratio of incorrect answers by modality across all options for the Low and Medium difficulty levels. For example, in the Low difficulty level, if the correct options are Text and Table but the model predicts Text and Chart, both the Table and Chart modalities are considered incorrect for that question.

Difficulty	Shot	Total	line	bar	scatter	pie
Easy	0-shot	82.0	85.2	84.9	75.7	81.4
	1-shot	80.2	82.8	84.9	71.0	83.7
Medium	0-shot	58.9	61.0	59.0	56.3	-
	1-shot	65.8	68.7	69.4	58.0	-
Hard	0-shot	34.9	36.6	40.2	27.1	-
	1-shot	37.9	42.5	37.9	32.2	-

Table 3: Accuracy by chart type, based on GPT-4. All values are presented as percentages. Pie charts are only used in the Ranking option type of the Easy difficulty, as they are unsuitable for the Trend option type.

4.2 Dataset Quality

Since the base templates requiring mathematical calculations (Table 5)—such as values exceeding certain thresholds, continuous increases, maximum values, and arithmetic operations—in each option type are based on Python scripts, they are error-free. However, errors could occur due to LLM hallucinations after the two-stage paraphrasing. To address this, human annotators with NLP expertise review all 1152 data samples. GPT-4o achieves a 91% success rate in paraphrasing, and we perform filtering on failed cases by identifying patterns.

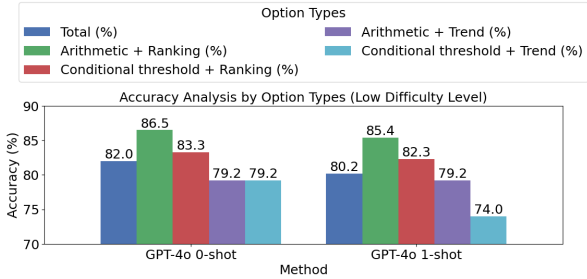


Figure 4: Accuracy analysis by option types for the Low difficulty level, which includes four option types.

5 Experiments

5.1 Overall Performance Evaluation

We conduct experiments to assess whether GPT-4o can effectively solve the FM^{3}QA dataset in both zero-shot and one-shot settings. The prompts can be found in Figure 6 and Figure 7.

Table 2 presents the evaluation results on the FM^{3}QA dataset. When comparing total accuracy, both the zero-shot and one-shot settings show higher accuracy in the order of Easy, Medium, and Hard, indicating that the difficulty levels are effectively set. In the Easy difficulty, the zero-shot model achieves a high accuracy of 82%, likely because LLMs possess the ability to decompose multiple-choice questions into individual options and solve them sequentially. However, the accuracy drops to 58.9% in Medium and 34.9% in Hard difficulties. In the one-shot setting, while there is a slight performance improvement in the Medium and Hard levels, even the state-of-the-art GPT-4 struggles, demonstrating that the FM^{3}QA dataset presents a significant challenge. Furthermore, when measuring accuracy based on the number of correct answers, we observe considerable variance. In all difficulty levels, cases with multiple correct answers exhibit lower accuracy—by over 20%—compared to cases with zero or one correct answer. This indicates that the number of correct options significantly impacts model evaluation. These findings suggest that Type K multiple-choice questions, which require selecting all correct options, are more effective than Type A questions in evaluating whether a language model truly understands the task.

5.2 Difficulty Analysis by Modality

In the Easy and Medium difficulties, we can automatically analyze which modalities the language model finds more challenging in the task. For the

Hard difficulty, where each option is a cross-modal 3-hop QA with a lengthy reasoning process, we manually analyze incorrect cases by sampling. Figure 3 shows the proportion of incorrect modality options out of the total number of options in the Easy and Medium difficulties, respectively.

In the Easy difficulty, the model struggles most with charts, while tables are slightly easier to analyze than text. This indicates that for simple 1-hop options, the model's reasoning difficulty increases in the order of Chart, Text, and Table. In the Medium difficulty, options involving charts have higher proportions of incorrect answers. This indicates that when dealing with complex cross-modal 2-hop options, the model finds reasoning more difficult in the order of Chart, Table, and Text. In the Hard difficulty, incorrect modalities are 4.4% for text, 17.8% for tables, and 28.9% for charts, with the chart modality posing the greatest challenge, consistent with Easy and Medium levels. Common errors in charts include confusing companies, misreading numbers, or referencing values from incorrect years. Errors in tables often involve incorrect arithmetic operations (especially with negative numbers) or only partially identifying entities that satisfy given conditions. In the text modality, failures are frequently due to not correctly inferring the company's report in question.

5.3 Difficulty Analysis by Chart Type

To determine which chart types the model finds most challenging, we conduct an accuracy analysis by chart type, as shown in Table 3. Experimental results show that, across all difficulty levels, accuracy is lowest for scatter charts, suggesting that the language model has particular difficulty analyzing scatter charts among chart modalities. This may be because scatter charts are less structured and more complex than other chart types like line or bar charts. Line and bar charts have visually clear trends and comparisons, making interpretation relatively straightforward. In contrast, scatter charts demand more advanced visual reasoning to extract meaningful insights, making it challenging for the language model to accurately interpret the information.

5.4 Difficulty Analysis by Option Type

We also analyze accuracy across different option types—Conditional Threshold, Arithmetic, Trend, and Ranking—with the results presented in Figure 4. In both zero-shot and one-shot settings, the model consistently finds the 'Trend' option type more challenging than the 'Ranking' type in the chart modality. This is likely because the Ranking type requires simply identifying the largest or smallest company, whereas the Trend type necessitates determining whether a column's values in a chart are continuously increasing or decreasing, requiring more detailed visual structure reasoning.

6 Expansion to Another Domain

M^{3}QAGen, our proposed dataset construction methodology presents a generalizable framework that can be easily applied to new domains. With the ease of replacing the Modular Part in our framework, as long as aligned entities exist between text and table sources, a high-quality M^{3}QA can be generated through an automated pipeline. To demonstrate this, we apply the framework to the domain of Material Science, further constructing a domain-specific M^{3}QA dataset. In this case, tables containing material properties and corresponding textual descriptions can be a Modular Part. Then, the M^{3}QAGen can seamlessly generate 1-hop, cross-modal 2-hop, and cross-modal 3-hop answer choices, resulting in a Material-M^{3}QA dataset.[6] The reason for selecting the material science domain is to demonstrate that our M^{3}QAGen framework can be applied to generate cross-modal 3-hop questions in various domains of interest. This contribution offers a significant advancement by moving beyond the limited scope of Wikipedia-centered cross-modal 3-hop QA tasks.

7 Conclusion

We introduce M^{3}QAGen, a framework designed for creating Multimodal Multi-Hop Multiple-Choice QA datasets. Utilizing this framework, we create FM^{3}QA, the first benchmark to assess the cross-modal 3-hop capabilities of LLMs in the financial domain. We specifically evaluate the performance of GPT-4, showing that current LLMs still struggle with integrating information from different modalities. Additionally, we demonstrate the scalability of M^{3}QAGen by applying it not only to the financial domain but also to material science. We expect the adoption of our framework by specialists across various fields will uncover the current capabilities and limitations of LLMs in processing information within specialized domains.

[6]The detailed process of creating Material-M^{3}QA is outlined in Appendix A.2.

Limitations

Our study has three limitations. First, despite the scalability of the proposed data generation framework, we have applied it only to the financial domain. Second, although there are various LLMs such as Claude and Gemini, we have conducted experiments solely on GPT-4o. Lastly, due to computational resource constraints, we have not trained models like Llama.

References

Mubashara Akhtar, Chenxi Pang, Andreea Marzoca, Yasemin Altun, and Julian Martin Eisenschlos. 2024. Tanq: An open domain dataset of table answered questions. *arXiv preprint arXiv:2405.07765*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021a. Open question answering over tables and text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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: EMNLP 2020*, pages 1026–1036, Online. 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. 2021b. 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.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4220–4231.

Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. 2024. Uncovering limitations of large language models in information seeking from tables. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1388–1409. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Revant Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, et al. 2022. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11200–11208.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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, ACL/IJCNLP 2021*, pages 3277–3287. Association for Computational Linguistics.

A Appendix

A.1 Source of Data

Annual Simplified Financial Statement The Annual Simplified Financial Statement consists of 80 columns, which are broadly classified into four categories: Identifying Information, Balance Sheet Variables, Income Statement Variables, and Statement of Cash Flows Variables. The components of each category are presented in Figure 8, which correspond to the columns of the Annual Simplified Financial Statement Table.

10-K Report The 10-K report is a comprehensive document filed annually by U.S. public companies with the Securities and Exchange Commission (SEC). It provides a detailed overview of the company's business operations, financial condition, and any associated risks. The report is structured to include mandatory sections, such as a description of the company's primary business activities, financial results, and management's analysis. This information is critical for investors, analysts, and other stakeholders who wish to gain a thorough understanding of a company's performance and risks.

A.2 Details of Material-M^{3}QA Generation

Each 10-K report includes several key items that are vital for corporate analysis. For instance, Item 1 provides a description of the company's business model, its products or services, and its primary markets. Item 7, often referred to as the Management's Discussion and Analysis (MD&A), allows company executives to discuss operational results, providing insight into trends, risks, and strategies. In addition, Item 7A covers quantitative and qualitative disclosures regarding market risks, while Item 8 presents the audited financial statements, offering a transparent view of the company's financial health. These items, along with other sections, make the 10-K an essential document for evaluating a company's long-term viability and strategy.

Among the various items of the 10-K report, we use ITEM 1 (Business), ITEM 2 (Legal Proceedings), ITEM 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations), ITEM 7A (Quantitative and Qualitative Disclosures about Market Risk), and ITEM 8 (Financial Statements and Supplementary Data), which are most commonly used in actual corporate analysis. Further details can be found in the document provided by the SEC[7].

We construct a Table Source containing material properties and a Text Source describing the materials through The Materials Project. The Materials Project[8] is an initiative that accelerates materials discovery by providing researchers with computational data and tools to predict material properties, enabling more targeted and efficient experimental research. Specifically, we transform the entity corresponding to companies in FM^{3}QA into materials. Following the same steps proposed in our framework, we build Anchor Tables and Text Chunks for Material-M^{3}QA. The Anchor Tables include columns for material properties such as band gap, density, volume, and formation energy per atom, while the Text Chunks provide descriptions of the material's crystal structure, structural characteristics, and properties. Using these Text Chunks and Anchor Tables, we create datasets combining text, table, and chart modality data. Subsequently, we generate 1-hop, cross-modal 2-hop, and cross-modal 3-hop options through templates, categorizing them by difficulty level. A two-stage paraphrasing process is employed to maximize diversity. An example of the dataset instance is shown in Figure 5.

[7] https://www.sec.gov/files/reada10k.pdf
[8] https://next-gen.materialsproject.org/

Text

$NaMnF_3$ is Orthorhombic Perovskite structured and crystallizes in the orthorhombic Pnma space group. Na^{1+} is bonded in a 8-coordinate geometry to eight F^{1-} atoms...

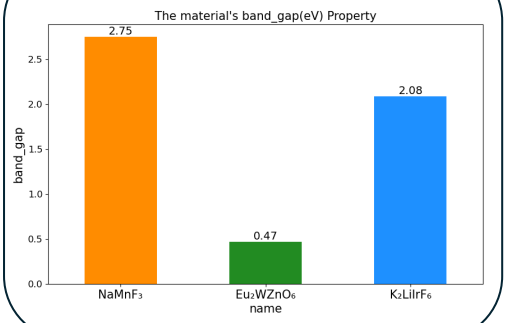
Eu_2WZnO_6 is (Cubic) Perovskite-derived structured and crystallizes in the cubic Fm̄3m space group. Eu^{2+} is bonded to twelve equivalent O^{2-} atoms to form EuO_{12}...

K_2LiIrF_6 is (Cubic) Perovskite-derived structured and crystallizes in the cubic Fm̄3m space group. K^{1+} is bonded to twelve equivalent F^{1-} atoms to form KF_{12} cuboctahedra

Table

name	density ($g \cdot cm^{-3}$)	volume ($Å^3$)	formation_energy_per_atom (eV/atom)
$NaMnF_3$	3.5897	249.651	-2.9655
Eu_2WZnO_6	8.6393	124.7754	-2.6097
K_2LiIrF_6	4.6546	139.6134	-2.5189

Chart



Question: Select all the statements that are true. If none of the statements are true, answer "None"

1. The material with the smallest volume contains Eu^{2+}, which is bonded to twelve equivalent O^{2-} atoms.
2. Material with a density lower than 4.1729 g/cm^3 is orthorhombic perovskite-structured and crystallize in the orthorhombic Pnma space group.
3. The material with a formation_energy_per_atom (eV/atom) less than -2.7215 has the smallest band_gap (eV).

Answer: 1, 2

Difficulty Type: Medium

Option Type & Order:
1. [Text+Chart]: Fact-Checking + Ranking
2. [Text+Table]: Fact-Checking + Conditional Threshold
3. [Table+Chart]: Conditional Threshold + Ranking

Chart Type: Bar

Figure 5: An example from Material-M^{3}QA.

Option Types	Example
FC	In May 2020, WORLD KINECT CORP modified and refreshed its asset-backed debt financing facility.
CT	The firm COSTCO WHOLESALE CORP. disclosed that its cogs values surpassed 92,765 during the year 2021.
AR	The company where the 2023 txt value minus the 2022 txt value equals 359.0 is WORLD KINECT CORP.
TR	Over the period from 2020 through 2023, SALESFORCE INC. consistently experienced an increase in its xsga values.
RK	During 2019, US FOODS HOLDING CORP. possessed the lowest txt value compared to other companies.
FC+CT	In 2023, the company whose xint values are less than 849.5 owns a terminal facility located at Fort Mifflin, which includes two docks for ships and has a total storage capacity of approximately 570 MBbls.
FC+AR	The company where the 2020 spi value minus the 2022 spi value equals 1036.0 recorded $39 million and $40 million in prior service credit amortization in 2018 and 2017, respectively.
FC+TR	The firm that showed a steady increase in dpc values between 2019 and 2023 must comply with the detailed regulations set by the Department of Transportation (DOT) regarding its pipeline infrastructure.
FC+RK	In 2022, the company, which reports fincf values greater than -2009, spreads out the amortization of its capitalized costs tied to new revenue contracts across four years.
CT+TR	Over the years, MONDELEZ INTERNATIONAL INC. consistently reports a fincf value exceeding -19575.5, while the ceq value demonstrates a continuous increase.
CT+RK	When the seq figure for WARNER BROS DISCOVERY INC falls below 10177.5, VERIZON COMMUNICATIONS INC records the lowest cogs value.
AR+TR	Throughout all periods, ENTERPRISE PRODCT PARTNRS LP is the company in which the cumulative nopi values surpass 1553.35, while the ao values have persistently increased.
AR+RK	During the years when lt value minus ibc value for UNITED PARCEL SERVICE INC is greater than 55265, the company with the lowest sale value is UNITED PARCEL SERVICE INC.
FC+CT+TR	Over the years, the company, which keeps its ao value below 15582.5 and consistently experiences a rise in its sale value, carefully manages credit and counterparty risk as part of its risk management strategy.
FC+CT+RK	During the years when the ibc value for WORLD KINECT CORP drops under 63.95, the company having the lowest sale value declares that it is not required to publicly revise forward-looking statements, except when required by legislation.
FC+AR+TR	Over the years, the company with cumulative txt figures higher than -674 and steadily increasing sale figures participates mainly in securities lending activities to generate net investment income.
FC+AR+RK	During the years when oibdp value minus lt value for UNITED PARCEL SERVICE INC is less than -47126.5, the company with the highest ppent value was granted market-based rate authority by the FERC in May 2018.

Table 4: Examples for each option type. FC refers to Fact-Checking, CT refers to Conditional Threshold, AR refers to Arithmetic, TR refers t o Trend, and RK refers to Ranking.

Option Types	Template
CT	The company with (column) values (greater than, less than) (threshold) in (Year) is company.
AR	The company where the (Year1) (column) value (plus, minus) the (Year2) (column) value equals (results) is (company).
TR	The company that showed a continuously (increasing, decreasing) trend in (column) values from (Year1) to (Year2) is (company).
RK	The company with the (highest, lowest) (column) value in (Year) is (company).
CT+TR	For all years, the company with the (column1) value (greater than, less than) (threshold) and the (column2) value continuously (increased, decreased) is (company).
CT+RK	During the years when the (column1) value for (company) is (greater than, less than) (threshold), the company with the (highest, lowest) (column2) value is (company).
AR+TR	For all years, the company with the cumulative sum of (column1) values (greater than, less than) (threshold) and the (column2) values continuously (increased, decreased) is (company).
AR+RK	During the years when (column1) value (plus, minus) (column2) value for (company1) is (greater than, less than) (threshold), the company with the (highest, lowest) (column3) value is (company2).

Table 5: Base templates of option types. FC refers to Fact-Checking, CT refers to Conditional Threshold, AR refers to Arithmetic, TR refers to Trend, and RK refers to Ranking. In the case of Option Types that include FC, a new template is generated by combining them with other Option Types and Facts, where they share a common company entity.

Difficulty	Modality Types	Option Types
Easy	Text	Fact-Checking
	Table	Conditional Threshold
		Arithmetic
	Chart	Trend
		Ranking
Medium	Text + Table	Fact-Checking + Conditional Threshold
		Fact-Checking + Arithmetic
	Text + Chart	Fact-Checking + Trend
		Fact-Checking + Ranking
		Conditional Threshold + Trend
	Table + Chart	Conditional Threshold + Ranking
		Arithmetic + Trend
		Arithmetic + Ranking
Hard	Text + Table + Chart	Fact-Checking + Conditional Threshold + Trend
		Fact-Checking + Conditional Threshold + Ranking
		Fact-Checking + Arithmetic + Trend
		Fact-Checking + Arithmetic + Ranking

Table 6: Detailed Option Types by Difficulty and Modality Types. For the Easy level, all three answer options are 1-hop options, while for the Medium level, all three options are cross-modal 2-hop options. At the Hard level, all three options consist of cross-modal 3-hop options. Even if each answer option is a 1-hop option, the overall question remains a cross-modal 3-hop QA. Specific examples can be found in Table 4

	MMQA	**FM^{3}QA(Ours)**
Percentage of cross-modal 3-hop QA(%)	0.8	100
Average Cost per question($)	0.33	0.004
# of Multi-Hop Operator Type	3	5
Adjustable QA Difficulty level	✗	✓

Table 7: Comparison between MMQA and FM^{3}QA. MMQA includes three operator types: Compose, Intersect, and Compare, while FM^{3}QA consists of five operator types: Fact-Checking, Conditional Threshold, Arithmetic, Trend, and Ranking.

You are provided with the following materials:
[Text Reports: Detailed excerpts from company reports of three companies]
{text}
[Table Data: A table containing financial data for the same three companies]
{table}
[Chart Image]
{chart}

Your Task:

Determine whether each of the following three statements is true or false based solely on the provided materials. For each statement:

1. {option1}
2. {option2}
3. {option3}

Provide a detailed reasoning process that references specific data or information from the text reports, table data, or chart images.
Do not use general knowledge or external information beyond what is provided in the materials.
If there is insufficient information to determine the truthfulness of a statement, or if the statement relies on information not present in the materials, consider it false.
Final Answer Format:

After your reasoning, provide the final answer by listing the numbers of the statements that are true.
For example: "Answer: 1 or 1,2 or 2,3".
If none of the statements are true, write: "Answer: None".

Figure 6: GPT-4o zero-shot prompt.

Your Task:
Your task is to determine the truth or falsehood of three statements given tables from financial statements, report chunks from three companies, and charts.
Provide a detailed reasoning process that references specific data or information from the text reports, table data, or chart images.
Do not use general knowledge or external information beyond what is provided in the materials.
If there is insufficient information to determine the truthfulness of a statement, or if the statement relies on information not present in the materials, consider it false.
Final Answer Format:
After your reasoning, provide the final answer by listing the numbers of the statements that are true.
For example: "Answer: 1 or 1,2 or 2,3".
If none of the statements are true, write: "Answer: None".
{Example}
You are provided with the following materials:
[Text Reports: Detailed excerpts from company reports of three companies]
{text}
[Table Data: A table containing financial data for the same three companies]
{table}
[Chart Image]
{chart}
Determine whether each of the following three statements is true or false based solely on the provided materials. For each statement:
1. {option1}
2. {option2}
3. {option3}
Provide a detailed reasoning process that references specific data or information from the text reports, table data, or chart images. Do not use general knowledge or external information beyond what is provided in the materials. If there is insufficient information to determine the truthfulness of a statement, or if the statement relies on information not present in the materials, consider it false.
Final Answer Format:
After your reasoning, provide the final answer by listing the numbers of the statements that are true.
For example: "Answer: 1 or 1,2 or 2,3".
If none of the statements are true, write: "Answer: None".

Figure 7: GPT-4o one-shot prompt.

Identifying Information

CUSIP (cusip)Ticker Symbol (tic)
CIK Number (cik)
Company Name (conm)
Fiscal Year-end Month (fyr)
State/Province (state)
Standard Industry Classification Code (sic)
North American Industry Classification Code (naics)

Balance Sheet Variables

Cash and Short-Term Investments (che)
Receivables - Total (rect)
Inventories - Total (invt)
Current Assets - Other - Total (aco)
Current Assets - Total (act)
Property, Plant and Equipment - Total (Net) (ppent)
Depreciation, Depletion and Amortization (Accumulated) (dpact)
Investment and Advances - Equity (ivaeq)
Investment and Advances - Other (ivao)
Intangible Assets - Total (intan)
Assets - Other (ao)
Assets - Total (at)
Debt in Current Liabilities - Total (dlc)
Accounts Payable - Trade (ap)
Income Taxes Payable (txp)
Current Liabilities - Other - Total (lco)
Current Liabilities - Total (lct)
Long-Term Debt - Total (dltt)
Deferred Taxes and Investment Tax Credit (txditc)
Liabilities - Other - Total (lo)
Liabilities - Total (lt)
Noncontrolling Interest (Balance Sheet) (mib)
Preferred/Preference Stock (Capital) - Total (pstk)
Common/Ordinary Equity - Total (ceq)
Stockholders Equity - Parent (seq)

Statement of Cash Flows Variables - Investing Activities

Increase in Investments (ivch)
Sale of Investments (siv)
Short-Term Investments - Change (ivstch)
Capital Expenditures (capx)
Sale of Property (sppe)
Acquisitions (aqc)
Investing Activities - Other (ivaco)
Investing Activities - Net Cash Flow (ivncf)

Income Statement Variables

Sales/Turnover (Net) (sale)
Cost of Goods Sold (cogs)
Selling, General and Administrative Expense (xsga)
Operating Income Before Depreciation (oibdp)
Operating Income After Depreciation (oiadp)
Interest and Related Expense - Total (xint)
Nonoperating Income (Expense) (nopi)
Special Items (spi)
Pretax Income (pi)
Income Taxes - Total (txt)
Income Before Extraordinary Items (ib)
Net Income Adjusted for Common/Ordinary Stock (Capital) Equivalents (niadj)
Earnings Per Share (Basic) - Excluding Extraordinary Items (epspx)
Earnings Per Share (Diluted) - Excluding Extraordinary Items (epsfx)

Statement of Cash Flows Variables - Operating Activities

Income Before Extraordinary Items (Cash Flow) (ibc)
Extraordinary Items and Discontinued Operations (Cash Flow) (xidoc)
Depreciation and Amortization (Cash Flow) (dpc)
Deferred Taxes (Cash Flow) (txdc)
Equity in Earnings - Unconsolidated Subsidiaries (esub)
Sale of Property, Plant and Equipment and Investments - Gain (Loss) (sppiv)
Funds from Operations - Other (fopo)
Accounts Receivable - Decrease (Increase) (recch)
Inventory - Decrease (Increase) (invch)
Accounts Payable and Accrued Liabilities - Increase/(Decrease) (apalch)
Income Taxes - Accrued - Increase/(Decrease) (txach)
Assets and Liabilities - Other - Net Change (aoloch)
Operating Activities - Net Cash Flow (oancf)

Statement of Cash Flows Variables - Financing Activities

Sale of Common and Preferred Stock (sstk)Excess Tax Benefit of Stock Options - Cash Flow Financing (txbcof)Purchase of Common and Preferred Stock (prstkc)Cash Dividends (Cash Flow) (dv)Long-Term Debt - Issuance (dltis)Long-Term Debt - Reduction (dltr)Current Debt - Changes (dlcch)Financing Activities - Other (fiao)Financing Activities - Net Cash Flow (fincf)

Figure 8: Description of each column in the Annual Simplified Financial Statement.