FCMR: Robust Evaluation of Financial Cross-Modal Multi-Hop Reasoning

Anonymous ACL submission

Abstract

Real-world decision-making often requires integrating and reasoning over information from multiple modalities. While recent multimodal large language models (MLLMs) have shown promise in such tasks, their ability to perform multi-hop reasoning across diverse sources remains insufficiently evaluated. Existing benchmarks, such as MMQA, face challenges due to (1) data contamination and (2) a lack of complex queries that necessitate operations across more than two modalities, hindering accurate performance assessment. To address this, we present Financial Cross-Modal Multi-Hop Reasoning (FCMR), a benchmark created to analyze the reasoning capabilities of MLLMs by urging them to combine information from textual reports, tables, and charts within the financial domain. FCMR is categorized into three difficulty levels—Easy, Medium, and Hard—facilitating a step-by-step evaluation. In particular, problems at the Hard level require precise cross-modal three-hop reasoning and are designed to prevent the disregard of any modality. Experiments on this new benchmark reveal that even state-of-the-art MLLMs struggle, with the best-performing model (Claude 3.5 Sonnet) achieving only 30.4% accuracy on the most challenging tier. We also conduct analysis to provide insights into the inner workings of the models, including the discovery of a critical bottleneck in the information retrieval phase.

1 Introduction

Despite the recent progress in AI (Touvron et al., 2023; Anthropic, 2024; OpenAI, 2024; *inter alia*), developing systems capable of human-level reasoning remains a significant challenge. Human cognition involves integrating information from multiple modalities to comprehend and make decisions in the real world. A domain that primarily requires such a comprehensive understanding is finance,

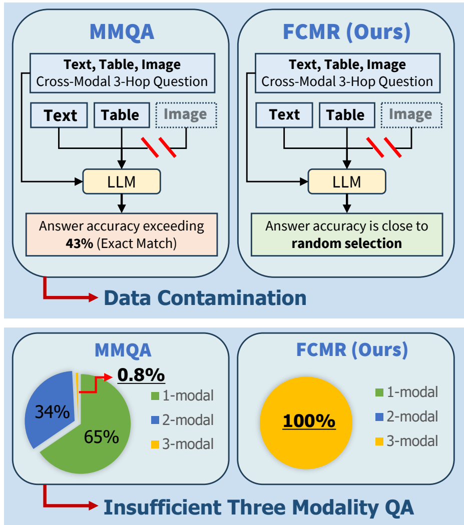
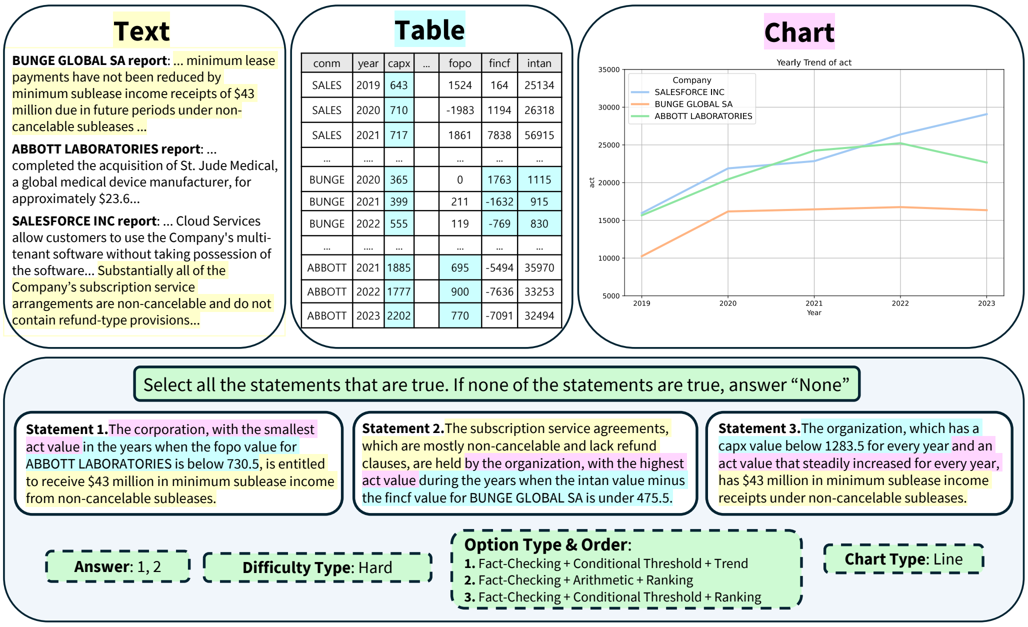


Figure 1: Prior benchmarks for cross-modal multi-hop reasoning, such as MMQA (Talmor et al., 2021), exhibit some flaws. MMQA's cross-modal three-hop questions are often solvable without images, and its complexity is limited, with only 0.8% of instances having three modalities. In contrast, FCMR addresses these issues.

where analysts often need to simultaneously examine textual reports, tabular data (e.g., balance sheets), and visual data (e.g., charts). For example, verifying the statement from Figure 2—*"The corporation, with the smallest act value in the years when the fopo value for ABBOTT LABORATORIES is below 730.5, is entitled to receive $43 million in minimum sublease income from non-cancelable subleases."*—one must consider all the relevant clues provided by each source, an ability we refer to as **cross-modal multi-hop reasoning**.

While the literature (Chen et al., 2020; Hannan et al., 2020; Talmor et al., 2021; Chang et al., 2022) presents initial attempts to evaluate the cross-modal multi-hop reasoning capabilities of multimodal large language models (MLLMs), these efforts exhibit several critical shortcomings that undermine their robustness. First, the heavy reliance



conm	year	capx	...	fopo	fincf	intan
SALES	2019	643		1524	164	25134
SALES	2020	710		-1983	1194	26318
SALES	2021	717		1861	7838	56915
...	....	...		...	...	...
BUNGE	2020	365		0	1763	1115
BUNGE	2021	399		211	-1632	915
BUNGE	2022	555		119	-769	830
...	....	...		...	...	...
ABBOTT	2021	1885		695	-5494	35970
ABBOTT	2022	1777		900	-7636	33253
ABBOTT	2023	2202		770	-7091	32494

Figure 2: An example from FCMR at the Hard difficulty level, where all statements require cross-modal three-hop reasoning. Highlights in yellow, cyan, and pink denote information from text, tables, and charts, respectively. The model must list all true statements and is correct only if its final prediction ("1, 2" in this case) is accurate. Information within dashed lines is used only for data generation and excluded from actual instances.

on Wikipedia as the foundation for most benchmarks raises concerns. As Wikipedia is widely known to be a key resource in the pretraining of many recent models, evaluations using Wikipedia-based datasets risk introducing inherent biases. These biases may skew results in favor of models that simply recall memorized knowledge, rather than accurately assessing reasoning abilities on unseen data. Moreover, the scope of validation needs to expand to encompass professional domains, such as finance and science.

Second, current benchmarks are largely focused on testing straightforward problems, such as single- and two-hop reasoning. As shown in Figure 1, MMQA (Talmor et al., 2021)—one of the leading benchmarks in this field—features only about 0.8% of the queries that explicitly require *three-hop* cross-modal reasoning. Furthermore, in preliminary experiments, we discovered that GPT-4o (OpenAI, 2024) can solve the MMQA's most challenging problems with a 43% exact match accuracy, even without access to visual clues. This result highlights the urgent need to establish a higher standard for evaluating cross-modal multi-hop understanding in a more robust and reliable manner.

In this work, we propose **Financial Cross-Modal Multi-Hop Reasoning (FCMR)**, a novel benchmark designed to address the limitations of existing datasets in cross-modal multi-hop reasoning. FCMR provides multiple-choice QA samples that test the integration of facts from text, table, and chart modalities. For systematic assessment, it consists of three levels of difficulty—Easy, Medium, and Hard. As shown in Figure 1, every instance in FCMR necessitates understanding all three modalities to be answered correctly. In addition, problems at the Hard level explicitly demand cross-modal three-hop reasoning, making them more challenging (see Figure 2). Since FCMR is built using data sources from the financial domain, it is relatively free from the risk of data contamination.

Experiments on FCMR confirm that it poses challenges even for state-of-the-art MLLMs, e.g., GPT-4o and Claude 3.5 Sonnet (Anthropic, 2024), encouraging research efforts to develop systems capable of reasoning across multiple modalities. For analysis, we define four fine-grained procedures of cross-modal multi-hop reasoning—Planning, Modality Identification, Information Retrieval, and Information Reasoning—and probe diverse models. We reveal that models particularly struggle with the Information Retrieval phase, implying that MLLMs often fail to extract precise information from a specific modality, even when they successfully identify where the required information is located. We also present valuable findings from other analyses, such as the observation that MLLMs still have difficulty with adding negative numbers.

Benchmarks	Cross-Modal 2-Hop?	Cross-Modal 3-Hop?	Contain Table?	Contain Image?	Domain Specific?
ManyModalQA	✗	✗	✓	✓	✗
CT2C-QA	✗	✗	✓	✓	✓
WebQA	✓	✗	✗	✓	✗
MuMuQA	✓	✗	✗	✓	✓
FinQA	✓	✗	✓	✗	✓
TAT-QA	✓	✗	✓	✗	✓
HybridQA	✓	✗	✓	✗	✗
OTT-QA	✓	✗	✓	✗	✗
TANQ	✓	✗	✓	✗	✗
MMQA	✓	✓✗	✓	✓	✗
FCMR (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison of datasets based on cross-modal reasoning and modality coverage. MMQA's ✓✗ shows that although it includes cross-modal three-hop reasoning, such instances constitute only 0.8% of the dataset.

2 Related Work

2.1 Cross-Modal Multi-Hop Reasoning

Benchmarking cross-modal multi-hop reasoning has received considerable attention. Efforts include WebQA (Chang et al., 2022) and MuMuQA (Reddy et al., 2022)—for two-hop cross-modal reasoning with text and images—as well as HybridQA (Chen et al., 2020), OTT-QA (Chen et al., 2021a), FinQA (Chen et al., 2021b), TAT-QA (Zhu et al., 2021), and TANQ (Akhtar et al., 2024), which include text and tables. However, all these datasets are limited to **only two modalities**, making them inadequate for evaluating more complex cases.

Meanwhile, datasets like ManyModalQA (Hannan et al., 2020) and CT2C-QA (Zhao et al., 2024) incorporate three modalities but lack an inherent focus on cross-modal multi-hop reasoning. MMQA (Talmor et al., 2021), in contrast, deals with **three modalities**—text, tables, and images—and requires **three-hop reasoning**, setting it apart from others. It has served as the de facto standard for evaluating related methods (Rajabzadeh et al., 2023; Yu et al., 2023; Luo et al., 2023; Zhang et al., 2024a; Abaskohi et al., 2024). MMCoQA (Li et al., 2022) and MMCV (Wang et al., 2024) have been developed as extensions of MMQA.

The characteristics of the benchmarks are summarized in Table 1. FCMR is crafted to address the limitations of previous ones, particularly MMQA.

2.2 Limitations of MMQA

We briefly revisit the drawbacks MMQA to emphasize the need for a new, robust benchmark for cross-modal multi-hop reasoning.

Dataset: MMQA	Image?	Exact Match (%)	F1 Score (%)
Random Selection	-	0.0	1.2
GPT-4o	✗	43.4	46.2
	✓	63.4	67.5

Table 2: Experiments on a subset of MMQA requiring cross-modal three-hop reasoning reveal that GPT-4o performs reasonably well even without images. This implies that it already contains information derivable from input images, questioning the rigor of MMQA. For more details, refer to Appendix C.2.

Data Contamination As MMQA is constructed from Wikipedia, it is vulnerable to data contamination. That is, the model being tested may already possess internalized knowledge of certain facts, reducing its dependence on the dataset's provided input. In particular, Table 2 shows that GPT-4o can achieve reasonable performance on the most challenging part of MMQA—questions intentionally tailored for requiring a combination of information from three modalities—*without* relying on visual hints.[1] This suggests that MMQA falls short of effectively measuring cross-modal multi-hop reasoning ability as it was originally intended.

Lack of Cross-Modal Three-Hop Cases Only about 0.8% (205 instances) of the MMQA dataset consists of cross-modal three-hop reasoning, while the majority comprises either one-hop or two-hop questions. This scarcity restricts its effectiveness in thoroughly evaluating a model's performance on complex reasoning tasks with intricate interactions across text, tables, and images.

3 Proposed Benchmark: FCMR

We introduce **Financial Cross-Modal Multi-Hop Reasoning (FCMR)**, a new benchmark created to alleviate the shortcomings of MMQA and enable a more comprehensive evaluation of cross-modal multi-hop reasoning. FCMR includes three modalities—text, tables, and charts—and presents questions that entail selecting all correct statements from a set of three. The tested model must identify all true statements in the problem and is considered correct only if its final prediction is accurate.

3.1 Datset Generation Framework: CMRGen

We propose **Cross-Modal Multi-Hop Reasoning Generator (CMRGen)**, a framework that facilitates the construction of cross-modal multi-hop reasoning datasets across various domains. CMRGen

[1]Figure 9 presents an example of data leak in MMQA.

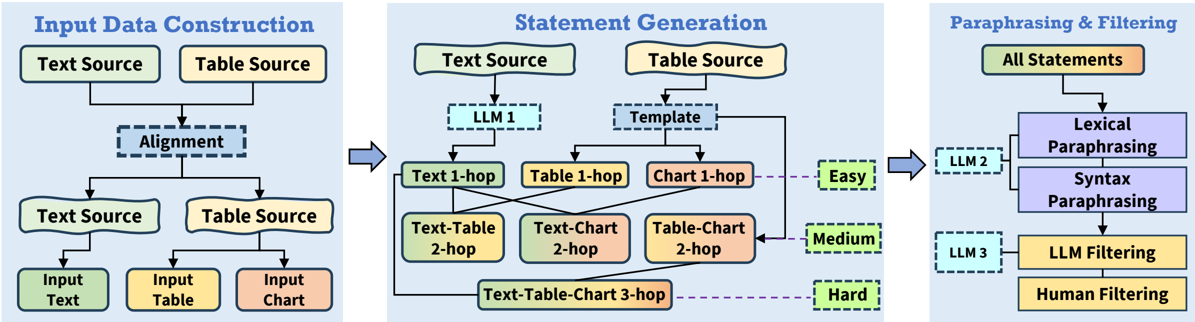


Figure 3: **CMRGen** is an automated and efficient framework for building cross-modal multi-hop reasoning datasets. **(1) Modality Data Construction** extracts text, table, and chart modalities from sources sharing common entities. **(2) Statement Generation** produces cross-modal one-, two-, and three-hop statements using LLM and templates. **(3) Paraphrasing & Filtering** applies two-stage paraphrasing with LLMs, followed by LLM & Human filtering.

distinguishes itself from other cross-modal multi-hop dataset generation methods with its highly automated and cost-effective pipeline. Notably, while generating a single question in MMQA incurs a cost of $0.33, our method reduces this to $0.004 per question. Furthermore, the framework demonstrates high adaptability to various domains and offers seamless control over difficulty levels, ranging from Easy to Hard. In this study, we focus on the financial domain, where complex interactions among text, tables, and charts frequently occur, making it an ideal testbed for evaluating cross-modal multi-hop reasoning. However, the proposed pipeline is also readily applicable to domains other than finance, demonstrating its flexibility.[2]

3.2 Procedure of CMRGen

CMRGen have three stages, as depicted in Figure 3: (1) Input Data Construction, (2) Statement Generation, and (3) Paraphrasing & Filtering. We explain each step using **FCMR** as an example. Details of the procedure can be found in Appendix B.

(1) Input Data Construction In the first stage, we prepare and organize data for the text, table, and chart modalities. As the origin of information, CMRGen utilizes two sources: Text Source and Table Source. For FCMR, the Text Source consists of Annual 10-K Reports collected from the SEC EDGAR database,[3] while the Table Source is derived from Annual Simplified Financial Statements provided by WRDS Compustat.[4] We then filter entries that share common company entities, aligning the two sources. Finally, we construct each data instance in FCMR, comprising a document, a table, and a chart about three companies. In the next step, this instance will be supplemented with three statements serving as questions about the contents generated in this stage. Note that the chart is created by plotting specific columns from the Table Source using custom scripts, and the columns used to create the chart are removed from the table.

(2) Statement Generation In the second phase, diverse forms of statements (i.e., questions) are crafted for each FCMR instance. We leverage GPT-4o-mini to generate text-based one-hop statements by extracting relevant facts from the Text Source. In addition, by leveraging various templates tailored to real-world financial scenarios—such as Trend, Ranking, Conditional Threshold, and Arithmetic—we create table-based and chart-based one-hop statements based on the Table Source. We then combine these single-modal one-hop statements across entities to construct cross-modal two-hop statements, which are further merged to create cross-modal three-hop statements. Each statement is categorized into Easy, Medium, or Hard based on the number of hops required for reasoning. The complete taxonomies of statement types and templates are presented in Table 8 and Table 9.

(3) Paraphrasing & Filtering In the final step, we apply two stages of lexical and syntactic paraphrasing using GPT-4o to enhance the diversity of expression and structure in the statements. Afterward, we conduct both LLM-based filtering (using Claude 3.5 Sonnet) and human-based filtering to ensure semantic accuracy. For Hard-level instances, human experts thoroughly review them to eliminate ambiguity and guarantee high quality. By uti-

[2]We showcase the application of the proposed method in material science. For more details, refer to Appendix A.

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

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

Dataset: FCMR (Hard)	Image?	Accuracy
Random Selection	-	12.28
GPT-4o	✗	14.71
	✓	24.37

Table 3: Replication of experiments from Table 2 with FCMR. Despite the inherently challenging nature of the benchmark, GPT-4o's performance drops to near random selection when charts are omitted, suggesting that FCMR is relatively robust against data contamination.

lizing distinct LLMs in various procedures, we aim to reduce unintended model-oriented biases.

3.3 Multiple-Choice Design

Previous research on cross-modal multi-hop reasoning has often employed descriptive or short-answer formats, evaluated with metrics like F1 or Exact Match. These approaches might result in inaccurate evaluations by penalizing semantically appropriate answers that slightly deviate in form. To address this, we adopt a multiple-choice format with three statements. Recent work (Pang et al., 2024) argues that single-choice question formats are more suitable for model evaluation than free-form answers, supporting our decision. In contrast to existing multimodal benchmarks (Li et al., 2023; Yue et al., 2024; Ying et al., 2024; Zhang et al., 2024b; Liu et al., 2024b) that typically rely on a single correct answer, our setup allows for zero to three correct statements. This strategy enhances the complexity of the reasoning process needed to answer the problem accurately, requiring a more comprehensive synthesis of all provided statements. It also allows for precise evaluation of models' cross-modal multi-hop reasoning capabilities.

3.4 Data Quality Control

To uphold high data quality, we implement multifaceted verification protocols.[5] Specifically, we utilize Word Position Deviation (WPD) and Lexical Deviation (LD) metrics (Liu and Soh, 2022) to evaluate paraphrasing quality and compare these values with those from MRPC (Dolan and Brockett, 2005) and PAWS (Zhang et al., 2019). The outcomes, presented in Table 6, confirm the superiority of our paraphrasing method. Furthermore, to verify that our dataset avoids the contamination issue identified in MMQA, we replicate the contamination experiment under the same conditions. As shown in Table 3, when the chart images are withheld, GPT-4o's performance approximates random guessing, alleviating the risk of data contamination in FCMR. We further mitigate potential biases by equalizing modality order, statement types, and answer distributions. Figure 20 demonstrates that our benchmark is well-balanced across various perspectives. The final dataset consists of 757 Easy, 728 Medium, and 714 Hard instances, demand 100% cross-modal three-hop reasoning, setting FCMR apart from MMQA and its variants.

[5]Refer to Appendix C for full details of our strategies.

Metric: Accuracy (%)	Easy	Medium	Hard	Avg
Random Selection	12.2	12.91	12.28	12.46
Multimodal Large Language Models (MLLMs)				
ChartInstruct-Llama2	11.49	12.64	10.78	11.64
llama3-llava-next-8b-hf	16.86	12.22	11.53	13.54
MiniCPM-V-2_6	16.38	11.68	13.17	13.74
Qwen2-VL-7B-Instruct	17.57	13.32	12.04	12.32
Llama 3.2 90B-Vision	42.47	21.60	13.73	25.94
GPT-4o mini	49.14	21.98	13.03	28.05
Gemini 1.5 Flash	57.33	26.65	13.43	32.80
Gemini 1.5 Pro	63.01	31.18	22.27	38.82
GPT-4o	64.20	43.70	24.37	44.09
Claude 3.5 Sonnet	**75.43**	**50.82**	**30.39**	**52.21**
Large Language Models (LLMs) with Deplot				
Qwen2-7B-Instruct	21.66	11.95	14.01	15.87
Llama 3.1 8B-Instruct	30.91	13.05	10.36	18.11
Llama 3.1 70B-Instruct	46.37	17.86	14.01	26.08
Llama 3.2 90B-Vision	50.20	22.39	11.90	28.16
GPT-4o mini	57.60	26.51	12.61	32.24
GPT-4o	**68.69**	**49.18**	32.91	**50.26**
Claude 3.5 Sonnet	66.84	46.15	**36.13**	49.71

Table 4: Performance of MLLMs and LLMs on FCMR. For LLMs, charts are converted into tables using Deplot. The best performance at each difficulty level and category is highlighted in **bold**.

4 Experiments

4.1 Experimental Setup

We evaluate a wide range of MLLMs and LLMs on FCMR under a zero-shot CoT setting, where no task-specific tuning or demonstration is provided. All models are prompted with the same template in Figure 22. Tables are represented in JSON format. For proprietary models, we employ the Claude version *claude-3-5-sonnet-20241022*, the GPT-4o version *gpt-4o-2024-08-06.*, and Gemini version *gemini-1.5-pro-002*. We also test several open-source models: Llama variants (Touvron et al., 2023), Qwen (Yang and Yang, 2024), MiniCPM (Yao and Yu, 2024), Llava (Liu et al., 2024a), and ChartInstruct (Masry and Shahmohammadi, 2024).

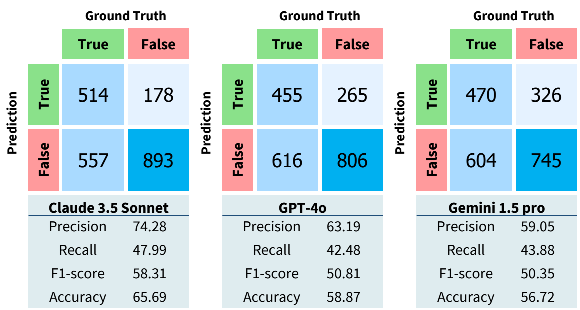


Figure 4: Confusion matrices for three advanced MLLMs, with metrics in percentages (%).

4.2 Main Results

Performance of MLLMs Table 4 reports the performance of various MLLM across different levels. Most open-source models perform just above random chance at the Easy level, focused on single-modal, one-hop reasoning, confirming FCMR as a challenging benchmark. Proprietary models perform better, demonstrating a remarkable gap in reasoning ability. However, at the Hard level, which necessitates full cross-modal three-hop integration, even sophisticated models, including Claude 3.5 Sonnet, achieve only around 30%. This result underscores FCMR's challenging nature and the need for developing more advanced reasoning strategies.

Performance of (M)LLMs + Deplot For image-blind standard large language models (LLMs), we use Deplot (Liu et al., 2023) to convert charts into tables, ensuring that all models can access chart information. We also explore applying the same heuristic to MLLMs, as the literature suggests that MLLMs tend to rely more on textual clues than visual ones (Rahmanzadehgervi et al., 2024).

Experimental results indicate that open-source models with fewer than 8B parameters continue to perform comparably to random selection for tasks at the Medium and Hard levels. However, for the Easy category, they demonstrate superior performance compared to similarly sized MLLMs. Surprisingly, even advanced MLLMs such as GPT-4o and Claude 3.5 Sonnet achieve performance gains in certain cases, suggesting that their visual interpretation capabilities are still not perfect. In Section 6.2, we dive deeper into this phenomenon.

5 Analysis

In this section, we analyze the inner workings of closed-source MLLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro—which outperform smaller open-source models. We focus on their performance at the Hard level, as this subset presents the most challenging questions for the models.

5.1 Statement-Level Analysis

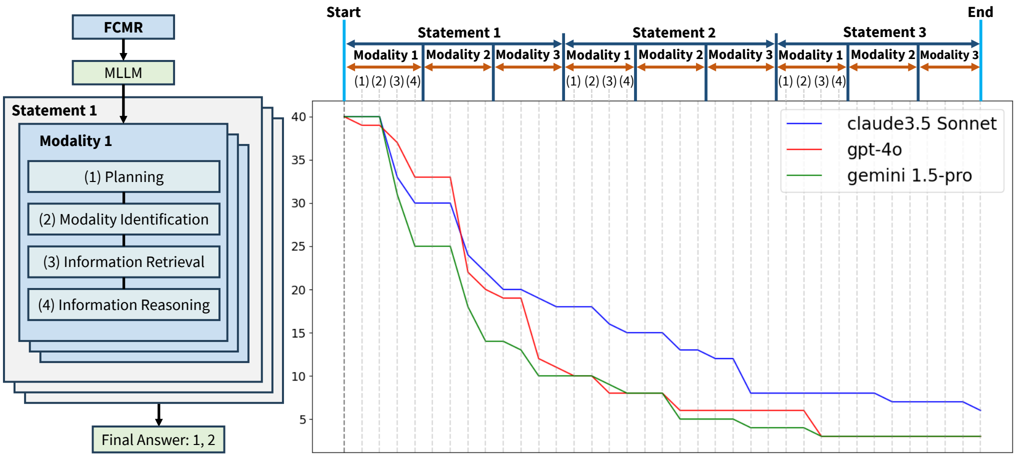
While each problem in FCMR requires models to draw an overall conclusion on three statements, their partial solutions for each statement can provide insight into how well each model handles diverse cases. We gather statistics on each model's predictions for every statement and construction confusion matrices for analysis. We have 714 $\times$ 3 = 2,142 statements with gold-standard answers, evenly split into true (1,071) and false (1,071) ones. Each model's prediction is annotated for these statements, forming the matrices in Figure 4.

While all three models show limitations in precision, recall, and F1-score, Claude 3.5 Sonnet achieves comparatively better performance. With a high precision of 74.27, Claude effectively minimizes false positives, reflecting an ability to accurately classify positive cases. However, its recall remains limited to 47.99, indicating a reduced capacity to capture all true positives. Despite this trade-off, Claude achieves the highest accuracy at 65.69 and an F1-score of 58.31, outperforming the other two models, GPT-4o and Gemini 1.5 Pro.

Moreover, all three models display a notable tendency to adopt a conservative decision-making strategy by defaulting to *false* in cases of uncertainty or low confidence. This behavior reflects a low-risk approach aimed at reducing false positive classifications, even if it results in a lower recall.

5.2 Stage-Based Analysis

In the main experiments, we observed advanced MLLMs follow a similar sequence of reasoning steps to solve problems in FCMR. Based on this, we define four fine-grained reasoning steps to identify where errors commonly occur. The four stages are specified as: **(1) Planning:** identifying the required values, **(2) Modality Identification:** recognizing which modality contains these values, **(3) Information Retrieval:** extracting relevant information from the identified modality, and **(4) Information Reasoning:** reasoning over the extracted information under the given conditions. Each instance includes three statements, each requiring the four-step process across all three modalities. Models must execute steps (1)-(4) three times per statement and repeat this for all three statements before answering (see Figure 5 (a)).



(a) Fine-grained reasoning stages. (b) The number of successful inferences after each fine-grained reasoning stage.

Figure 5: Fine-grained stage-based analysis of three advanced MLLMs. This unique strategy discovers several intriguing findings, including that the models fails most often at (3) Information Retrieval.

We manually monitor MLLMs' inference trajectories on 40 given samples. After each fine-grained stage, we record the number of problems successfully processed by the models, forming a success history diagram in Figure 5 (b). The visualization reveals an intriguing pattern: for most samples, MLLMs fail at some stage before completing the reasoning steps for the first statement. Specifically, we observe a notable performance drop at the [Statement 1, Modality 2, (3)-(4)] stage. This suggests that while the models handle the first modality relatively well, they struggle considerably when they encounter a second modality. Interestingly, GPT-4o outperforms Claude 3.5 Sonnet in the first modality reasoning step of the first statement, but Claude surpasses GPT-4o starting from the second modality phase. As the models progress to the second statement, Claude's performance diverges further from the others, showcasing more robust and sustained reasoning capabilities.[6]

We further explore model failures by identifying which of the four reasoning steps (1)-(4) these failures occur in, regardless of statement and modality. As illustrated in Figure 21, the most common cause of failure across MLLMs occurs at step (3) Information Retrieval—failing to extract the required information from the identified modality. The second most frequent failures arise at stage (4), Information Reasoning, where models struggle to correctly apply the retrieved information to the given conditions. Notably, Gemini 1.5 Pro exhibits a higher proportion of failures at step (4), meaning that even when it successfully retrieves information, it has difficulty reasoning over it.[7]

Level	Text	Table	Chart	Total
Easy	1 (4%)	5 (21%)	18 (75%)	24
Medium	5 (16%)	6 (19%)	20 (65%)	31
Hard	6 (14%)	13 (32%)	22 (54%)	41

Table 5: Error counts and proportions by modality for Claude 3.5 Sonnet across 90 statements per level.

While Claude and GPT make no modality identification errors (stage (2)), Gemini 1.5 Pro occasionally misidentifies modalities, such as confusing chart values with table values. Gemini 1.5 Pro has no failures at the Planning stage (stage (1)). In contrast, GPT and Claude sometimes skip planning for the third modality after successfully handling the first two, leading to task failure. This shows that while all models struggle at later reasoning steps, GPT and Claude particularly struggle to maintain a consistent strategy across modalities.

6 Case Study

Given Claude 3.5 Sonnet's effectiveness for FCMR, we conduct case studies to derive insights for enhancing cross-modal multi-hop reasoning.

6.1 Error Rate by Modality

Table 5 displays the numbers and proportions of statements Claude fails to interpret correctly, based on randomly selected 90 statements for each diffi-

[6]Fine-grained model answer examples are in Figure 10.

[7]Examples are shown in Figure 11 and Figure 12.

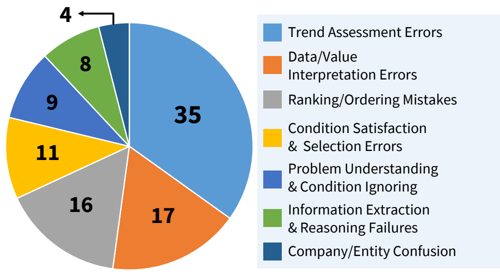


Figure 6: Error categories and their distributions across 100 manually labeled error cases for Claude 3.5 Sonnet.

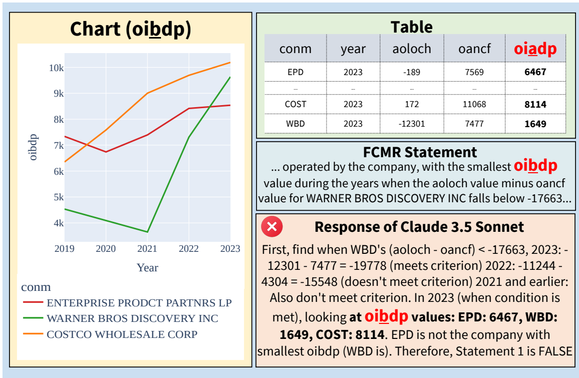


Figure 7: A case study on Claude's errors: The condition, "*the years when the aoloch value minus the oancf value for WARNER BROS DISCOVERY INC falls below -17663,*" is satisfied for 2023. This task thus requires finding a company with a small oibdp value for 2023. Although the oibdp value is in the chart, not the table, Claude ignores the chart and incorrectly substitutes the oiadp value, likely due to the similar column names.

culty level.[8] At the Easy level, Claude frequently struggles with interpreting charts, performing noticeably worse compared to its handling of text and tables. This disparity indicates that MLLMs generally exhibit weaker capabilities in interpreting charts than in processing textual or tabular data. However, as the difficulty level increases, errors in text and tables become more prominent. This shift is likely attributed to the increased complexity of reasoning chains, which raises the likelihood of errors in interpreting text and tables.

6.2 Chart Interpretability

Building on previous findings that even Claude struggles with chart interpretation, we analyze the specific conditions that pose the greatest challenges. Table 7 shows that among line, bar, scatter, and pie charts, scatter plots are the most challenging due to their less structured representations. In contrast, MLLMs find it easier to identify trends in line or bar charts, which provide clearer patterns. Ranking tasks appear simpler than trend analysis, as they involve identifying extremes, whereas trend detection demands more advanced inference.

6.3 Error Classification and Inspection

To manually inspect Claude's working patterns, we examine 100 error cases where it was unable to provide accurate answers. The distribution of these errors is visualized in Figure 6. Real error cases can be found in Figure 7 and Appendix D.

The most common issue is again about incorrectly assessing trends in charts, such as misidentifying whether values were increasing or decreasing (35 cases). The second most frequent type of error, with 16 instances, pertains to the misidentification of top-ranked entities or overall rankings. The model also fails in 11 cases to identify entities meeting the given conditions or applies the conditions incorrectly.

A total of 17 errors stem from misinterpreting data or values, including mishandling sums, negative values, subtle differences, or column confusion. In 4 cases, the model conflates the identities of companies or entities. It also makes 8 errors in information extraction and reasoning, such as inaccurately extracting facts from text or drawing unjustified conclusions despite correct information. Lastly, 9 instances involve misunderstanding problem instructions, ignoring required modalities, or reaching illogical conclusions.

These results highlight that successful cross-modal multi-hop reasoning demands both strong reasoning abilities and effective interpretative skills for each modality. Therefore, enhancing a model's capacity to interpret individual modalities is essential for enabling effective multi-hop reasoning.

7 Conclusion

We introduce FCMR, a new benchmark designed to evaluate the cross-modal multi-hop reasoning ability of MLLMs. We evaluate the performance of various state-of-the-art MLLMs, revealing that current models continue to struggle with reasoning across different modalities. As future work, we plan to develop methods to enhance the performance of MLLMs based on the observations and analyses presented in this study.

[8] Errors unrelated to modality, such as misinterpreting conditions, are excluded from this analysis.

Limitations

We present several points that can serve as the foundation for improving this work and initiating future research.

Heavy Reliance on Manual Analysis Our analysis required extensive manual sampling and verification to ensure high-quality insights. Future work could aim to automate this process, enhancing efficiency and scalability.

Potential for Extension to Other Domains While we have conducted extensive experiments and analyses in the financial domain using FCMR, the proposed dataset generation framework, CMRGen, has the potential to extend beyond the financial and material science domains, enabling the creation of datasets in fields such as law, biology, medicine, and electrical engineering. Future work can consider performing comprehensive performance evaluations of various models across these domains.

Room for Prompt Optimization Providing optimized prompts for each MLLM could improve performance in evaluation. However, our top priority in this work is to test all models under equal conditions. Additionally, recent models are known to be optimized with diverse prompts during the post-training phase, making them increasingly robust to variations in instructions and prompts. For example, our experiments show that questions requiring complex reasoning often naturally elicit strategies similar to CoT, even when such strategies are not explicitly mentioned.

References

Amirhossein Abaskohi, Spandana Gella, Giuseppe Carenini, and Issam H Laradji. 2024. Fm2ds: Few-shot multimodal multihop data synthesis with knowledge distillation for question answering. *arXiv preprint arXiv:2412.07030*.

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*.

Anthropic. 2024. Claude: Large language model by anthropic. Available online at https://www.anthropic.com. Accessed on 2024-10-15.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

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.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Timothy Liu and De Wen Soh. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pages 216–233, Cham. Springer Nature Switzerland.

Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore. Association for Computational Linguistics.

Ahmed Masry and Mehrad Shahmohammadi. 2024. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10387–10409, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.

Hossein Rajabzadeh, Suyuchen Wang, Hyock Ju Kwon, and Bang Liu. 2023. Multimodal multi-hop question answering through a conversation between tools and efficiently finetuned large language models. *Preprint*, arXiv:2309.08922.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2024. Piecing it all together: Verifying multi-hop multimodal claims. *arXiv preprint arXiv:2411.09547*.

An Yang and Baosong Yang. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yuan Yao and Tianyu Yu. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57116–57198. PMLR.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Unified language representation for question answering over text, tables, and images. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4756–4765, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Qing Zhang, Haocheng Lv, Jie Liu, Zhiyun Chen, Jianyong Duan, Hao Wang, Li He, and Mingying Xu. 2024a. An entailment tree generation approach for multimodal multi-hop question answering with mixture-of-experts and iterative feedback mechanism. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4814–4822.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024b. M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models. In

Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Bowen Zhao, Tianhao Cheng, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2024. Ct2c-qa: Multimodal question answering over chinese text, table and chart. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3897–3906.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3277–3287, Online. Association for Computational Linguistics.

A Details of Material Cross-Modal Multi-Hop Reasoning (Material-CMR)

The proposed CMRGen framework is easily adaptable to various domains, and as an example, we applied it to the field of Material Science to create the Material Cross-Modal Multi-Hop Reasoning (Material-CMR) dataset. We construct a Table Source containing material properties and a Text Source describing the materials through The Materials Project. The Materials Project[9] 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 FCMR into materials. Following the same steps proposed in CMRGen, we build Table Sources and Text Sources for Material-CMR. The Table Sources include columns for material properties such as band gap, density, volume, and formation energy per atom, while the Text Sources provide descriptions of the material's crystal structure, structural characteristics, and properties. Using these Text Sources and Table Sources, we create datasets combining text, table, and chart input data. Subsequently, we generate single-modal one-hop, cross-modal two-hop, and cross-modal three-hop statements through GPT-4o-mini and templates, categorizing them by difficulty level. Also, A two-stage paraphrasing process is employed to maximize diversity. An example of the dataset instance is shown in Figure 8.

B Details on the Procedure of CMRGen

This section provides a comprehensive explanation of the dataset generation process for FCMR, detailing the sources and preprocessing steps for input table, text, and chart modalities, as well as the construction of table sources, text sources, and distractors.

B.1 Table Source

WRDS Compustat: Annual Simplified Financial Statements Considering the practicality of cross-modal multi-hop reasoning, we utilize *annual financial statements*, an essential element in real corporate analysis, as the table data source. Wharton Research Data Services (WRDS) Compustat[10] 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. 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 24. 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.

B.2 Text Source

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).[11] 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.

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

[9] https://next-gen.materialsproject.org/

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

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

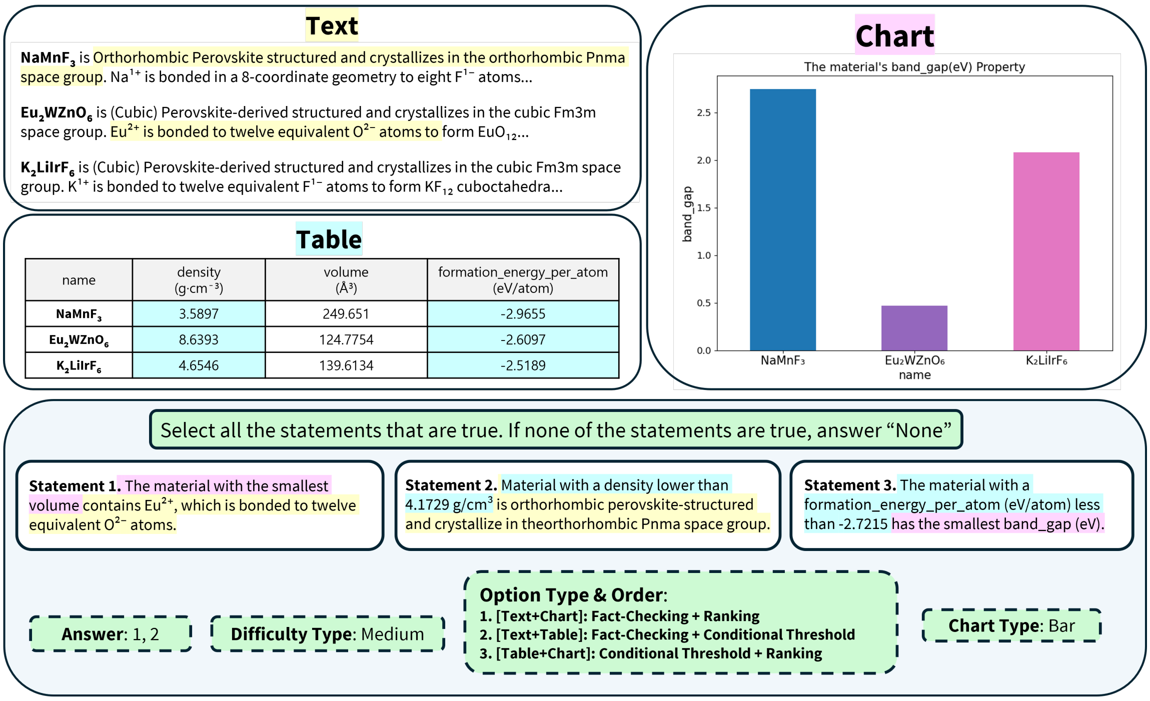


Figure 8: An example from Material-CMR.

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[12].

This 10-K report data will later be used to construct the input text.

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.

B.3 Table Source Construction

A table source serves as an intermediate bridge connecting the text, table, and chart modalities and is used as a base anchor for creating multi-hop statements. After sampling three companies from the Annual Simplified Financial Statement data of the 101 companies, we construct the table source 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 table source consists of the Annual Simplified Financial Statement data of three companies, each having five years, and is composed of ten columns.

B.4 Text Source Construction

The 10-K reports obtained through SEC EDGAR are too lengthy to use entirely at once as input text. Therefore, we divide the 10-K reports of each company into chunks of three consecutive paragraphs. Later, these chunks will be used as the input text.

B.5 Input Data Construction: Text, Tables, and Charts

A single data instance contains a total of three companies. From the table source with three companies, we select one column without NaN values as the chart column and convert the table source

[12] https://www.sec.gov/files/reada10k.pdf

excluding the chart column into the input table and the chart column into the input chart. To preserve the structural information of the table, the input table is constructed in JSON format, and to ensure data diversity, the input chart uses three different libraries and four chart types (line, bar, scatter, pie) commonly used in financial domains. The input text corresponds to the text sources of the three companies. All of these processes are automated through a Python script.

B.6 Distractor Generation

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

B.7 Input Chart Code Generation

The input chart in FCMR consists of four types: Line, Bar, Scatter, and Pie, generated using visualization libraries such as matplotlib, seaborn, and plotly. To enhance chart diversity and mitigate data bias, 16 font types, including [‘Arial’, ‘Verdana’, ‘Times New Roman’, ‘Courier New’, ‘Georgia’, ‘Comic Sans MS’, ‘Tahoma’, ‘Cambria’, ‘Microsoft YaHei’, ‘Nirmala UI’, ‘Calibri’, ‘Consolas’, ‘Segoe UI’, ‘Garamond’, ‘Century Schoolbook’, ‘Book Antiqua’], were applied to text within the charts. The font size for titles, labels, legends, and ticks was randomly selected within predefined minimum and maximum thresholds. To ensure clear visual distinction, the color palette consisted of seven colors: [‘#1f77b4’, ‘#ff7f0e’, ‘#2ca02c’, ‘#d62728’, ‘#9467bd’, ‘#8c564b’, ‘#e377c2’]. The thickness of lines and bars was also randomly selected within predefined thresholds. To clearly visualize trends and rankings in charts, we introduced controlled variance in yearly data values, ensuring the design avoids cases where differentiation is visually ambiguous.

C Details of Data Quality Control

C.1 Paraphrasing Quality

To evaluate the quality of Lexical-Syntax 2-Stage Paraphrasing, we employed the Word Position Deviation (WPD) and Lexical Deviation (LD) metrics proposed in (Liu and Soh, 2022). WPD assesses the syntactic diversity of paraphrased sentences, while LD evaluates lexical diversity. For an objective comparison, as shown in Table 6, we benchmarked the WPD and LD metrics of FCMR against prominent paraphrasing datasets such as MRPC (Dolan and Brockett, 2005) and PAWS (Zhang et al., 2019), demonstrating the superior quality of our paraphrasing. Additionally, we validated semantic consistency using Claude 3.5 Sonnet to filter out samples where the paraphrased sentences were flagged as semantically altered. For hard statements with longer sentence lengths, we considered the potential for ambiguity introduced by paraphrasing. Consequently, all instances were manually reviewed, and sentences with ambiguous meanings were revised accordingly.

Dataset	WPD	LD
MRPC	0.12	0.42
PAWS	0.07	0.13
FCMR (Ours)	**0.2**	**0.45**

Table 6: WPD (Word Position Deviation) represents syntactic diversity, and LD (Lexical Deviation) reflects lexical diversity. Both metrics indicate that higher scores correspond to greater diversity in paraphrasing.

C.2 Verification of Data Contamination

To ensure a fair comparison of data contamination between MMQA and FCMR under identical conditions, we evaluated instances requiring cross-modal three-hop reasoning from each dataset using the GPT-4o model under the following settings: (1) Random Selection, (2) Without Image Input, and (3) With Image Input. For MMQA, *Random Selection* involves randomly selecting a single word from the question, text, or table. In contrast, for FCMR, *Random Selection* involves randomly selecting one of the eight possible answers, ranging from none to (1, 2, 3).

C.3 Bias Mitigation Strategies

Due to the design requiring reasoning over three statements, there is a potential for bias to arise from specific factors. To minimize bias, we implemented several strategies. First, we ensured a balanced distribution of modality order types to prevent bias toward specific order configurations. Second, we adjusted the distribution of statement types to avoid overrepresentation of particular types. Third, we maintained an even distribution across the eight

Difficulty	Total	line	bar	scatter	pie
Easy	75.43	74.89	78.60	71.01	84.31
Medium	50.82	52.70	50.00	49.79	-
Hard	30.39	39.22	29.20	23.44	-

Table 7: Accuracy by chart type, based on Claude 3.5 Sonnet. 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.

answer types to reduce bias toward any specific answer type. The distributions of answer type, statement type, and library type across all difficulty levels are visualized in Figure 20.

D Case Study Examples

D.1 Trend Assessment Error

As shown in Figure 13, Claude struggles to identify increasing trends. This difficulty is particularly pronounced when interpreting cumulative bar charts or charts with ranges that include negative values, where the success rate of interpretation is significantly lower.

D.2 Ranking/Ordering Mistake

Figure 14 illustrates a case where the Claude model fails to accurately determine the ranking for a specific year from a chart. While the model performs better in identifying rankings compared to recognizing increasing or decreasing trends, its success rate remains significantly lower when interpreting cumulative bar charts or charts with ranges that include negative values.

D.3 Condition Satisfaction & Selection Error

The model sometimes fails to correctly identify a company or element that meets given conditions, or asserting that no such entity exists when one does. An example is in Figure 15.

D.4 Data/Value Interpretation Error

The model occasionally fails in calculations involving addition when negative numbers are included or when the number of terms exceeds three. Additionally, there are instances where it fails to correctly compare the magnitude of numbers. Figure 16 illustrates one such case. Considering that addition and magnitude comparison are simple operations for humans, this highlights the need for improvement in the arithmetic reasoning capabilities of MLLMs.

D.5 Company/Entity Confusion

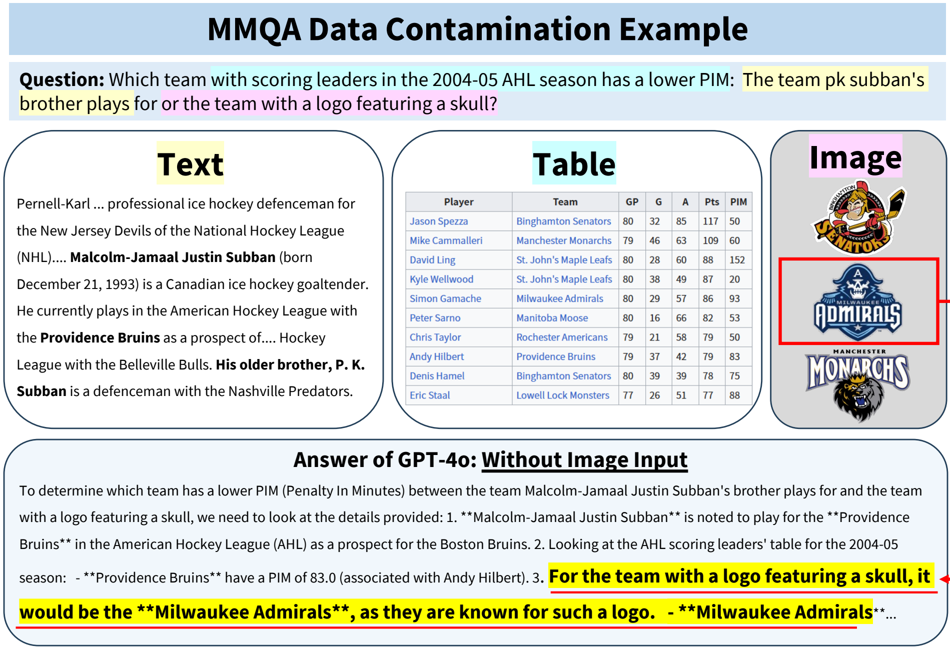
Errors in this category involve mixing up one company or entity with another. Even when companies are distinguished by unique colors, labels, or legends, the model may incorrectly assign data from one company to a different one, thus undermining the validity of its reasoning and final answers. An example is in Figure 17.

D.6 Information Extraction & Reasoning Failure

There are the cases incorrectly extracting facts, misunderstanding textual information, drawing unjustified conclusions, or logical missteps after gathering correct details. An example is in Figure 18.

D.7 Problem Understanding & Condition Ignoring

Claude sometimes makes incorrect judgments by considering only a subset of the required conditions. This issue is particularly prominent in Hard-level tasks that require deep reasoning. An example of this case is in Figure 19.



Player	Team	GP	G	A	Pts	PIM
Jason Spezza	Binghamton Senators	80	32	85	117	50
Mike Cammalleri	Manchester Monarchs	79	46	63	109	60
David Ling	St. John's Maple Leafs	80	28	60	88	152
Kyle Wellwood	St. John's Maple Leafs	80	38	49	87	20
Simon Gamache	Milwaukee Admirals	80	29	57	86	93
Peter Sarno	Manitoba Moose	80	16	66	82	53
Chris Taylor	Rochester Americans	79	21	58	79	50
Andy Hilbert	Providence Bruins	79	37	42	79	83
Denis Hamel	Binghamton Senators	80	39	39	78	75
Eric Staal	Lowell Lock Monsters	77	26	51	77	88

Figure 9: Example of data contamination in MMQA. Despite the requirement to infer the pink-highlighted information, *"team with a logo featuring a skull"*, using image data, GPT-4o deduced the answer, *"Milwaukee Admirals"*, without being provided with the image. This suggests the model leveraged internal knowledge, rather than cross-modal reasoning, highlighting a limitation in evaluating the cross-modal multi-hop reasoning capabilities of MMQA.

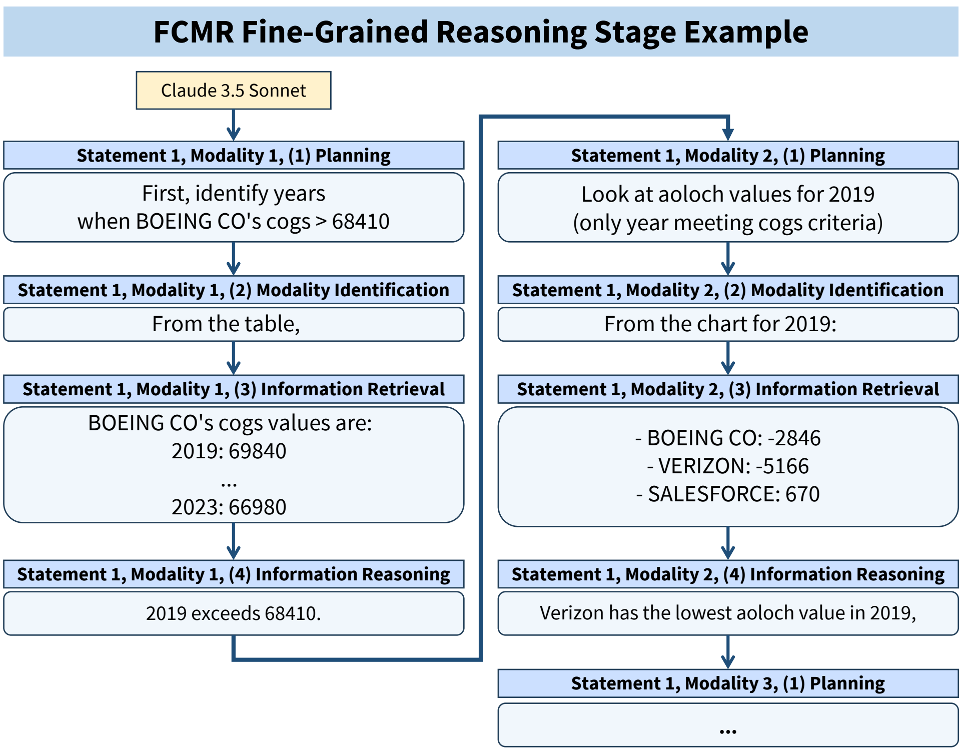


Figure 10: An example of decomposing the reasoning process of the Claude 3.5 Sonnet's response into fine-grained, stage-based steps.

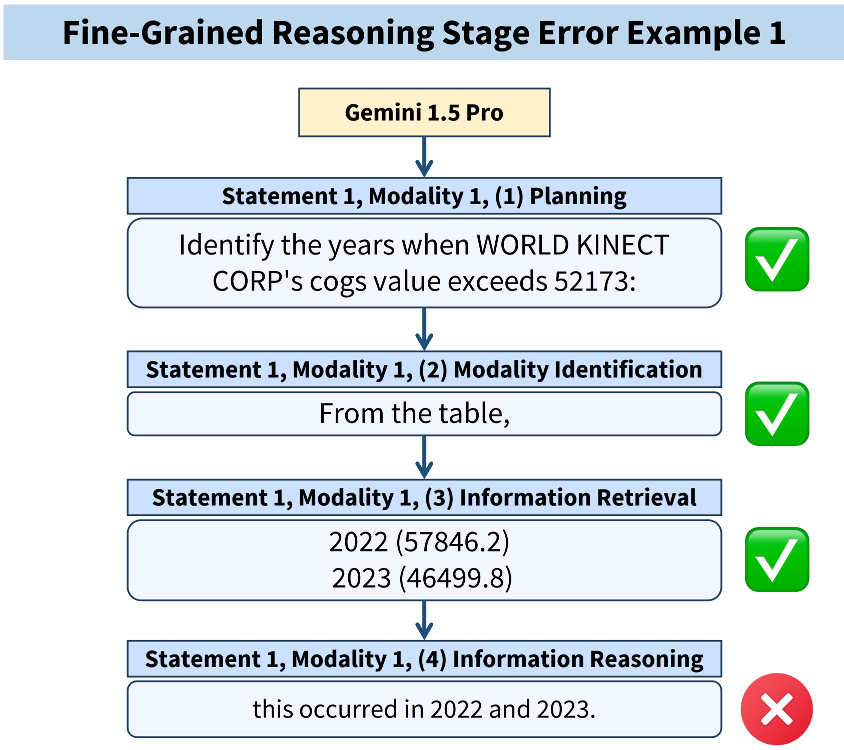


Figure 11: An example where the Gemini 1.5 Pro succeeds in (3) Information Retrieval but fails in (4) Information Reasoning. While the model successfully extracts the information that the cogs value is 57846.2 for 2022 and 46499.8 for 2023 from the table, it incorrectly reasons that the cogs values for both 2022 and 2023 are greater than 52173, resulting in a failure.

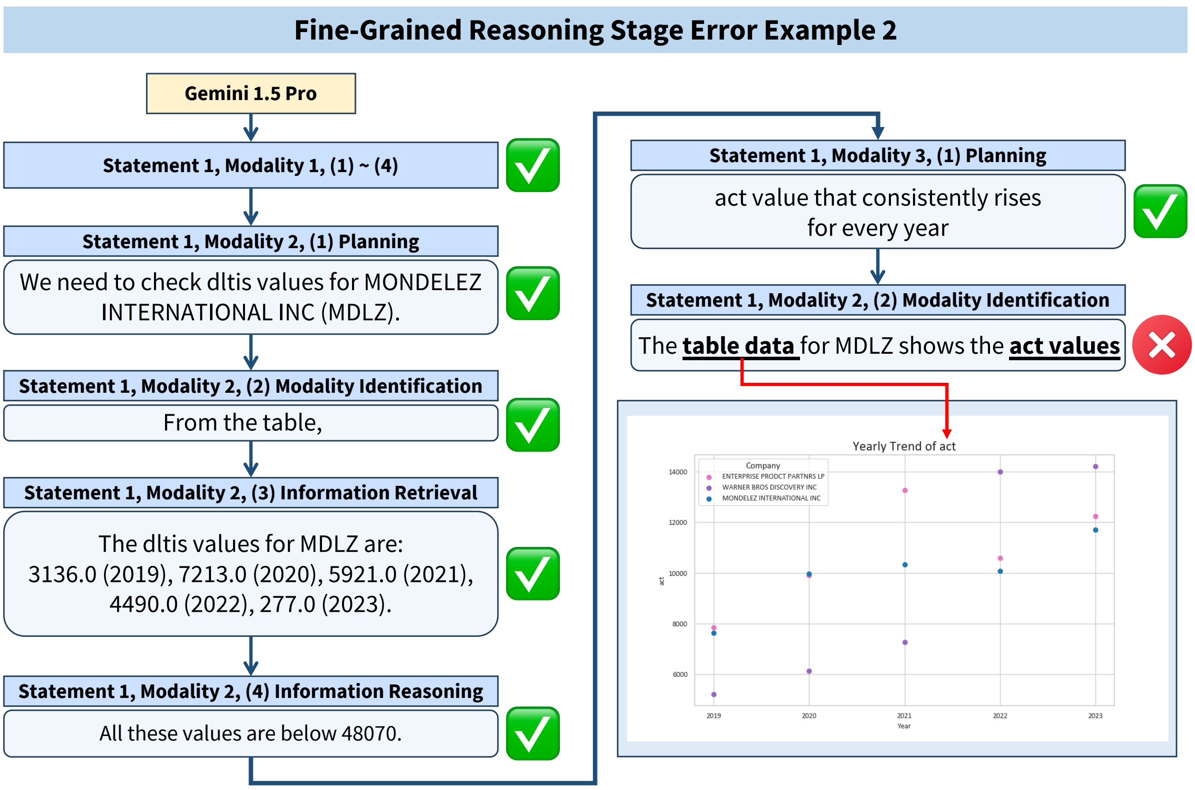
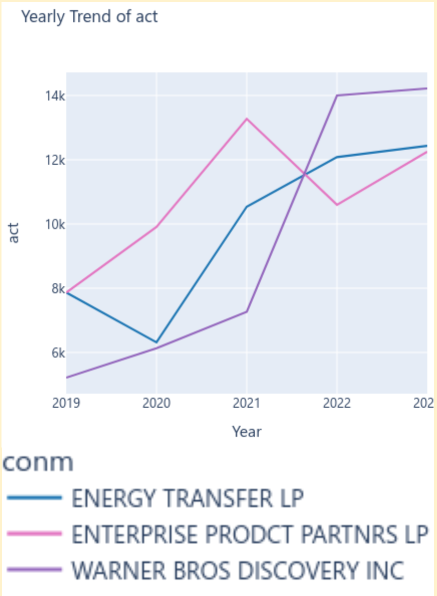


Figure 12: An example where the Gemini 1.5 Pro succeeds in (1) Planning but fails in (2) Modality Identification. While the model successfully plans that the act value is needed, it fails by identifying the modality as Table instead of Chart, where the act value is actually presented.

Trend Assessment Error

Chart (act)



Table

conm	year	dltr	oiadp	aoloch
WBD	2019	2939	3187	-3088
...	...	...	...	...
WBD	2022	7740	115	-11244
WBD	2023	8210	1649	-12301

FCMR Statement

The company, where the dltr value is below 48984.5 and the **act value consistently rising throughout all the years**, 's largest single expense, associated with content including costs for content amortization, impairment, and ...

Response of Claude 3.5 Sonnet

Statement 3: Need to check dltr < 48984.5 and consistently rising act values. For dltr < 48984.5: WBD qualifies (all values well below threshold). Looking at WBD's act trend in chart: Shows steep increase after 2021, but **not consistent rise throughout**. Since the second condition fails, no need to verify expense claim Result: FALSE

Figure 13: In the chart, WBD shows a consistently increasing trend in act values across all years, but Claude 3.5 Sonnet fails to recognize this correctly, resulting in an error.

Ranking/Ordering Mistake

Chart (oibdp)

Table

conm	year	lct	fiao	ivncf
EPD	2019	9129.7	437.5	-4575.5
...	...	...	...	...
EPD	2022	12265	-244	-4954
EPD	2023	13131	-221	-3197

FCMR Statement

The Seminole NGL Pipeline of the company, with the **top oibdp value** in the years when the fiao value minus ivncf value for ENTERPRISE PRODCT PARTNRS LP is greater than 4861.5, transports NGLs from the Hobbs hub and the ...

Response of Claude 3.5 Sonnet

1. Need to find years when fiao - ivncf > 4861.5 for EPD
2. Calculating fiao - ivncf for EPD: 2019: 437.5 - (-4575.5) = 5013, 2020: -242.4 - (-3120.7) = 2878.3, ,2021: -93.7 - (-2134.6) = 2040.9, 2022: -244 - (-4954) = 4710
2023: -221 - (-3197) = 2976. Only 2019 exceeds 4861.5
3. From the chart, **EPD doesn't have the highest oibdp** in 2019 4. Therefore false

Figure 14: An example where the Claude model fails to interpret a cumulative graph: Although the company with the highest OIBDP value in 2019 is indeed EPD (Enterprise Product Partners LP), Claude incorrectly states otherwise.

Condition Satisfaction & Selection Errors

Table

conm	year	dpc
US Foods	2019	369
...	...	...
Mondelez	2023	1215

Chart

FCMR Statement

The company, with a dpc value below 1953.5 for every year and a sale value that shows a continual rise for every year, computes constant currency operating results to evaluate the effect of currency exchange rates on financial...

Response of Claude 3.5 Sonnet

- Let's identify the company with dpc below 1953.5 for every year and continually rising sales
- Looking at dpc values in the table:
  * Abbott: 3014-3243
  * US Foods: 369-412
  * Mondelez: 1047-1215
- US Foods has dpc below 1953.5 all years
- Looking at US Foods sales from the chart:
  * 2019: 25939
  * 2020: 22885 (decrease)
  * 2021: 29487
  * 2022: 34057
  * 2023: 35597
- Sales don't show continual rise (dropped in 2020)

FALSE

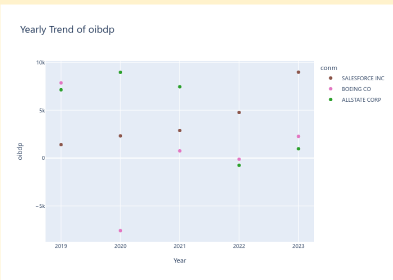
Figure 15: The condition that the dpc value is less than 1953.5 for all years is satisfied by both US Foods and Mondelez, but Claude 3.5 Sonnet recognizes only US Foods and fails to consider Mondelez, resulting in an error.

Data / Value Interpretation Errors

Table

conm	year	fincf
BOEING CO	2019	5739
BOEING CO	2020	34955
BOEING CO	2021	-5600
BOEING CO	2022	-1266
BOEING CO	2023	-5487
...	...	...

Chart



FCMR Statement

An extensive risk management approach for pension assets, which encompasses overseeing ... by the firm, with the smallest oibdp value during the years when the fincf value for BOEING CO falls below -5543.5.

Response of Claude 3.5 Sonnet

2. Analyzing:

- From Boeing's report: there is indeed an extensive risk management approach for pension assets that includes interest rate risk and market risk
- Looking at Boeing's fincf values: 5739, 34955, -5600, -1266, -5487
- The statement specifies "smallest oibdp value during years when fincf falls below -5543.5"
- Two years have fincf **below -5543.5**: 2021 (-5600) and **2023 (-5487)**

However, we don't have clear oibdp values in the data to verify this specific condition. The statement cannot be verified completely, so it's FALSE.

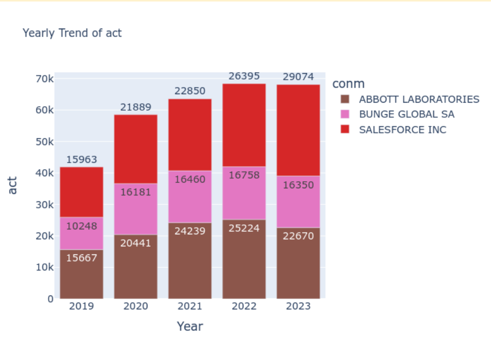
Figure 16: -5487 is not smaller than -5543.5, but Claude failed to correctly compare the two numbers and consequently reached the wrong conclusion.

Company / Entity Confusion

FCMR Statement

Reports are provided in U.S. dollars by the organization, possessing the largest act value in those years when the sum of the lco value and the spi value for SALESFORCE INC is below 11484, even though its overseas subsidiaries generally operate using their respective local currencies.

Chart



Response of Claude 3.5 Sonnet

First, let's find when SALESFORCE INC's (lco + spi) < 11484:
2019: 10662 - 158 = 10504 < 11484
2020: 12607 - 143 = 12464 > 11484
2021: 15628 - 54 = 15574 > 11484
2022: 17376 - 828 = 16548 > 11484
2023: 19003 - 988 = 18015 > 11484
So only in 2019 does it fall below 11484.
From the chart, in 2019:
ABBOTT LABORATORIES: 15667
BUNGE GLOBAL SA: 10248
SALESFORCE INC: 15963
SALESFORCE INC had the largest act value in 2019. From Bunge's report, ...

Figure 17: Although charts and tables were successfully interpreted to determine that Salesforce had the highest act value in 2019, the company was ultimately confused in the final step by Claude, resulting in an incorrect conclusion.

Information Extraction & Reasoning Failures

FCMR Statement

The organization, with the minimal aoloch value during the years when the ao value minus the ivncf value for VERIZON COMMUNICATIONS INC is below 25063.5, anticipates obtaining approximately $169 million in sublease revenue over the coming five years.

Text

The return on our investment portfolios is an important component of our ability to offer good value to customers, fund business improvements and create value for shareholders. ...
The Protection Services portfolio is focused on protection of principal and consistent income generation, within a total return framework. The portfolio is largely comprised of fixed income securities with a lesser allocation to equity securities and short-term investments.

Response of Claude 3.5 Sonnet

3. To verify this:
a) First identify years when (ao - ivncf) for VERIZON < 25063.5
2019: 3875 - (-17581) = 21456
...
2023: 5239 - (-23432) = 28671

Then find minimum aoloch value during qualifying years (when < 25063.5):
Only 2019 qualifies (21456 < 25063.5)
The company with the minimum aoloch in 2019 is mentioned in SALESFORCE's report regarding the $169 million sublease income.
This statement appears to be TRUE.

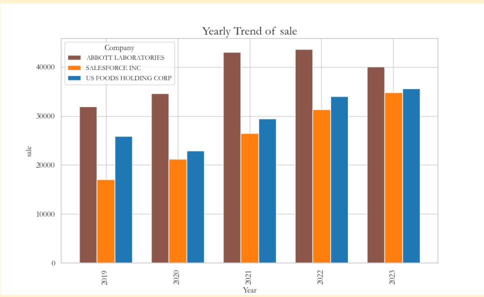
Figure 18: To compare the aloloch value, the chart should have been considered. Instead, having relied solely on the text data, Claude led to an incorrect conclusion.

Problem Understanding & Condition Ignoring

FCMR Statement

The company, with the lo value below 3493.5 and a sale value that continuously rose for every year, gained an advantage in its attrition rate, by shifting towards enterprise and international markets with extended contract durations.

Chart



Response of Claude 3.5 Sonnet

Statement 2:
- Need to identify which company this refers to by checking lo values below 3493.5
- Both Salesforce and US Foods have lo values below 3493.5
- Looking at sales trends in the chart:
  - Salesforce shows consistent rise
  - US Foods shows fluctuations
- Only Salesforce's text mentions attrition rate benefits from enterprise/international markets
- Cannot definitively match this statement to a single company, making it ambiguous and therefore false

Figure 19: Although Claude found that Salesforce satisfied all the conditions, it misunderstood the statement and consequently made an incorrect judgment.

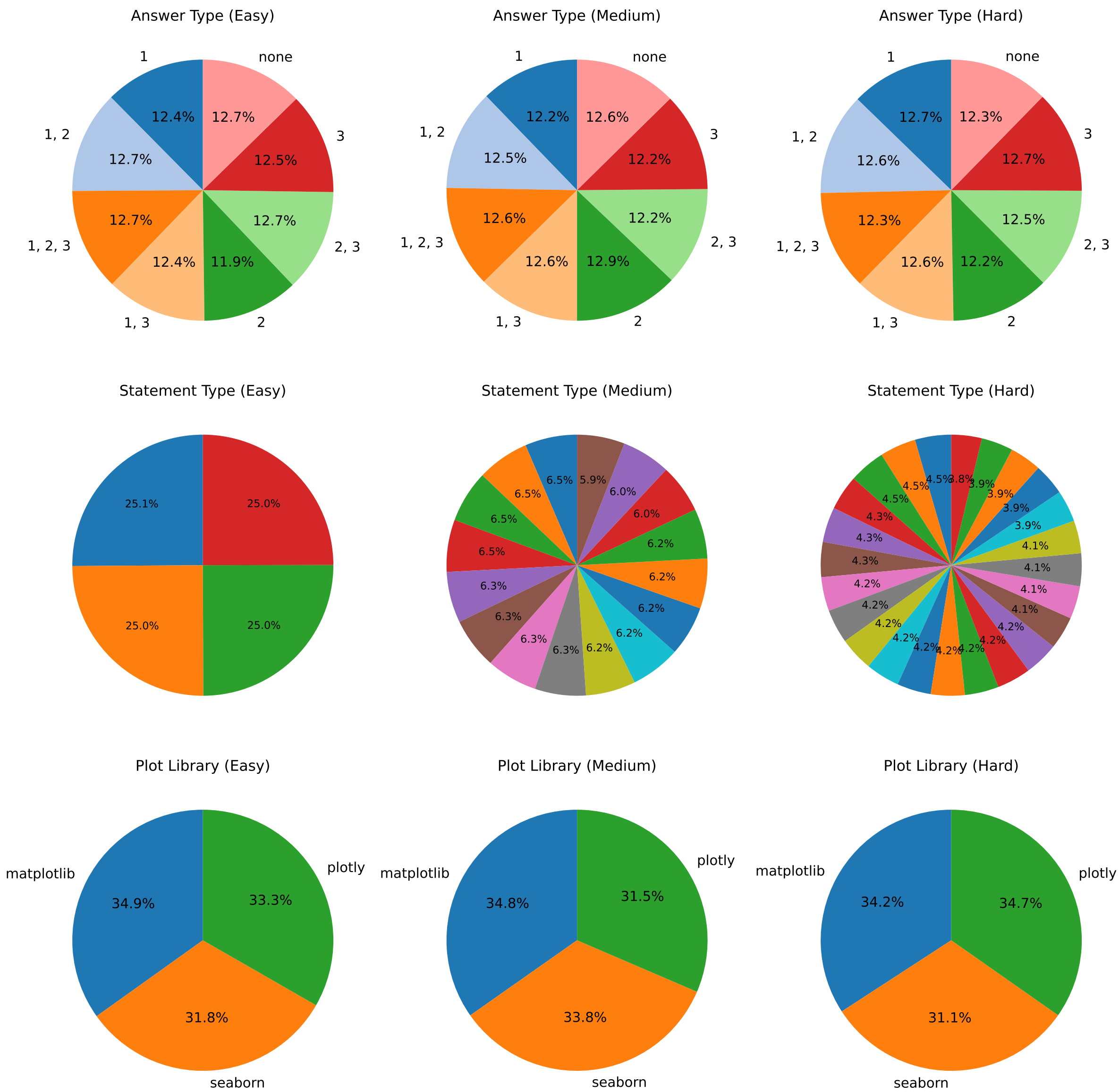


Figure 20: Pie charts for answer types, statement types, and library usage categorized by difficulty levels.

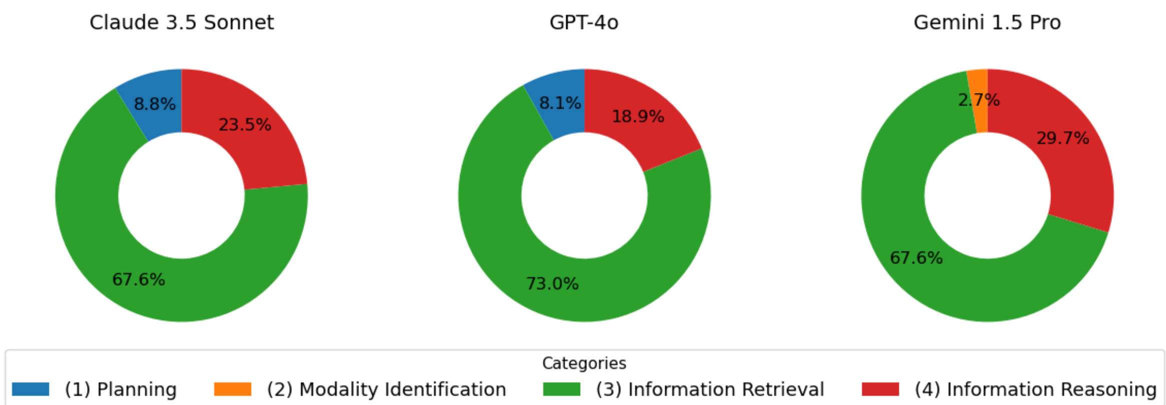


Figure 21: Reasons for Inference Failures by Model Across Fine-Grained Reasoning Stages

Statement 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	Professional services are provided by the organization, which has the nopi value below 2796 for all years and the aoloch value that consistently declines for all years, to help clients with digital transformations using Salesforce solutions.
FC+CT+RK	The firm, with the highest act value during the years when ENERGY TRANSFER LP's intan value dips below 8059, acquired a controlling interest in USAC through a $250 million cash transaction.
FC+AR+TR	The organization, with the cumulative sum of nopi values below 2840 and continuously increasing act values for every year, has provided put rights to certain consolidated subsidiaries. These put rights are omitted from the contractual obligations table due to unpredictability in payment.
FC+AR+RK	The business, with the minimal aoloch value in the years when the ivao value minus the ao value for VERIZON COMMUNICATIONS INC exceeds 9312.5, provides expert services to support customers in executing digital transformations leveraging Salesforce solutions.

Table 8: Examples for each statement type. FC refers to Fact-Checking, CT refers to Conditional Threshold, AR refers to Arithmetic, TR refers to Trend, and RK refers to Ranking.

Statement 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 9: Base templates of statement 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 Statement Types that include FC, a new template is generated by combining them with other Statement Types and Facts, where they share a common company entity.

Difficulty	Modality Types	Statement Types
Easy	Text Table Chart	Fact-Checking Conditional Threshold Arithmetic Trend Ranking
Medium	Text + Table Text + Chart Table + Chart	Fact-Checking + Conditional Threshold Fact-Checking + Arithmetic Fact-Checking + Trend Fact-Checking + Ranking Conditional Threshold + Trend 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 10: Detailed Statement Types by Difficulty and Modality Types. For the Easy level, all three answer statements are single-modal one-hop, while for the Medium level, all three statements are cross-modal two-hop. At the Hard level, all three statements consist of cross-modal three-hop. Even if each answer statement is a one-hop, the overall question remains a cross-modal three-hop QA. Specific examples can be found in Table 8

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 22: MLLMs zero-shot prompt.

You are provided with the following materials:
[Text Reports: Detailed excerpts from company reports of three companies]
{text}
[Table Data 1: A table containing financial data for the same three companies]
{table}
[Table Data 2: A table containing financial data for the same three companies]
{chart_to_table}

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 from the text reports, table data.
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 23: (M)LLMs + Deplot zero-shot prompt.

Identifying Information	Income Statement Variables
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)	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 24: Description of each column in the Annual Simplified Financial Statement.