SECQUE: A Benchmark for Evaluating Real-World Financial Analysis Capabilities

Noga Ben Yoash * **Meni Brief**

Oded Ovadia **Gil Shenderovitz** **Moshik Mishaeli**

Rachel Lemberg **Eitam Sheetrit**

Microsoft Industry AI

Abstract

We introduce SECQUE, a comprehensive benchmark for evaluating large language models (LLMs) in financial analysis tasks. SECQUE comprises 565 expert-written questions covering SEC filings analysis across four key categories: comparison analysis, ratio calculation, risk assessment, and financial insight generation. To assess model performance, we develop SECQUE-Judge, an evaluation mechanism leveraging multiple LLM-based judges, which demonstrates strong alignment with human evaluations. Additionally, we provide an extensive analysis of various models' performance on our benchmark. By making SECQUE publicly available[1], we aim to facilitate further research and advancements in financial AI.

1 Introduction

Recent advances in large language models (LLMs) have demonstrated their potential across diverse domains, including law (Huang et al., 2023), medicine (Singhal et al., 2023; Wu et al., 2024), and finance (Cheng et al., 2023; Wu et al., 2023). However, as these models are increasingly adopted for specialized applications, the need for domain-specific evaluation has become more pressing. While general-purpose benchmarks assess a wide range of capabilities, they often fail to capture the nuances and challenges inherent in domain-specific tasks (Yang et al., 2024).

While domain-specific evaluation is challenging across many fields, the financial domain presents unique challenges in assessing LLM capabilities. Financial analysts routinely analyze complex datasets, extract meaningful insights from textual and numerical data, and answer high-stakes questions about companies, industries, and market trends. These tasks require models to excel in financial reasoning, numerical computation, and the synthesis of information from lengthy, multi-format documents. Yet, many existing benchmarks for financial LLMs often focus on isolated downstream tasks, such as sentiment analysis or named entity recognition, and do not adequately reflect the breadth of questions analysts face in real-world scenarios (Xie et al., 2024a; Brief et al., 2024; Islam et al., 2023).

To address this gap, we introduce SECQUE, a benchmark specifically designed to evaluate LLMs on the types of questions financial analysts pose while analyzing SEC[2] filings. SECQUE includes questions spanning four key categories: Comparison and Trend Analysis, Ratio Analysis, Risk Factors, and Analyst Insights, thus representing essential components of financial analysis. For each question, we present a ground truth answer and variations of the supporting data from the SEC filings, representing different textual pre-processing methods. The benchmark consists of 565 questions curated to challenge models' abilities to comprehend, reason, and synthesize information within the context of corporate filings.

Our benchmark offers several key advantages. First, SECQUE is designed to reflect real-world financial tasks, moving beyond basic text processing to assess reasoning over long unstructured data. Second, it emphasizes long-context questions, requiring models to extract relevant information from complex and detailed inputs, such as financial tables with varied structures. Third, SECQUE addresses limitations identified in FinanceBench (Islam et al., 2023) by introducing cross-company comparisons and high-difficulty questions.

Additionally, following (Zheng et al., 2023),

*Corresponding author: nogabenyoash@microsoft.com

[1] https://huggingface.co/datasets/nogabenyoash/SecQue

[2] SEC is the common name for the U.S. Securities and Exchange Commission

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 212–230

Table 1: Summary Statistics of the SEC filings used in SECQUE.

Statistic	Value
Unique Accessions	45
Unique Companies	29
Unique Filing Years	4
Companies with Multiple Filings	12
Earliest Filing Date	7/25/2018
Latest Filing Date	8/8/2024

LLM judges have become a central component of open-ended question evaluation, and SECQUE significantly relies on the ability to use LLMs for evaluation accordingly. The questions in SECQUE are of high complexity and therefore present difficulty for LLM judging. To address this difficulty, we present SECQUE-judge that, following (Gu et al., 2024), leverages multiple LLM judges evaluations. We perform a thorough investigation of SECQUE-judge and demonstrate its alignment with human evaluation. Using our validated SECQUE-judge, we have performed a thorough analysis of SECQUE. Finally, we conduct an ablation study to examine how different configurations, such as prompt choice, affect the results.

2 SECQUE Benchmark

The SECQUE benchmark was developed as a tool to evaluate the performance of large language models (LLMs) specializing in the financial domain in real-world financial scenarios. Our evaluation focuses on key use cases where LLMs could significantly impact the work of financial professionals in general, and financial analysts in particular. Financial analysts rely on diverse documents in their work, and we focused on the primary publicly available financial reports[3]: 10-K and 10-Q SEC filings. A 10-K is a company's annual financial report filed with the SEC, while a 10-Q is a quarterly update on its financial performance. These documents include textual and tabular data about publicly traded companies, covering sections such as risk factors, income statements, balance sheets, and cash flow statements.

Benchmark Creation: The SECQUE benchmark was created by three subject matter experts (SMEs) specializing in financial analysis. To ensure high standards, all questions and answers were iteratively refined and reviewed both by the SMEs and by two additional financial experts with expertise in LLM systems.

Benchmark Composition: The benchmark consists of 565 open-ended questions representing real-world financial analysts' questions in terms of complexity, jargon, and type. Each entry in the benchmark includes a question, supporting data (also referred to as context), and a ground truth answer. Additionally, references to the supporting data (e.g., metadata specifying accession numbers, page numbers, and relevant sections from the filings that indicate the source of the context) and a question type label are provided.

Following is an example data point from SECQUE benchmark (for full context see Appendix A).

Input:

- **Question:** `How has NVIDIA's Interest Coverage Ratio changed from 2023 to 2024?`
- **Context:** `NVIDIA Corporation and Subsidiaries...`

Ground Truth Answer:

Interest Coverage Ratio is calculated using the formula: Interest Coverage Ratio = EBIT / Interest Expenses.

For 2024:
EBIT (Operating Income) = $32,972 million
Interest Expenses = $257 million
Interest Coverage Ratio = $32,972 / $257 = **128.3**

For 2023:
EBIT (Operating Income) = $4,224 million
Interest Expenses = $262 million
Interest Coverage Ratio = $4,224 / $262 = **16.1**

Metadata:

- **Question Type:** Ratio Analysis
- **Accession Number:** 0001045810-24-000029
- **Page:** 50
- **Item:** *Item 15. Exhibit and Financial Statement Schedules*

Table 1 provides summary statistics for the underlying SEC filings. In total, the questions reference 45 SEC filings from 29 different companies, fully listed in Appendix D. The supporting data spans multiple documents and may reach significant lengths, with some entries requiring tens of thousands of tokens[4].

[3] https://sec.gov/edgar/search

[4] All token counting was done with `tiktoken.get_encoding("cl100k_base")`

Table 2: SECQUE breakdown by question type.

Question Type	Count
Comparison and Trend Analysis	220
Ratio Analysis	188
Risk Factors	85
Analyst Insights	72

Table 3: Token statistics by representation type.

Type	Mean	Std	Median	Max
HTML	5.4K	5.6K	3.9K	32.6K
Markdown	2.9K	2.9K	2.2K	16.9K

SECQUE Questions: The SMEs were instructed to write questions following three main guidelines: I) They represent real-world questions that are interesting to a financial analyst. II) The answers rely solely on the information provided in the reference supporting data; no external data is needed. III) The questions can be answered objectively, based on the provided context. The benchmark addresses four types of questions, reflecting core tasks performed by financial analysts:

(1) *Risk Questions:* Financial analysts assess potential risks impacting companies based on the "Risk Factors" section of SEC filings. This task requires text analysis skills.

(2) *Ratio Questions:* Analysts examine financial statements to understand a company's financial position, performance, and cash flow. This involves extracting data from tables, defining formulas, and performing calculations.

(3) *Comparison Questions:* Analysts identify trends and differences across multiple documents to evaluate a company's performance relative to peers or previous records.

(4) *Analyst Insights Questions:* Analysts synthesize multiple data points to generate conclusions and provide financial explanations. Insight questions require deep financial understanding.

Table 2 shows a breakdown of the benchmark's questions by subject.

References to the Supporting Data: The **context** of a question is the portion of text from an SEC filing (or multiple filings) that the SMEs have identified as relevant to answering the question. The **references to the supporting data**, indicating the pages and items to be used from each **accession number** (the unique ID of a filing), are provided in the benchmark.

We define a **chunk** of data to be the text corresponding to a single page of the filing. If multiple chapters are covered on the same page, the chunk is divided into smaller, coherent chunks. The chunks are then concatenated to form the final context of the question, with each question requiring, on average, five chunks as context.To preserve contextual clarity when concatenating chunks, each chunk may also include a brief **header** with key information (e.g., company name, filing type, and filing date). This header slightly increases the number of tokens required to execute a question.

Context: SEC filings are available for download both in XBRL and in HTML formats, and their content is composed of text and tables. We used the Markdown representation of the texts, and formatted the tables in two ways: 1) Markdown, a straightforward text-based representation that is more concise, but less expressive. 2) HTML, a structured representation using separate tags for each attribute, and styling elements removed. Table 3 provides key statistics about the number of tokens needed for HTML and Markdown representations, respectively.

Since any change in the context may impact performance on SECQUE, we provide four slightly different versions of the context for each question in the SECQUE benchmark. These versions correspond to HTML and Markdown table representations, with and without headers. Fig. 1 illustrates the available choices for text representation.

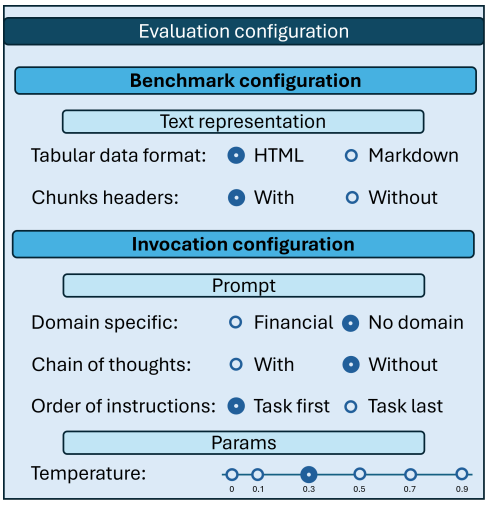

Figure 1: Configuration for executing the SECQUE benchmark. This configuration specifies the format of the text extracted from SEC filings, along with other relevant parameters. Only one radio button can be selected within each configuration category.

3 Evaluating Judge Performance

Manual evaluation of the entire benchmark is impractical, therefore, we have implemented SECQUE-judge, an automated comparison for various model outputs with the SECQUE ground truth answers (denoted as $\langle \tilde{y}, y \rangle$, respectively). In this section we describe our SECQUE-judge implementation and verify alignment with human evaluation.

3.1 SECQUE-judge Implementation

For SECQUE evaluation, our primary goal is to ensure that it properly distinguishes between fully correct answers (i.e., answers acceptable for a financial analyst) and those that are partially correct or incorrect. To this end, we use *Single-judge*, employing a scoring system of $\{0, 1, 2\}$, representing incorrect, partially correct, and correct answers, respectively. Single-judge's implementation follows the judging prompt presented in (Brief et al., 2024), which similarly handles free-text comparisons categorized into three classes. We use GPT-4o (OpenAI, 2024) as the underlying judging model.

Since an LLM judge can be inconsistent due to its stochastic nature, we utilize a 'panel of judges', following LLM-as-a-judge best practices outlined in (Gu et al., 2024). We form our final SECQUE-judge by aggregating several Single-judge scores: for each $\langle \tilde{y}, y \rangle$ pair, we invoke Single-judge five times (using the exact same prompt and parameters). The summed score of these five individual evaluations is denoted by S. SECQUE-judge maps S to a final categorical score with same $\{0, 1, 2\}$ scoring system using two fixed thresholds, U_T (upper threshold) and L_T (lower threshold), as defined in Eq. (1). We aim to compute the optimal thresholds U_T and L_T for our SECQUE evaluation.

$$\text{score} := \begin{cases} 2, & \text{if } S \geq U_T, \\ 1, & \text{if } U_T > S \geq L_T, \\ 0, & \text{if } S < L_T, \end{cases} \quad (1)$$

3.2 Human Evaluation Experiment Setup

We conducted an experiment to assess the alignment between our SECQUE-judge and expert human evaluation. First, we ran our benchmark and generated answers using GPT-4o and Llama-3.3-70B-Instruct (Dubey et al., 2024). Due to the high cost of human evaluation, we manually selected a subset of 62 questions from all four question categories that were scored differently by several automated judges (described in Section 3.3). Since each question was answered by two LLM models, this resulted in 124 generated answers for evaluation, 62 from GPT-4o and 62 from Llama-3.3-70B-Instruct.

Next, we presented the 124 answers to financial experts and asked them to independently compare each generated $\tilde{y}$ to its corresponding y using the same $\{0, 1, 2\}$ scale as described earlier. This setup allows us to find a lower bound on the alignment between SECQUE-judge and human evaluation.

For most questions, all human evaluators assigned the same score. In cases where the evaluation was a mix of 1 and 2, we set the final human-score to 2, as such an answer could be deemed acceptable for a financial analyst. Similarly, when scores of 0 and 1 were assigned, the final human-score was set to 0, as the answer was considered mostly incorrect. In the only four cases where evaluators disagreed entirely (with the full range of scores assigned), we set the final human-score to 1.

Since we are primarily interested in verifying that SECQUE-judge properly distinguishes fully correct answers from others, we use the following $F_1(2)$ metric as our optimization objective:

$$F_1(2) := 2 \cdot \frac{\text{precision}(2) \cdot \text{recall}(2)}{\text{precision}(2) + \text{recall}(2)}, \quad (2)$$

i.e., the standard multi-class F_1, precision, and recall scores, when 2 is the target class.

3.3 Analyzing SECQUE-judge

We begin by evaluating the stability of Single-judge scoring on the answer set. In all cases, the five Single-judge scores differed by at most 1, meaning that we did not observe both scores of 0 and 2 for the same $\langle \tilde{y}, y \rangle$ pair. In 85.5% of cases, the five Single-judge scores were unanimous. Fig. 2 presents a histogram of S, the summed Single-judge scores for the 62 questions, showing that the most common sums are 0, 5, and 10, representing unanimous scores of 0, 1, and 2, respectively.

We then used human-scores and Single-judge summed scores S to calculate the optimal U_T and L_T (defined in Eq. (1)) maximizing our objective function $F_1(2)$ presented in Eq. (2). We finalized $U_T = 6$ and $L_T = 4$ to be the threshold used in SECQUE-judge, which resulted in a maximal $F_1(2) = 0.85$ (the full confusion matrix is presented in Appendix C). Thus, Eq. (3) represents our final SECQUE-judge. It is interesting to note that $U_T = 6$ implies that at least one Single-judge

Table 4: Comparison of LLM-based judges, assessing their alignment with human judgment across multiple alignment metrics. A judge is defined both by its methodology and by the LLM used to perform the judging. The best scores for each alignment metric are indicated by underlining.

Judge Methodology	Underlying Model	Alignment Metrics F1(2)	precision(2)	recall(2)	accuracy
Single-judge	GPT-4o	0.82	0.9	0.75	0.71
Majority vote	GPT-4o	0.8	0.9	0.73	0.69
SECQUE-judge	**GPT-4o**	<u>0.85</u>	0.905	0.8	<u>0.75</u>
SECQUE-judge	Llama-3.3-70B-Instruct	0.83	0.8	<u>0.86</u>	0.68
SECQUE-judge	GPT-4o-mini	0.62	<u>0.93</u>	0.465	0.515

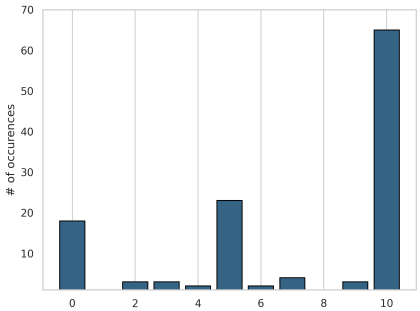

Figure 2: Histogram of S, the sum of five Single-judge scores, for all 124 answers.

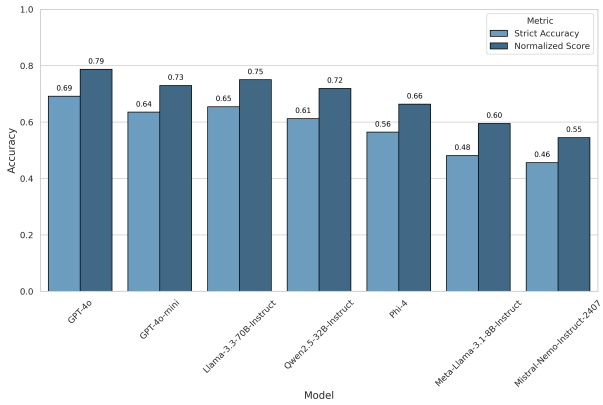

Figure 3: The performance of each model on the benchmark. Both Strict Accuracy and Normalized Accuracy are shown.

assigned a score of 2 to the answer. Similarly, $L_T = 4$ implies that at least one Single-judge assigned a score of 0.

$$\text{score} = \begin{cases} 2, & \text{if } S \geq 6, \\ 1, & \text{if } 4 \leq S < 6, \\ 0, & \text{if } S < 4. \end{cases} \quad (3)$$

Further analysis of SECQUE-judge is presented in Table 4. We first observe that precision$(2) = 0.905$ and accuracy $= 0.75$. We conclude that SECQUE-judge excels in identifying fully correct answers, while its ability to distinguish between partially correct and incorrect answers is less optimal.

SECQUE-judge also outperforms other evaluation methods in terms of alignment. Table 4 demonstrates that employing SECQUE-judge, a panel of judges, instead of Single-judge, improves performance across all metrics by up to 4%. Majority vote utilizes the same summed score S, but results in lower alignment with human evaluation. This further implies that one Single-judge score of 2 or 0 out of five Single-judge scores is enough to award a final score of 2 and 0, respectively.

Additionally, we changed the underlying judging model, both with Llama-3.3-70B-Instruct and GPT-4o-mini (OpenAI, 2024)). While the first performs almost like GPT-4o, for the second we observe a significant decrease in the alignment between the judge and human evaluation. We also provide a breakdown by which model generated the answer is provided in Appendix C, to mitigate possible concerns around self-enhancement bias (Zheng et al., 2023).

4 Evaluation and Results

4.1 Setup

We evaluated the performance of seven models on SECQUE, representing diverse model sizes and providers, to assess their ability to answer complex financial questions effectively. The models we chose are GPT-4o and GPT-4o-mini, Meta-Llama-3.3-70B-Instruct and Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-32B-Instruct (Qwen, 2024), Mistral-Nemo-Instruct-2407(12B) (Mistral, 2024), and Phi-4(14B) (Abdin et al., 2024)[5].

All answers were scored using our SECQUE-judge. Each response was given a score according

[5]Phi-4 has a limited context length of just 16K, resulting in lower performance, as longer questions remained unanswered.

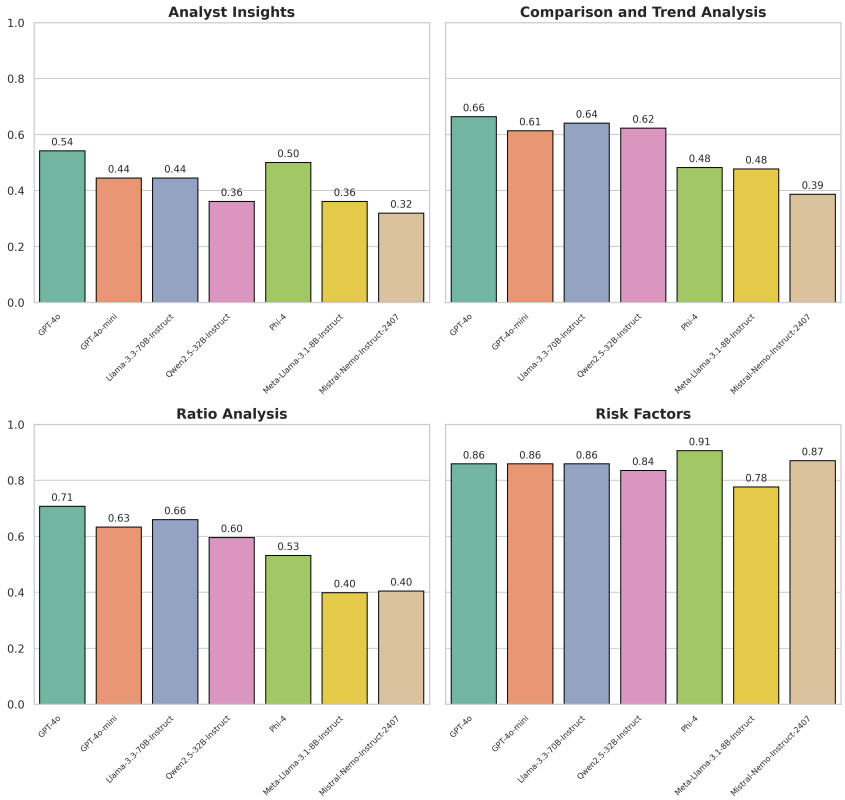

Figure 4: Model performance across different question types. Each subplot represents one question type, comparing the Strict Accuracy of all models.

	Baseline	Financial	Baseline CoT	Financial CoT	Flipped	Avg Tokens by Model
GPT-4o	**0.69**/0.79	0.62/0.71	0.67/0.76	0.63/0.73	0.68/0.78	319.84
GPT-4o-mini	0.64/0.73	0.38/0.47	0.60/0.72	0.56/0.65	0.62/0.73	289.76
Llama-3.3-70B-Instruct	0.65/0.75	0.60/0.71	0.63/0.74	0.60/0.72	0.62/0.74	341.63
Qwen2.5-32B-Instruct	0.61/0.72	0.49/0.58	0.60/0.71	0.55/0.67	0.65/0.75	331.34
Phi-4	0.56/0.66	0.55/0.64	0.57/0.67	0.56/0.66	0.57/0.67	294.33
Meta-Llama-3.1-8B-Instruct	0.48/0.60	0.41/0.54	0.44/0.56	0.40/0.53	0.47/0.59	338.38
Mistral-Nemo-Instruct-2407	0.46/0.55	0.32/0.42	0.45/0.56	0.44/0.55	0.44/0.54	231.52
Avg Tokens by Prompt	283.04	151.97	437.38	334.71	317.57	304.93

Table 5: Performance metrics across prompt ablations. In each column, the left score indicates Strict Accuracy, the right Normalized Accuracy. The average number of output tokens used for each model and prompt type is included. The best score per model is underlined, and best overall is in bold

to Eq. (3), which was then aggregated into two scores:

- **Strict Accuracy**: $\frac{1}{2n}\sum_i 2\mathbf{I}_{\{\text{score}=2\}}$ (2 points if score = 2 else 0).
- **Normalized Accuracy**: $\frac{1}{2n}\sum_i \text{score}$ (use score directly).

Both scores were divided by 2 to maintain a $[0, 1]$ scale.

To mitigate any issues arising from the sensitivity of LLMs to input perturbations, particular attention was given to standardizing data representations and prompts. Fig. 1 illustrates the possible configurations for an experiment using the SECQUE benchmark and identifies the 'baseline' configuration (simple prompt, `temperature=0.3`, and HTML tables with headers) that results in the highest overall performance across models. In the rest of this section we analyze the performance of the described models using the 'baseline' configuration, except for the ablation studies where we evaluate the effect of text representation, prompt and temperature configurations, both on quality and on the number of tokens produced.

4.2 Overall Performance

The performance of each model on the benchmark is shown in Fig. 3. GPT-4o leads with 0.69 and

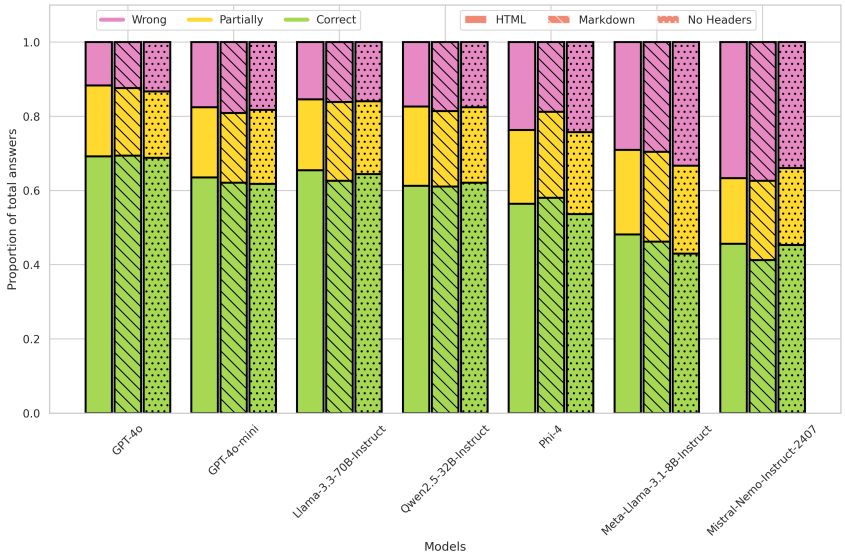

Figure 5: A comparison of all models' performance for each data representation configuration (HTML, Markdown, HTML with no headers), as well as a breakdown of scores achieved by each model. Note that the leftmost column for each model is equivalent to the baseline shown in Fig. 3

0.79 in Strict and Normalized accuracy, respectively. GPT-4o-mini and Llama-3.3-70B-Instruct have very similar performance, both slightly under GPT-4o and slightly above Qwen2.5-32B-Instruct. The smaller models perform significantly worse with Mistral-Nemo-Instruct-2407 being the furthest behind. It is interesting to note that while the absolute difference between Strict and Normalized accuracies remains similar across all models, the ratio of these accuracies is significantly higher for smaller models. This trend is more clearly illustrated in Fig. 5.

4.3 Performance by Question Type

The various models' Strict Accuracy scores across the four SECQUE question categories are shown in Fig. 4. Results highlight significant variability across categories:

Risk Factors: Phi-4 performed best, with almost all the other models achieving similar scores. All models achieved high scores, implying that answering such questions should be a minimum requirement for any financial model.

Ratio Analysis: This category proved more challenging, with GPT-4o achieving the highest score. The results indicate both correct usage of formulas and superior mathematical reasoning abilities.

Comparison and Trend Analysis: The results for this category were very similar to Ratio Analysis. Smaller models exhibited difficulty reasoning over data points from long contexts, while the rest of the models had roughly equivalent performance.

Analyst Insights: These questions had the lowest scores across almost all models, with GPT-4o significantly ahead, followed by Phi-4. These questions are more difficult in nature due to combining numerical reasoning and financial insights, but also involve slightly more nuanced answers, and therefore the evaluation of this category may be less reliable than the other categories.

4.4 Ablation Study

Text Representation: The choice of text representation i.e., HTML, Markdown, and removing headers, had a small impact on overall performance. Fig. 5 shows the performance of the models across two important dimensions, both comparing the representation format, and also showing a breakdown of the scores for each model. The results indicate Markdown tables were slightly harder for smaller models to interpret, indicating a trade-off between using fewer tokens and a more explicit representation format. The exception is Phi-4, gaining a boost from the token reduction due to its limited context length. The inclusion of headers is not conclusively helpful, but in most cases appears to be beneficial.

Prompt Variations: Altering the prompt had the most significant impact of the various ablations. Switching from the baseline prompt to a more financial and targeted one proved to be very detrimental to performance, although better from a token usage perspective. Interestingly, while including chain-

of-thought (CoT) reasoning in the baseline prompt resulted in a slight decrease in performance, incorporating CoT in the financial prompt led to a modest improvement. These findings are surprising since generally providing clearer instructions, as well as explicitly requesting the use of CoT have been shown to improve results in various reasoning tasks (Wei et al., 2023). Changing the order within the prompt (context followed by question vs. question followed by context) had minimal impact, which contrasts with the findings of (Islam et al., 2023). This discrepancy can be attributed to our use of newer and more advanced models. All prompts can be found in Appendix B.

Temperature Settings: Temperature adjustments $\{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ were evaluated only for GPT-4o. The change in temperature had almost no impact, with less than 2% fluctuations between values, thus we cannot conclude that the choice of temperature matters for evaluation.

5 Related Work

Recent advances in large language models (LLMs) have spurred considerable research in domain-specific benchmarks and evaluation frameworks, particularly in finance. In this section, we briefly review work on financial benchmarks and the use of LLMs for evaluation.

Financial Benchmarks and Datasets A variety of benchmarks have been introduced to assess LLM performance on financial tasks. Comprehensive evaluation frameworks such as FinBen (Xie et al., 2024b), PIXIU (Xie et al., 2024a), and BBT-Fin (Lu et al., 2023) aggregate diverse tasks to measure general financial skills. Other datasets target specialized skills: FinEval (Zhang et al., 2023) focuses on textbook-based financial knowledge, SuperCLUE-Fin (Xu et al., 2024) decomposes real-world financial tasks into fine-grained subtasks, and FinDABench (Liu et al., 2024) emphasizes financial analysis and reasoning. In parallel, several financial QA datasets have been proposed. Early efforts include FiQA (Maia et al., 2018) for sentiment analysis and opinionated QA, while FinQA (Chen et al., 2021) and its conversational extension ConvFinQA (Chen et al., 2022) offer more realistic, multi-turn interactions. Datasets such as TAT-QA (Zhu et al., 2021) incorporate numerical reasoning over tabular and textual data from financial reports. Despite these efforts, many of the existing benchmarks do not fully capture the retrieval, analysis and reasoning challenges inherent to day-to-day financial analysis (Brief et al., 2024; Islam et al., 2023), which are necessary for real-world financial work.

Evaluation Paradigms: LLM-as-a-Judge Traditional benchmark evaluation has evolved with the emergence of LLMs. Beyond standard multiple-choice or completion tasks where easy evaluation is possible, recent approaches leverage LLMs (notably GPT-4 (Achiam et al., 2023)) as automated judges for assessing generation quality. For example, Li et al. (Li et al., 2023) and Zheng et al. (Zheng et al., 2023) have demonstrated the effectiveness of using LLMs to score answers in open-ended question setups, while (Gu et al., 2024) employed majority voting from multiple judges. (Gu et al., 2024) and others have conducted extensive studies around the alignment of LLM evaluators with human annotators, yet a single optimal setup has not been identified, prompting the need for further case-by-case optimization.

6 Conclusions

We have presented SECQUE, a comprehensive benchmark for evaluating LLMs in financial analysis tasks. Our results demonstrate that while leading models show promising capabilities in financial analysis, significant challenges remain, particularly in complex reasoning tasks and analyst insights generation. The benchmark reveals important differences in model performance across question types and highlights the critical role of configurations in evaluation results. These findings provide valuable guidance for future development of financial LLMs and evaluation frameworks.

7 Limitations

Limitations of our work include potential biases in the LLM-based evaluation system, the need for broader coverage of financial document types. Another key limitation is that there could be more than one correct way to calculate some of the analysis questions. This is an inherent part of the domain, as there are potentially more than one way for analysts to interpret financial information.

Future work should address these limitations by allowing for multiple correct ways to answer questions and expanding the benchmark to cover additional financial tasks and document types.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Meni Brief, Oded Ovadia, Gil Shenderovitz, Noga Ben Yoash, Rachel Lemberg, and Eitam Sheetrit. 2024. Mixing it up: The cocktail effect of multi-task fine-tuning on llm performance–a case study in finance. *arXiv preprint arXiv:2410.01109*.

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

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. *arXiv preprint arXiv:2210.03849*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yongxin Huang, Kexin Wang, Sourav Dutta, Raj Nath Patel, Goran Glavaš, and Iryna Gurevych. 2023. Adasent: Efficient domain-adapted sentence embeddings for few-shot classification. *arXiv preprint arXiv:2311.00408*.

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

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, Qingquan Wu, and Chong Yang. 2024. Findabench: Benchmarking financial data analysis ability of large language models. *Preprint*, arXiv:2401.02982.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

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

Mistral. 2024. Mistral nemo. Accessed: 2024-11-21.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

OpenAI. 2024. Hello, gpt-4. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-09-23.

Qwen. 2024. Qwen2.5: A party of foundation models.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2024a. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, et al. 2024b. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Liang Xu, Lei Zhu, Yaotong Wu, and Hang Xue. 2024. Superclue-fin: Graded fine-grained analysis of chinese llms on diverse financial tasks and applications. *arXiv preprint arXiv:2404.19063*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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

A Question Examples

Ratio Analysis:

Input:

- **Question:** `How has NVIDIA's Interest Coverage Ratio changed from 2023 to 2024?`
- **Context:**

NVIDIA CORP 10-K form for the fiscal year ended 2024-01-28, page 50:

NVIDIA Corporation and Subsidiaries
Consolidated Statements of Income
(In millions, except per share data)

	Year Ended		
	Jan 28, 2024	Jan 29, 2023	Jan 30, 2022
Revenue	$ 60,922	$ 26,974	$ 26,914
Cost of revenue	16,621	11,618	9,439
Gross profit	44,301	15,356	17,475
Operating expenses			
Research and development	8,675	7,339	5,268
Sales, general and administrative	2,654	2,440	2,166
Acquisition termination cost	-	1,353	-
Total operating expenses	11,329	11,132	7,434
Operating income	32,972	4,224	10,041
Interest income	866	267	29
Interest expense	(257)	(262)	(236)
Other, net	237	(48)	107
Other income (expense), net	846	(43)	(100)
Income before income tax	33,818	4,181	9,941
Income tax expense (benefit)	4,058	(187)	189
Net income	$ 29,760	$ 4,368	$ 9,752
Net income per share:			
Basic	$ 12.05	$ 1.76	$ 3.91
Diluted	$ 11.93	$ 1.74	$ 3.85
Weighted average shares used in per share computation:			
Basic	2,469	2,487	2,496
Diluted	2,494	2,507	2,535

See accompanying notes to the consolidated financial statements.

Ground Truth Answer:

Interest Coverage Ratio is calculated using the formula: Interest Coverage Ratio = EBIT / Interest Expenses.

For 2024:
EBIT (Operating Income) = $32,972 million
Interest Expenses = $257 million
Interest Coverage Ratio = $32,972 / $257 = **128.3**

For 2023:
EBIT (Operating Income) = $4,224 million
Interest Expenses = $262 million
Interest Coverage Ratio = $4,224 / $262 = **16.1**

Metadata:

- **Question Type:** Ratio Analysis
- **Accession Number:** 0001045810-24-000029
- **Page:** 50
- **Item:** *Item 15. Exhibit and Financial Statement Schedules*

Risk Factors:

Input:

- **Question:** `What are the potential financial and operational impacts of climate change on The Coca-Cola Company?`
- **Context:**

COCA COLA CO 10-K form for the fiscal year ended 2023-12-31, page 25:
and oceans, as well as inefficient use of resources when packaging materials are not included in a circular economy. We and our bottling partners sell certain of our beverage products in plastic bottles and use other packaging materials that, while largely recyclable, may not be regularly recovered and recycled due to lack of collection and recycling infrastructure. If we and our bottling partners do not, or are perceived not to, act responsibly to address plastic materials recoverability and recycling concerns and associated waste management issues, our corporate image and brand reputation could be damaged, which may cause some consumers to reduce or discontinue consumption of some of our beverage products. In addition, from time to time we establish and publicly announce goals and targets to reduce the Coca-Cola system's impact on the environment by, for example, increasing our use of recycled content in our packaging materials; increasing our use of packaging materials that are made in part of plant-based renewable materials; expanding our use of reusable packaging (including refillable or returnable glass and plastic bottles, as well as dispensed and fountain delivery models where consumers use refillable containers for our beverages); participating in programs and initiatives to reclaim or recover bottles and other packaging materials that are already in the environment; and taking other actions and participating in other programs and initiatives organized or sponsored by nongovernmental organizations and other groups. If we and our bottling partners fail to achieve or improperly report on our progress toward achieving our announced environmental goals and targets, the resulting negative publicity could adversely affect consumer preference for our products. In addition, in response to environmental concerns, governmental entities in the United States and in many other jurisdictions around the world have adopted, or are considering adopting, regulations and policies designed to mandate or encourage plastic packaging waste reduction and an increase in recycling rates andor recycled content minimums, or, in some cases, restrict or even prohibit the use of certain plastic containers or packaging materials. These regulations and policies, whatever their scope or form, could increase the cost of our beverage products or otherwise put the Company at a competitive disadvantage. In addition, our increased focus on reducing plastic containers and other packaging materials waste has in the past and may continue to require us or our bottling partners to incur additional expenses and to increase our capital expenditures. A reduction in consumer demand for our products andor an increase in costs and expenditures relating to production and distribution as a result of these environmental concerns regarding plastic bottles and other packaging materials could have an adverse effect on our business and results of operations.
Water scarcity and poor quality could negatively impact the Coca-Cola system's costs and capacity. Water is a main ingredient in substantially all of our products, is vital to the production of the agricultural ingredients on which our business relies and is needed in our manufacturing process. It also is critical to the prosperity of the communities we serve and the ecosystems in which we operate. Water is a limited resource in many parts of the world, facing unprecedented challenges from overexploitation, increasing demand for food and other consumer and industrial products whose manufacturing processes require water, increasing pollution and emerging awareness of potential contaminants, poor management, lack of physical or financial access to water, sociopolitical tensions due to lack of public infrastructure in certain areas of the world and the effects of climate change. As the demand for water continues to increase around the world, and as water becomes scarcer and the quality of available water deteriorates, the Coca-Cola system may incur higher costs or face capacity constraints and the possibility of reputational damage, which could adversely affect our profitability.
Increased demand for food products, decreased agricultural productivity and increased regulation of ingredient sourcing due diligence may negatively affect our business.
As part of the manufacture of our beverage products, we and our bottling partners use a number of key ingredients that are derived from agricultural commodities such as sugarcane, corn, sugar beets, citrus, coffee and tea. Increased demand for food products; decreased agricultural productivity in certain regions of the world as a result of changing weather patterns; loss of biodiversity; increased agricultural regulations, including regulation of ingredient sourcing due diligence; and other factors have in the past, and may in the future, limit the availability andor increase the cost of such agricultural commodities and could impact the food security of communities around the world... Climate change and legal or regulatory responses thereto may have a long-term adverse impact on our business and results of operations.
There is increasing concern that a gradual increase in global average temperatures due to increased concentration of carbon dioxide and other greenhouse gases in the atmosphere is causing significant changes in weather patterns around the globe and an increase in the frequency and severity of natural disasters. Decreased agricultural productivity in certain regions of the world as a result of changing weather patterns may limit the availability or increase the cost of key agricultural commodities, such as sugarcane, corn, sugar beets, citrus, coffee and tea, which are important ingredients for our products, and could impact the food security of communities around the world. Climate change may also exacerbate extreme weather, resulting in water scarcity or flooding, and cause a further deterioration of water quality in affected regions, which could limit water availability for the Coca-Cola system's bottling operations. Increased frequency or duration of extreme weather conditions could also impair 25

COCA COLA CO 10-K form for the fiscal year ended 2023-12-31, page 26:
production capabilities, disrupt our supply chain or impact demand for our products. Increasing concern over climate change also may result in additional legal or regulatory requirements designed to reduce or mitigate the effects of carbon dioxide and other greenhouse gas emissions on the environment, andor may result in increased disclosure obligations. Increased energy or compliance costs and expenses due to increased legal or regulatory requirements may cause disruptions in, or an increase in the costs associated with, the manufacturing and distribution of our beverage products. The physical effects and transition costs of climate change and legal, regulatory or market initiatives to address climate change could have a long-term adverse impact on our business and results of operations. In addition, from time to time we establish and publicly announce goals and targets to reduce the Coca-Cola system's carbon footprint by increasing our use of recycled packaging materials, expanding our renewable energy usage, and participating in environmental and sustainability programs and initiatives organized or sponsored by nongovernmental organizations and other groups to reduce greenhouse gas emissions industrywide. If we and our bottling partners fail to achieve or improperly report on our progress toward achieving our carbon footprint reduction goals and targets, the resulting negative publicity could adversely affect consumer preference for our beverage products.
Adverse weather conditions could reduce the demand for our products.
The sales of our products are influenced to some extent by weather conditions in the markets in which we operate. Unusually cold or rainy weather during the summer months may have a temporary effect on the demand for our products and contribute to lower sales, which could have an adverse effect on our results of operations for such periods.

Ground Truth Answer:

Climate change poses several financial and operational risks to The Coca-Cola Company. Changes in weather patterns and increased frequency of extreme weather events can disrupt production and supply chains. For example, severe droughts or floods can impact water availability and quality, affecting manufacturing processes.

Metadata:

- **Question Type:** Risk Factors
- **Accession Number:** 0000021344-24-000009
- **Page:** 25, 26
- **Item:** *ITEM 1A. RISK FACTORS*

Comparison and Trend Analysis:

Input:

- **Question:** `Compare the deposit balances for Goldman Sachs and Bank of New York Mellon as of June 30, 2024.`
- **Context:**

  GOLDMAN SACHS GROUP INC 10-Q form for quarterly period ended 2024-06-30, page 2:

  THE GOLDMAN SACHS GROUP, INC. AND SUBSIDIARIES
  Consolidated Balance Sheets
  (Unaudited)

	As of	
$ in millions	June 2024	December 2023
Assets		
Cash and cash equivalents	$ 206,326	$ 241,577
Collateralized agreements:		
Securities purchased under agreements to resell (includes $198,360 and $223,543 at fair value)	198,626	223,805
Securities borrowed (includes $45,819 and $44,930 at fair value)	204,621	199,420
Customer and other receivables (includes $23 and $23 at fair value)	142,000	132,495
Trading assets (at fair value and includes $117,586 and $110,567 pledged as collateral)	521,981	477,510
Investments (includes $86,855 and $75,767 at fair value)	160,924	146,839
Loans (net of allowance of $4,808 and $5,050, and includes $6,035 and $6,506 at fair value)	184,127	183,358
Other assets (includes $243 and $366 at fair value)	34,708	36,590
Total assets	$ 1,653,313	$ 1,641,594
Liabilities and shareholders' equity		
Deposits (includes $32,042 and $29,460 at fair value)	$ 433,105	$ 428,417
Collateralized financings:		
Securities sold under agreements to repurchase (at fair value)	238,139	249,887
Securities loaned (includes $10,775 and $8,934 at fair value)	63,935	60,483
Other secured financings (includes $22,868 and $12,554 at fair value)	23,123	13,194
Customer and other payables	242,986	230,728
Trading liabilities (at fair value)	199,660	200,355
Unsecured short-term borrowings (includes $49,579 and $46,127 at fair value)	76,769	75,945
Unsecured long-term borrowings (includes $88,361 and $86,410 at fair value)	234,632	241,877
Other liabilities (includes $142 and $266 at fair value)	21,501	23,803
Total liabilities	1,533,850	1,524,689
Commitments, contingencies and guarantees		
Shareholders' equity		
Preferred stock; aggregate liquidation preference of $12,753 and $11,203	12,753	11,203
Common stock; 927,414,906 and 922,895,030 shares issued, and 316,162,882 and 323,376,354 shares outstanding	9	9
Share-based awards	5,058	5,121
Nonvoting common stock; no shares issued and outstanding	–	–
Additional paid-in capital	61,350	60,247
Retained earnings	148,652	143,688
Accumulated other comprehensive loss	(2,900)	(2,918)
Stock held in treasury, at cost; 611,252,026 and 599,518,678 shares	(105,459)	(100,445)
Total shareholders' equity	119,463	116,905
Total liabilities and shareholders' equity	$ 1,653,313	$ 1,641,594

  See accompanying notes to the consolidated financial statements.

Bank of New York Mellon Corp 10-Q form for quarterly period ended 2024-06-30, page 52:

The Bank of New York Mellon Corporation (and its subsidiaries)
Consolidated Balance Sheet (unaudited)

(dollars in millions, except per share amounts)	June 30, 2024	Dec. 31, 2023
Assets		
Cash and due from banks, net of allowance for credit losses of $27 and $18	$ 5,311	$ 4,922
Interest-bearing deposits with the Federal Reserve and other central banks	116,139	111,550
Interest-bearing deposits with banks, net of allowance for credit losses of $1 and $2 (includes restricted of $2,026 and $3,420)	11,488	12,139
Federal funds sold and securities purchased under resale agreements	29,723	28,900
Securities:		
Held-to-maturity, at amortized cost, net of allowance for credit losses of $1 and $1 (fair value of $41,287 and $44,711)	46,429	49,578
Available-for-sale, at fair value (amortized cost of $94,566 and $80,678, net of allowance for credit losses of $5 and less than $1)	90,421	76,817
Total securities	136,850	126,395
Trading assets	9,609	10,058
Loans	70,642	66,879
Allowance for credit losses	(286)	(303)
Net loans	70,356	66,576
Premises and equipment	3,267	3,163
Accrued interest receivable	1,253	1,150
Goodwill	16,217	16,261
Intangible assets	2,826	2,854
Other assets, net of allowance for credit losses on accounts receivable of $3 and $3 (includes $1,577 and $1,261, at fair value)	25,500	25,909
Total assets	$ 428,539	$ 409,877
Liabilities		
Deposits:		
Noninterest-bearing deposits (principally U.S. offices)	$ 58,029	$ 58,274
Interest-bearing deposits in U.S. offices	149,115	132,616
Interest-bearing deposits in non-U.S. offices	97,167	92,779
Total deposits	304,311	283,669
Federal funds purchased and securities sold under repurchase agreements	15,701	14,507
Trading liabilities	3,372	6,226
Payables to customers and broker-dealers	17,569	18,395
Commercial paper	301	-
Other borrowed funds	280	479
Accrued taxes and other expenses	4,729	5,411
Other liabilities (including allowance for credit losses on lending-related commitments of $73 and $87, also includes $63 and $195, at fair value)	10,208	9,028
Long-term debt	30,947	31,257
Total liabilities	387,418	368,972
Temporary equity		
Redeemable noncontrolling interests	92	85
Permanent equity		
Preferred stock – par value $0.01 per share; authorized 100,000,000 shares; issued 43,826 and 43,826 shares	4,343	4,343
Common stock – par value $0.01 per share; authorized 3,500,000,000 shares; issued 1,409,173,568 and 1,402,429,447 shares	14	14
Additional paid-in capital	29,139	28,908
Retained earnings	40,999	39,549
Accumulated other comprehensive loss, net of tax	(4,900)	(4,893)
Less: Treasury stock of 671,216,069 and 643,085,355 common shares, at cost	(28,752)	(27,151)
Total The Bank of New York Mellon Corporation shareholders' equity	40,843	40,770
Nonredeemable noncontrolling interests of consolidated investment management funds	186	50
Total permanent equity	41,029	40,820
Total liabilities, temporary equity and permanent equity	$ 428,539	$ 409,877

See accompanying unaudited Notes to Consolidated Financial Statements

Ground Truth Answer:

As of June 30, 2024, Goldman Sachs' deposits were $433,105 million, up from $428,417 million as of December 31, 2023, marking a 1.1% increase. Bank of New York Mellon's total deposits were $304,311 million as of June 30, 2024, up from $283,669 million as of December 31, 2023, marking a 7.3% increase.

Metadata:

- **Question Type:** Comparison and Trend Analysis
- **Accession Number:** 0000886982-24-000022; 0001390777-24-000105
- **Page:** 2; 52
- **Item:** *Item 1. Financial Statements (Unaudited)*; *Item 1. Financial Statements:*

Analyst Insights:

Input:

- **Question:** `How does DFS Debt-to-Equity Ratio for 2023 reflect on the company's financial stability?`
- **Context:**

Discover Financial Services 10-K form for the fiscal year ended 2023-12-31, page 85:

DISCOVER FINANCIAL SERVICES
Consolidated Statements of Financial Condition
(dollars in millions, except for share amounts)

	December 31 2023	2022
Assets		
Cash and cash equivalents	$ 11,685	$ 8,856
Restricted cash	43	41
Investment securities (includes available-for-sale securities of $13,402 and $11,987 reported at fair value with associated amortized cost of $13,451 and $12,167 at December 31, 2023 and 2022, respectively)	13,655	12,208
Loan receivables		
Loan receivables	128,409	112,120
Allowance for credit losses	(9,283)	(7,374)
Net loan receivables	119,126	104,746
Premises and equipment, net	1,091	1,003
Goodwill	255	255
Other assets	5,667	4,597
Total assets	151,522	131,706
Liabilities and Stockholders' Equity		
Liabilities		
Deposits		
Interest-bearing deposit accounts	107,493	90,151
Non-interest-bearing deposit accounts	1,438	1,485
Total deposits	108,931	91,636
Short-term borrowings	750	-
Long-term borrowings	20,581	20,108
Accrued expenses and other liabilities	6,432	5,618
Total liabilities	136,694	117,362
Commitments, contingencies and guarantees (Notes 15, 18 and 19)		
Stockholders' Equity		
Common stock, par value $0.01 per share; 2,000,000,000 shares authorized; 570,837,720 and 569,689,007 shares issued at December 31, 2023 and 2022, respectively	6	6
Preferred stock, par value $0.01 per share; 200,000,000 shares authorized; 10,700 shares issued and outstanding at December 31, 2023 and 2022, respectively	1,056	1,056
Additional paid-in capital	4,553	4,468
Retained earnings	30,448	28,207
Accumulated other comprehensive loss	(225)	(339)
Treasury stock, at cost; 320,734,860 and 302,305,216 shares at December 31, 2023 and 2022, respectively	(21,010)	(19,054)
Total stockholders' equity	14,828	14,344
Total liabilities and stockholders' equity	151,522	131,706

The table below presents the carrying amounts of certain assets and liabilities of Discover Financial Services' consolidated variable interest entities (ẄIEsẅ), which are included in the consolidated statements of financial condition above. The assets in the table below include those assets that can only be used to settle obligations of the consolidated VIEs. The liabilities in the table below include third-party liabilities of consolidated VIEs only and exclude intercompany balances that eliminate in consolidation. The liabilities also exclude amounts for which creditors have recourse to the general credit of Discover Financial Services.

	December 31 2023	2022
Assets		
Restricted cash	$ 43	$ 41
Loan receivables	$ 30,590	$ 25,937
Allowance for credit losses allocated to securitized loan receivables	$ (1,347)	$ (1,152)
Other assets	$ 3	$ 3
Liabilities		
Short- and long-term borrowings	$ 11,743	$ 10,259
Accrued expenses and other liabilities	$ 19	$ 14

See Notes to the Consolidated Financial Statements

Ground Truth Answer:

Increase in Leverage: The ratio increased from 8.2 in 2022 to 9.2 in 2023, indicating higher reliance on debt relative to equity.
Financial Risk: The higher ratio suggests greater financial risk due to increased debt obligations.
Impact on Stability: Greater leverage could affect financial stability, especially in adverse economic conditions or with rising interest rates.

Metadata:

- **Question Type:** Analyst Insights
- **Accession Number:** 0001393612-24-000010
- **Page:** 85
- **Item:** *Item 8. Financial Statements and Supplementary Data*

B Instruction Prompts

The various prompts from Table 5 are included here.

Baseline Prompt

You are given a financial question and a financial document. Your task is to answer the question based on the document.

Input:

- **Document:** `{document}`
- **Question:** `{question}`

Output:

- *A response answering the question based on the provided document.*

Financial Prompt

You are given a financial text extracted from 10-K or 10-Q files and a question written by domain experts. Your task is to answer the question based only on the provided context. Do not use any additional context. Your answer should be concise and accurate. In case you are unable to answer the question, you should state that you can't answer the question. Do not guess and do not suggest your own solutions.

Input:

- **Document:** `{document}`
- **Question:** `{question}`

Output:

- *A response answering the question based on the provided document.*

Baseline Prompt with CoT

You are given a financial question and a financial document. Your task is to answer the question based on the document. Think step-by-step, and describe your reasoning process clearly before providing the final answer. You must provide the correct answer in a clear manner. Begin by describing your detailed reasoning process in a step-by-step manner, and then provide the final answer.

Input:

- **Document:** `{document}`
- **Question:** `{question}`

Output:

- *A response answering the question based on the provided document, including a step-by-step reasoning process.*

Financial Prompt with CoT

You are given a financial text extracted from 10-K or 10-Q files and a question written by domain experts. Your task is to answer the question based only on the provided context. Do not use any additional context. Your answer should be concise and accurate. In case you are unable to answer the question, you should state that you can't answer the question. Do not guess and do not suggest your own solutions. Think step-by-step, and describe your reasoning process clearly before providing the final answer. You must provide the correct answer in a clear manner. Begin by describing your detailed reasoning process in a step-by-step manner, and then provide the final answer.

Input:

- **Document:** `{document}`
- **Question:** `{question}`

Output:

- *A response answering the question based on the provided document, including a step-by-step reasoning process.*

C Human Evaluation Experiment results

We provide additional details about our judge alignment experiment. Fig. 6 displays the detailed confusion matrix of our LLM judge relative to human scores, and Table 6 show the stability of the LLM judge across two different models' outputs.

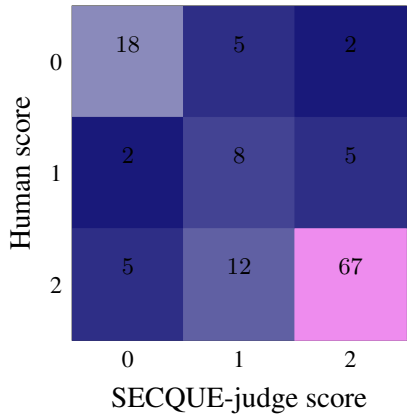

Figure 6: Confusion matrix heatmap comparing human scores to SECQUE-judge scores.

Table 6: Stability test for SECQUE-Judge for the 62 outputs from each model. *Both* is the average for all 124 (as shown in Table 4.)

Data source	#Answers	Alignment Metrics			
		F1(2)	precision(2)	recall(2)	accuracy
Both	124	0.85	0.905	0.8	0.75
GPT-4o	62	0.86	0.895	0.83	0.76
Llama-3.3-70B	62	0.835	0.915	0.77	0.74

D Full List of Accessions

Table 7 lists the exact filings used in SECQUE.

Table 7: Accession Numbers and Filing Periods

Accession Number	Company Name	From	Filing Date
0000004962-24-000052	AMERICAN EXPRESS CO	10-Q	2024-07-19
0000004962-24-000013	AMERICAN EXPRESS CO	10-K	2024-02-09
0000732717-24-000009	AT&T INC.	10-K	2024-02-23
0000320193-24-000081	Apple Inc.	10-Q	2024-08-02
0000320193-24-000069	Apple Inc.	10-Q	2024-05-03
0000320193-23-000106	Apple Inc.	10-K	2023-11-03
0000320193-22-000108	Apple Inc.	10-K	2022-10-28
0000070858-24-000208	BANK OF AMERICA CORP /DE/	10-Q	2024-07-30
0000070858-24-000156	BANK OF AMERICA CORP /DE/	10-Q	2024-04-30
0001390777-24-000105	Bank of New York Mellon Corp	10-Q	2024-08-02
0000093410-24-000040	CHEVRON CORP	10-Q	2024-08-07
0000811156-24-000084	CMS ENERGY CORP	10-Q	2024-04-25
0000021344-24-000044	COCA COLA CO	10-Q	2024-07-29
0000021344-24-000009	COCA COLA CO	10-K	2024-02-20
0001393612-24-000047	Discover Financial Services	10-Q	2024-07-31
0001393612-24-000010	Discover Financial Services	10-K	2024-02-23
0000034088-24-000050	EXXON MOBIL CORP	10-Q	2024-08-05
0001262039-24-000037	Fortinet, Inc.	10-Q	2024-08-08
0001262039-24-000014	Fortinet, Inc.	10-K	2024-02-26
0001562762-24-000034	Frontier Communications Parent, Inc.	10-K	2024-02-23
0001193125-24-168943	GENERAL MILLS INC	10-K	2024-06-26
0001193125-23-177500	GENERAL MILLS INC	10-K	2023-06-28
0000886982-24-000022	GOLDMAN SACHS GROUP INC	10-Q	2024-08-02
0000886982-24-000016	GOLDMAN SACHS GROUP INC	10-Q	2024-05-03
0000886982-23-000011	GOLDMAN SACHS GROUP INC	10-Q	2023-11-03
0000045012-24-000007	HALLIBURTON CO	10-K	2024-02-06
0000773840-24-000051	HONEYWELL INTERNATIONAL INC	10-Q	2024-04-25
0000051143-24-000012	INTERNATIONAL BUSINESS MACHINES CORP	10-K	2024-02-26
0000091419-24-000054	J M SMUCKER Co	10-K	2024-06-18
0000091419-22-000049	J M SMUCKER Co	10-K	2022-06-16
0000200406-24-000013	JOHNSON & JOHNSON	10-K	2024-02-16
0000019617-24-000453	JPMORGAN CHASE & CO	10-Q	2024-08-02
0000019617-24-000326	JPMORGAN CHASE & CO	10-Q	2024-05-01
0000019617-24-000225	JPMORGAN CHASE & CO	10-K	2024-02-16
0000753308-24-000008	NEXTERA ENERGY INC	10-K	2024-02-16
0000320187-18-000142	NIKE INC	10-K	2018-07-25
0001045810-24-000029	NVIDIA CORP	10-K	2024-02-21
0000078003-24-000166	PFIZER INC	10-Q	2024-08-05
0000080424-24-000083	PROCTER & GAMBLE Co	10-K	2024-08-05
0000080424-23-000073	PROCTER & GAMBLE Co	10-K	2023-08-04
0001560327-24-000021	Rapid7, Inc.	10-K	2024-02-26
0001558370-24-001532	SIMON PROPERTY GROUP INC /DE/	10-K	2024-02-22
0001628280-24-002390	Tesla, Inc.	10-K	2024-01-29
0000950170-22-000796	Tesla, Inc.	10-K	2022-02-07
0000899689-24-000005	VORNADO REALTY TRUST	10-K	2024-02-12