Systematic Multi-Aspect Evaluation of Time Series-Based Report Generation: The Case of Financial Analysis from Stock Data

Elizabeth Fons[1*], **Elena Kochkina**[1*], **Rachneet Kaur**[1], **Zhen Zeng**[1], **Berowne Hlavaty**[2], **Charese Smiley**[1], **Svitlana Vyetrenko**[1†], **Manuela Veloso**[1†]
[1]J.P. Morgan AI Research,
[2]J.P. Morgan Chase
[*]Equal Contribution, [†]Work done while at J.P. Morgan AI Research
{name}.{surname}@jpmorgan.com

Abstract

This paper explores the capability of large language models (LLMs) to generate coherent textual reports from time series data, using financial reports from stock data as the use case. We conduct a comprehensive multi-aspect evaluation across four model families, including linguistic quality, content source attribution, automated metrics, and expert human assessment. We evaluate models using four major stock indices and two synthetic time series to assess generalization. We assess reports based on single and multiple time series data, and experiment with plain text and multi-modal prompting. We examine temporal effects by analyzing report quality as data approaches model knowledge cutoffs and testing synthetic future intervals. Our evaluation shows that LLMs are capable of creating high-quality financial analyst reports, with larger models demonstrating superior performance, however even those require human oversight and have potential for temporal logic errors. Our findings reveal model-specific behavioral patterns that enable tailored generation pipelines and inform future research about model pitfalls in time series-to-text generation tasks.

Keywords: Time series, Large Language Models, Report Generation, Report Evaluation

1. Introduction

The generation of coherent textual reports from time series is a fundamental challenge in natural language processing, requiring models to interpret temporal patterns, extract meaningful insights, and communicate findings in domain-appropriate language. While, recent studies show the proficiency of large language models (LLMs) (Abdin et al., 2024; OpenAI, 2023; Team et al., 2023; Touvron et al., 2023) in processing financial texts and time series data (Fons et al., 2024), the task of generating comprehensive reports from time series with LLMs is largely unexplored.

This paper systematically evaluates LLM-based time series-to-text generation, focusing on financial analysis from stock data (Figure 1). Unlike domains such as healthcare, financial time series reporting lacks publicly available gold standard reports and established evaluation benchmarks, making it a novel and challenging testbed for LLMs. We assess two report types: short, high-level summaries based on single time series, and longer, detailed reports incorporating multiple time series of technical indicators. This allows us to evaluate model performance across varying analytical demands. Our multi-aspect evaluation spans four model families and uses data from four major stock indices—S&P 500, Nasdaq, Dow Jones Industrial Average, and Nikkei 225 (FRED, 2024)—as well as synthetic time series to test model generalization and temporal reasoning beyond the scope of real-world data.

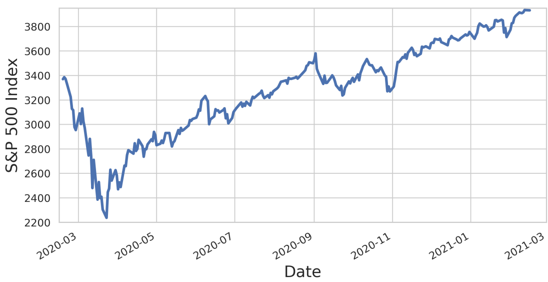
We combine automated metrics, expert human assessment, linguistic analysis, and content source attribution to achieve high quality evaluations in the absence of ground truth. We also examine temporal effects by analyzing report quality near model knowledge cutoffs and on synthetic future intervals, evaluating generalization across both real and synthetic data. We make the following contributions:

- We investigate financial time series report generation as an underexplored Natural Language Generation task in the space of LLMs, providing baseline evaluations and methodological insights across both short single-series and longer multi-indicator formats.
- We provide comparative analysis across four model families, identifying model-specific patterns and common pitfalls in financial data interpretation.
- We systematically assess linguistic quality, content source attribution, automated metrics, and expert human evaluation, tailored for financial time series reporting without ground truth.
- We analyze how model performance varies near training data cutoffs and on synthetic future intervals, revealing insights into temporal reasoning.
- We propose categorizing report segments by information source (direct reference, financial interpretation, external knowledge), enabling analysis of model reasoning and factual grounding.

2. Related Work

Automated Report Generation While generative tasks based on textual data is an active research

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3833–3850
11–16 May 2026.



Short report: The period from February 2020 to February 2021 marks a tumultuous yet ultimately bullish trend for the S&P 500. At the beginning of this interval, the index was at a relatively high point, reaching levels just under 3400. However, with the onset of the COVID-19 pandemic, the index experienced a significant and rapid decline, hitting a low of 2237.4 on March 23, 2020. This represented a precipitous drop of approximately 33% from its high observed in February 2020. This sharp decline highlights a period of intense volatility, reflecting market uncertainty as investors grappled with the global implications of the pandemic. Following the March low, the S&P 500 embarked on a robust recovery, seemingly buoyed by a mix of fiscal stimulus, accommodative monetary policy, and optimism around economic recovery.

Figure 1: Example of time series and short report generated based on it, with highlights indicating **Direct Reference (blue)**, **Financial Interpretation (green)**, and **External Knowledge (red)** using our proposed highlighting system.

area (Nishida et al., 2023; Assis et al., 2024; Liu et al., 2024), report generation based on time series data remains underexplored. Liu et al. (2024) address this by generating professional data analysis plans from chart data in Chinese, which are then manually evaluated by experts. However, their task differs from ours as it focuses on creating analysis plans rather than reports based on chart analysis. Several studies employ LLMs for stock market predictions. For example, Yang et al. (2023) introduced FinGPT Forecaster, which uses market news and basic financial data to predict stock price movements and provide analysis summaries. This approach, however, still relies mostly on textual data. Similarly, Li et al. (2024) use LLMs for stock trend prediction and generate reports to justify these predictions, with human annotators evaluating the model's outputs. In our work, we focus on using time series data to leverage LLMs for generating informative reports for human users, rather than making decisions or predictions. We also explore evaluation strategies in the absence of gold standard answers.

Generation Evaluation Evaluating the quality of generations is a critical aspect of natural language generation (NLG) research. The majority of works on report generation (Messina et al., 2022; Nishida et al., 2023; Sloan et al., 2024; Assis et al., 2024) apply traditional evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics, while well-established and easy to compute, require access to ground truth answers, which are not available to us in this work. Moreover, they may also lack the sensitivity to account for the specific requirements of financial reporting, such as factual accuracy and domain-specific terminology. Human evaluation is another major approach (Messina et al., 2022; Li et al., 2024; Chiang et al., 2024) involving human judges who assess various aspects of the generated reports, such as fluency, factual accuracy, and grammatical correctness. While this is considered the gold standard, it is time-consuming and expensive, requiring significant human resources. Due to this, here we provide human-evaluation on a subset of the data. Recently, LLM-based evaluation methods have been proposed as a suitable alternative in the absence of ground truth (Gao et al., 2024). Liu et al. (2023) introduced G-eval, an LLM-based metric that aligns well with human judgment. This approach, adapted in this work, provides a scalable and consistent evaluation compared to traditional approaches.

LLMs in Finance. The use of LLMs in Finance is a growing area of research, aiding automation of a wide range of tasks such as text classification tasks, time series, financial reasoning, and agent-based modelling (Nie et al., 2024; Li et al., 2023). For example, Callanan et al. (2023) showed that GPT-4 is likely to pass a professional CFA exam. Aguda et al. (2024) demonstrated that LLMs can be used for financial data annotation outperforming untrained crowd-workers. BloombergGPT (Wu et al., 2023) enabled a finance-specialized LLM by training an LLM on extensive financial data.

With the increased interest in their multimodal capabilities, LLMs are rapidly being applied to time series analysis, particularly in finance. Recent works explored using LLMs for forecasting (Xue and Salim, 2023; Yu et al., 2023), imputation and classification (Zhou et al., 2023). Kawarada et al. (2024) prompted LLMs with time-series data to obtain market comments, while Fons et al. (2024) evaluate LLM understanding on time series across a taxonomy of time series features.

3. Methodology

We outline our methodology for data preparation, report generation, analysis and evaluation (Figure 2). This methodology can be re-purposed as a generation framework with automated feedback, for model comparison, prompt tuning, and re-generation if evaluation metrics fall outside acceptable ranges.

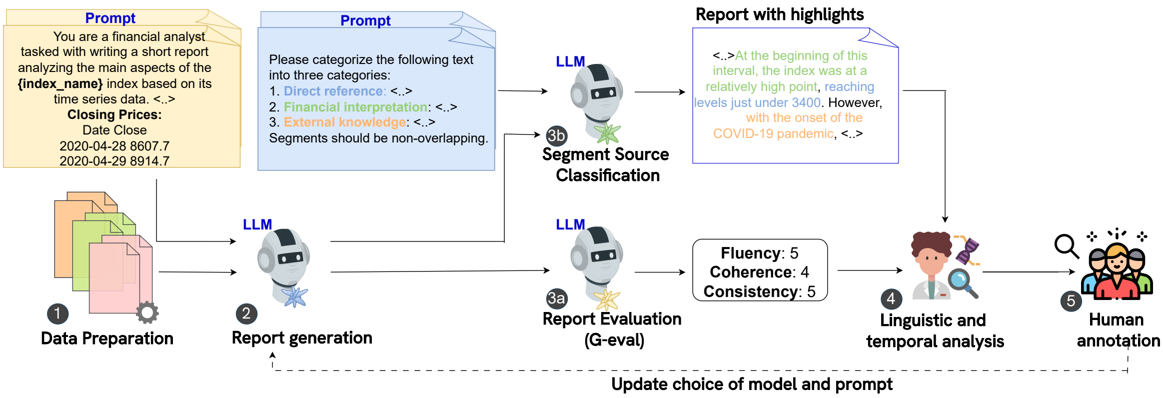


Figure 2: Methodology used for report generation, evaluation and analysis

3.1. Data Preparation

We experiment with two types of data: real and synthetic. Synthetic time series allow us to test model generalization beyond the scope of real-world data.

Real indices We compile a comprehensive dataset spanning five years (2019–2024), encompassing the S&P 500, Nasdaq, Dow Jones Industrial Average, and Nikkei 225 indices. The data is divided into overlapping one-year windows with a one-month stride, facilitating the analysis of both short-term and longer-term trends. This resulted in a total of 196 time series (49 time windows for 4 indices). For each one-year window, we compute a set of standard technical indicators, including the 50-day Simple Moving Average (SMA), 50-day Exponential Moving Average (EMA), Volatility rolling window, Relative Strength Index (RSI) (Murphy and Murphy, 1999), Moving Average Convergence Divergence (MACD) (Appel, 2005), Bollinger Bands (Murphy and Murphy, 1999) and Fibonacci Retracement levels (Malkiel, 1973).

Synthetic indices We generate synthetic time series data using Geometric Brownian Motion that simulate five years of daily stock prices for two distinct periods: 2019-2024, aligning with the period of real-world data, and 2024-2029, extending beyond the LLMs' training data. This results in 90 time series (45 time windows for 2 indices). This entirely unseen data enables a robust evaluation of the models' generalization and time series processing capabilities.

3.2. Report Generation

For each financial time series (196 real and 90 synthetic) that contain the close price, and/or technical indicators, the LLM is prompted to describe and summarize the patterns and trends observed in the data in an analyst report style. We generate two types of reports: (1) **short**, a 1-paragraph high level summary of the main trends given the close price over the observed period, and (2) **technical indicator reports (TI)**, 2-3 paragraphs covering the main patterns as well as conclusions made from analyzing the close price and technical indicators provided.

Generation quality and properties depend on the choice of the model and prompt, which can be determined through the iterations that incorporate feedback from automated and/or human evaluation.

Prompt Engineering: We developed two main types of prompts tailored to the short and TI reports. The prompt for generating short reports includes a time series of daily close price and emphasizes a high-level overview of the time series data, guiding the model to focus on key trends, major price movements, and volatility. For TI reports, the prompt is designed to incorporate time series data for the close price and technical indicators. We explore two methods for integrating data into the prompt: (1) text-based time series and (2) data plot images. The resulting report should provide insights into how these indicators signal trends or shifts in the index's performance, offering a more technical perspective on the market movements. The prompts can be found in Appendix B.

Models: We evaluate our framework with GPT-4o, GPT-4o-mini (OpenAI, 2023), Gemini (Team et al., 2023), LLama3.2-11B-Vision-Instruct (Dubey et al., 2024) and Phi-3-3B (Abdin et al., 2024). These models are chosen for their multimodal abilities, enabling both textual and visual time series representations. The selection offers diversity in model size and family. We use temperature 0 across all models and experiments[1].

3.3. Report Evaluation

We perform a comprehensive evaluation of report generation that combines fast, cost-effective automated metrics with in-depth linguistic analysis and

[1]Reports are available for research purposes upon request.

selective human review, which is more resource-intensive. This approach allows us to assess report quality and typical characteristics, identify which features are desirable and which need improvement— across all reports or for specific cases.

3.3.1. Automated Evaluation

We adopt the G-Eval (Liu et al., 2023) framework for automated evaluation. Originally, G-Eval assesses summaries against source texts. We adapt it for financial report generation from time series data, using the data itself as the “source". LLM evaluator receives a prompt with task and evaluation criteria definitions, and scores the report on a 1-5 scale on three key criteria below.

- **Consistency**: ensures factual accuracy by comparing the report to the original time series data, penalizing discrepancies.
- **Coherence**: assesses logical flow and clarity, focusing on structure and transitions.
- **Fluency**: evaluates grammar, clarity, readability.

We use two different models (GPT-4o and Gemini) as LLM evaluators to assess and mitigate potential bias from LLMs evaluating their own outputs. The prompts can be found in Appendix B.

3.3.2. Segment Source Classification

We categorize report segments by the type of information or reasoning they contain. Segment categorization is used to show the distribution of statement types in reports, to flag unsupported external references, and to guide annotators on required data for verification. Figure 1 presents an example of a highlighted report. To achieve this, we split each report into sentences and use an LLM (GPT-4o) with a tailored prompt (Appendix B.) to assign each sentence to one of three categories:

- **Direct Reference (DR):** segments that directly cite specific data points or trends in the time series, like index values, dates, or changes.
- **Financial Interpretation (FI):** segments offering analysis or inference on financial data without relying on external input, such as market trends or fluctuation explanations.
- **External Knowledge (EK):** segments referencing information outside the time series, e.g. economic factors, geopolitical or industry events.

To estimate the accuracy of segmentation and segment categorization, we sample a set of 52 reports for manual evaluation with one expert annotator, who assigned a correct label to each segment. The annotator also indicates if the given segment should be split into two or more categories as it contains several labels.

3.3.3. Linguistic analysis

We analyze linguistic properties of the generated financial reports. These are common features that help evaluate the generations by identifying relevant linguistic phenomena, allowing us to detect anomalies like excessive sentiment or subjectivity, which is crucial for maintaining professionalism in formal reports. In our linguistic analysis we included (1) report and sentence length; (2) lexical diversity within and across reports calculated using Token-to-Type Ratio; (3) sentiment and degree of subjectivity conducted using TextBlob[2]; (4) readability score evaluated using the Flesch Reading Ease score[3]; (5) TI mentions; and (6) the use of hedging language. For the latter we define a lexicon of hedging words that indicate speaker uncertainty, such as "potentially" and "possibly". We track the usage of these words in reports over time, with each word contributing a -1 to the count. The average count of hedge words is calculated across indices.

3.3.4. Human Annotation

To verify the reliability of our automated evaluation approach, we conduct human evaluation of a sample of reports, limited by the cost of financial experts. Human evaluation follows G-Eval’s dimensions, with annotators scoring each report from 1 to 5 for consistency, coherence, and fluency. Each report is assessed by 3 annotators. For the real data reports, we focus on the S&P 500 index, sampling 4 reports per model and report/prompt type. For the synthetic data reports, we evaluate reports based on both of the indices, sampling 2 reports per model and report/prompt type. We then assess how annotator scores align with those assigned by G-eval.

4. Results and Analysis

4.1. Report Quality Evaluation

Table 1 presents the G-Eval (using GPT-4o) and Human Evaluation scores for all models (GPT-4o, GPT-4-mini, Gemini, Llama3.2 , and Phi-3) across all report types (Short, TI, and TI plots), using both real and synthetic indices.

4.1.1. Reports from real time series

GPT-4o consistently generated the highest quality reports across all report types and evaluation metrics (both human and G-Eval). This model demonstrated superior ability to analyze time series data

[2]The polarity score is a float in the range [-1.0, 1.0]. The subjectivity is a float in the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

[3]https://pypi.org/project/textstat/

		Real Data						Synthetic Data					
		G-Eval			*Human scores*			*G-Eval*			*Human scores*		
Report type	**Model**	**Con**	**Coh**	**Flu**	**Con**	**Coh**	**Flu**	**Con**	**Coh**	**Flu**	**Con**	**Coh**	**Flu**
Short	GPT-4o	**3.8 ± 0.5**	**4.1 ± 0.2**	**5.0 ± 0.1**	3.8 ± 0.4	**4.3 ± 0.5**	**5.0 ± 0.0**	**3.7 ± 0.4**	**4.1 ± 0.2**	**4.9 ± 0.1**	4.3 ± 0.5	**4.8 ± 0.3**	**5.0 ± 0.0**
	GPT-4-mini	3.5 ± 0.5	4.0 ± 0.2	5.0 ± 0.1	**4.0 ± 0.9**	4.1 ± 0.6	4.9 ± 0.2	3.4 ± 0.4	3.9 ± 0.2	**4.9 ± 0.1**	4.0 ± 0.9	4.6 ± 0.6	**5.0 ± 0.0**
	Gemini	3.4 ± 0.4	3.9 ± 0.2	**5.0 ± 0.1**	3.5 ± 0.8	4.2 ± 0.2	4.7 ± 0.2	3.3 ± 0.5	3.9 ± 0.2	**4.9 ± 0.1**	**4.8 ± 0.2**	4.6 ± 0.3	**5.0 ± 0.0**
	Phi-3	2.7 ± 0.4	2.9 ± 0.4	4.5 ± 0.4	2.5 ± 0.4	2.5 ± 0.2	3.7 ± 0.3	2.3 ± 0.4	2.6 ± 0.5	4.3 ± 0.6	2.0 ± 0.0	2.1 ± 0.2	**5.0 ± 0.0**
	Llama3.2	2.6 ± 0.6	3.3 ± 0.6	4.7 ± 0.6	3.7 ± 0.4	4.5 ± 0.3	4.9 ± 0.1	2.7 ± 0.5	3.4 ± 0.4	4.8 ± 0.5	3.3 ± 0.5	3.2 ± 0.3	4.8 ± 0.2
TI	GPT-4o	**3.5 ± 0.4**	**4.0 ± 0.1**	4.9 ± 0.1	4.0 ± 0.7	**4.6 ± 0.5**	**5.0 ± 0.0**	**3.4 ± 0.3**	**4.0 ± 0.1**	**5.0 ± 0.1**	**4.4 ± 0.9**	**4.8 ± 0.3**	**5.0 ± 0.0**
	GPT-4-mini	3.2 ± 0.5	3.9 ± 0.2	4.9 ± 0.0	**4.2 ± 0.2**	**4.6 ± 0.2**	4.8 ± 0.3	3.1 ± 0.4	3.9 ± 0.2	4.9 ± 0.1	4.2 ± 0.6	4.5 ± 0.2	**5.0 ± 0.0**
	Gemini	2.8 ± 0.4	3.9 ± 0.3	**5.0 ± 0.0**	3.0 ± 0.6	**4.6 ± 0.4**	4.9 ± 0.2	2.8 ± 0.4	3.7 ± 0.3	**5.0 ± 0.1**	4.2 ± 0.4	4.7 ± 0.0	**5.0 ± 0.0**
	Llama3.2	2.4 ± 0.5	3.2 ± 0.5	4.7 ± 0.5	3.4 ± 0.5	4.1 ± 0.5	4.2 ± 0.4	2.5 ± 0.4	3.3 ± 0.4	4.7 ± 0.6	3.2 ± 0.4	3.7 ± 0.6	4.6 ± 0.2
TI (plots)	GPT-4o	**3.5 ± 0.4**	**4.0 ± 0.1**	4.9 ± 0.1	3.8 ± 0.6	**4.9 ± 0.2**	4.9 ± 0.2	**3.4 ± 0.4**	**3.9 ± 0.2**	**5.0± 0.0**	4.2 ± 1.0	**4.8 ± 0.3**	**5.0± 0.0**
	GPT-4-mini	3.1 ± 0.4	3.9 ± 0.2	5.0 ± 0.0	3.9 ± 0.5	4.7 ± 0.3	4.9 ± 0.2	3.1 ± 0.4	3.8 ± 0.2	4.9 ± 0.2	**4.6 ± 0.4**	4.4 ± 0.2	**5.0 ± 0.0**
	Gemini	3.1 ± 0.4	3.9 ± 0.2	**5.0 ± 0.0**	3.7 ± 0.6	4.7 ± 0.3	**5.0 ± 0.0**	2.9 ± 0.4	3.7 ± 0.3	4.9 ± 0.1	4.0 ± 0.1	4.6 ± 0.2	**5.0 ± 0.0**
	Llama3.2	2.2 ± 0.5	2.8 ± 0.5	4.2 ± 0.7	**4.0 ± 0.6**	4.3 ± 0.5	4.4 ± 0.4	2.2 ± 0.5	2.8 ± 0.5	4.2 ± 0.7	3.6 ± 0.5	3.3 ± 1.0	4.2 ± 0.6

Table 1: Comparative analysis of report quality (for both real and synthetic data reports) across diverse models and report types using G-Eval scores, which assess Consistency (Con), Coherence (Coh), and Fluency (Flu), alongside corresponding Human Evaluation scores from expert annotators. The best model for each report type is highlighted in **bold**.

and produce comprehensive, informative financial reports. Reports generated by Phi-3 often lacked in factual accuracy and coherence as evidenced by G-eval and manual evaluations. This suggests limitations in Phi-3's capacity to handle time series data. Models like GPT-4o-mini and Gemini generally performed well, with GPT-4o-mini often slightly outperforming Gemini in performance scores in Table 1. Llama3.2 also demonstrated strong performance, particularly in fluency, indicating its ability to generate readable and grammatically correct reports.

When generating short reports Phi-3 frequently produced repetitive prose and inaccurate descriptions of specific data points. As a result, it was excluded from further experiments. This highlights the need for models that can interpret time series data and generate insightful, cohesive reports, rather than just isolated descriptions.

For Technical Indicator (TI) reports, we compared results from two prompt formats: textual time series and visual plots. GPT-family models perform similarly across both formats. For , plot-based prompts yield better results on real data but sometimes lower scores on synthetic data. Llama3.2 performs best with purely textual prompts.

We observed context length limitations when generating TI reports. Multiple time series produce high input token counts, often near Llama3.2 's 8k-token limit, leaving little room for generation and causing incoherent or incomplete outputs. These cases were manually filtered to ensure fair evaluation. This underscores the need for larger context windows to handle complex numerical inputs.

4.1.2. Generalization to synthetic time series

Similar to the real data results, GPT-4o consistently achieves the highest scores in consistency and coherence, demonstrating its ability to generate high-quality reports even with unseen financial indices. However, absolute scores for both GPT-4o and GPT-4o-mini are generally lower with synthetic data, indicating a potential challenge in interpreting unfamiliar time series. This is not the case for Llama3.2 , whose scores improve on reports generated from synthetic time series.

Interestingly, the performance gap between GPT-4o and GPT-4o-mini is less pronounced with synthetic data, particularly for consistency. This suggests that the larger model's advantages may be less pronounced when analyzing unfamiliar financial patterns. Furthermore, Gemini exhibits a more noticeable performance drop with synthetic data, especially in consistency, implying a potential sensitivity to the nature of the input time series and a stronger reliance on real-world data for robust analysis. The lower overall scores and reduced inter-model differences may be attributed to the challenges of interpreting hypothetical future trends within the synthetic data, which requires a deeper understanding of market dynamics that may be harder to extract from simulated data.

4.1.3. G-eval Evaluator Model Choice

Prior studies have shown that LLM-based evaluations tend to favor generations from the same model family (Liu et al., 2023). To assess whether G-Eval is subject to evaluator bias, specifically, whether LLMs favor generations from their own model family, we conducted a control experiment using Gemini as the evaluator. Table 2 reports the average G-Eval scores for reports generated from real data, with Gemini as evaluator. Additionally, we provide correlation metrics (Spearman and Kendall-Tau) between the two evaluators.

The Gemini-based G-Eval scores broadly preserve the relative model rankings across all dimensions. We observe that Gemini assigns higher scores to GPT-4o-generated reports than to those generated by its own model for both coherence and

Report Type	Model	Con	Coh	Flu
Short	GPT-4o	**4.8 $\pm$ 0.3**	**4.7 $\pm$ 0.4**	**4.4 $\pm$ 0.6**
	GPT-4-mini	4.7 $\pm$ 0.4	4.3 $\pm$ 0.5	4.2 $\pm$ 0.6
	Gemini	4.7 $\pm$ 0.5	4.2 $\pm$ 0.5	3.5 $\pm$ 1.2
	Phi-3	3.0 $\pm$ 0.7	2.7 $\pm$ 0.6	4.0 $\pm$ 0.3
	Llama3.2	3.4 $\pm$ 0.9	3.2 $\pm$ 0.6	3.9 $\pm$ 0.6
TI	GPT-4o	**4.6 $\pm$ 0.4**	**4.7 $\pm$ 0.4**	4.1 $\pm$ 0.3
	GPT-4-mini	4.4 $\pm$ 0.5	4.3 $\pm$ 0.4	4.1 $\pm$ 0.3
	Gemini	4.2 $\pm$ 0.5	4.2 $\pm$ 0.5	**4.4 $\pm$ 0.5**
	Llama3.2	3.5 $\pm$ 0.8	3.5 $\pm$ 0.4	3.9 $\pm$ 0.6
TI (plots)	GPT-4o	**4.7 $\pm$ 0.4**	**4.6 $\pm$ 0.4**	4.3 $\pm$ 0.4
	GPT-4-mini	4.5 $\pm$ 0.5	4.2 $\pm$ 0.4	4.2 $\pm$ 0.4
	Gemini	4.4 $\pm$ 0.5	4.2 $\pm$ 0.4	**4.5 $\pm$ 0.6**
	Llama3.2	3.1 $\pm$ 0.9	3.0 $\pm$ 0.7	3.5 $\pm$ 0.9
Spearman (ρ)		0.93	0.95	0.64
Kendall-Tau (τ)		0.80	0.87	0.58

Table 2: G-Eval results using Gemini as evaluator for reports based on real data. The bottom two rows display correlations, capturing the alignment between G-Eval using Gemini and GPT-4o .

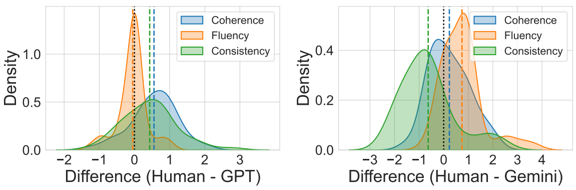


Figure 3: Density distributions of score differences between human and G-eval evaluations.

consistency. A similar, though less pronounced, pattern is observed with GPT-4o-mini, where Gemini typically assigns either slightly higher or comparable scores relative to its own outputs. These two dimensions exhibit strong alignment, with Spearman correlations of 0.95 (coherence) and 0.93 (consistency), and Kendall's τ of 0.87 and 0.80, respectively. This suggests that despite differences in absolute scoring levels, both evaluators produce highly similar rankings. This aligns with findings in Chen et al. (2025), who have shown that LLMs are less biased in fact-centric scenarios. In contrast, fluency shows weaker agreement, likely due to GPT-4o assigning uniformly high scores (often near 5) across models, reducing variance and limiting rank discrimination.

4.1.4. Alignment between G-eval and Human evaluation scores

To assess the reliability of the G-eval evaluation method, we examine its alignment with human scores, by analyzing their average differences and correlation.

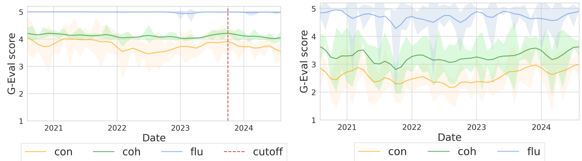
Score Differences We calculate the difference between each annotator's score and the G-eval score (from GPT or Gemini) for every report, then average these differences across annotators within each score category. Figure 3 presents kernel density estimates of these differences (Human minus G-eval): GPT on the left, Gemini on the right. Curves near zero indicate close agreement; positive values mean humans gave higher scores than G-eval. Dashed lines mark the mean.

For GPT, fluency scores closely match human ratings (mean: -0.05), while coherence (0.55) and consistency (0.43) show humans rate higher than GPT, which is preferable in this domain. Gemini's score differences are more spread out, but mean differences remain below 1 (Coh: 0.23; Con: -0.63; Flu: 0.74). The larger fluency gap is due to humans often assigning the highest score, while Gemini favored slightly lower ratings. These mean differences are also comparable to the standard deviations shown in Table 1.

Correlation Analysis Table 3 reports correlation metrics between the GPT-based G-eval and human evaluations. To account for systematic scale discrepancies, we apply isotonic regression separately for each dimension, learning a monotonic mapping between G-eval and human scores. This preserves rank order while aligning numeric scales more closely. We observe the strongest alignment for fluency, while coherence shows the weakest rank correlation, followed by consistency. This suggests that fluency is less subjective and easier to evaluate consistently, whereas coherence and consistency involve higher-level interpretive judgments.

	Con	Coh	Flu
Pearson	0.54	0.56	0.72
Spearman (ρ)	0.53	0.48	0.58
Kendall-Tau (τ)	0.41	0.38	0.50

Table 3: Spearman and Kendall-Tau, capturing the alignment between G-Eval (GPT) and Human Evaluation.



(a) GPT-4o (b) Llama3.2

Figure 4: The evolution of G-eval over time for short reports generated by GPT-4o and Llama3.2 .

4.1.5. Temporal analysis of report quality

We examined how G-Eval (GPT-4o) scores change over time. Figure 4 displays results for short reports generated by GPT-4o and Llama3.2 using real data.

GPT-4o scores remain stable over time, with only a slight decline in consistency after the models' knowledge cutoff date, indicating possible difficulty in interpreting newer data. In contrast, Llama3.2 scores show greater fluctuations throughout the entire period.

Report type	Model	Rep. len	Sent. len	TTR (W)	TTR (A)	Polarity	Subjectivity	Terms	Readability	MA	RSI	MACD	BB	Retracemt.
Short	GPT-4o	**226.6**	**27.6**	0.79	0.07	0.11	0.42	0.03	33.4	0.00	0.00	0.00	0.00	0.00
Short	GPT-4o -mini	204.7	25.8	0.80	0.06	0.11	0.42	0.03	35.8	0.01	0.00	0.00	0.00	0.00
Short	Gemini	150.1	24.7	**0.82**	0.07	0.13	0.47	0.04	38.4	0.01	0.00	0.00	0.00	0.00
Short	Phi-3	199.0	23.7	0.53	0.02	**0.24**	**0.52**	0.01	**62.5**	0.00	0.00	0.00	0.00	0.00
Short	Llama3.2	216.5	19.5	0.73	**0.10**	0.13	0.43	0.02	47.2	0.02	0.00	0.00	0.00	0.00
TI	GPT-4o	354.2	**27.2**	0.73	0.06	0.09	0.41	0.05	28.7	0.87	1.00	1.00	0.57	0.00
TI	GPT-4o -mini	328.7	27.0	**0.76**	0.06	0.07	0.42	0.06	26.6	0.99	1.00	0.96	0.20	0.00
TI	Gemini	291.7	22.8	0.70	0.05	0.09	**0.44**	0.06	35.6	0.93	1.00	1.00	0.52	0.00
TI	Llama3.2	**452.0**	11.3	0.60	**0.07**	**0.10**	0.42	0.04	**42.0**	0.91	0.94	0.87	0.21	0.00
TI (plots)	GPT-4o	**334.1**	**27.8**	0.75	0.06	0.09	0.41	0.04	28.8	1.00	1.00	0.02	0.00	0.00
TI (plots)	GPT-4o -mini	320.4	26.3	**0.76**	0.06	0.08	0.41	0.04	29.2	1.00	1.00	0.01	0.00	0.00
TI (plots)	Gemini	275.7	23.5	0.71	0.05	**0.10**	**0.44**	**0.07**	39.6	0.94	1.00	0.94	0.84	0.21
TI (plots)	Llama3.2	293.4	14.8	0.71	**0.12**	0.08	0.41	0.04	**45.2**	0.52	0.61	0.45	0.58	0.28

Table 4: Linguistic analysis of all report types: short, technical indicator (TI), technical indicator reports generated using time series and plots (TI (plots)) for real indexes. Table presents average report lengths (Rep. len), sentence lengths (Sent. len), Type-Token Ratio within each report (TTR (w)) and across reports (TTR (a)), sentiment polarity, report subjectivity, proportion of financial terms in the report, Flesch reading ease score (Readability), proportion of reports mentioning each of the technical indicators (MA, RSI, MACD, BB and Retracement).

4.2. Linguistic Analysis

Our findings for reports based on real indices are shown in Table 4.

Models have different interpretations of a paragraph length. While models were instructed to produce 1-2 paragraphs for short reports and 2-3 paragraphs for TI, we observe that models exhibited varying interpretations of “paragraph length”. GPT-4o and GPT-4o-mini generate the most verbose reports and Gemini producing the most concise. This difference in verbosity was also reflected in average sentence length. Llama3.2 is an outlier, producing lengthy reports with shorter sentences. Overall, this variation in report length among the models is reasonable and aligns with expectations.

Larger models produce reports with rich language, which remains reliably consistent across reports. To evaluate lexical diversity, we calculated two metrics: Type-Token Ratio for individual reports (TTR (W)), which measures the average proportion of unique words within each report, and Type-Token Ratio across all reports of a given type (TTR (A)), which measures the overall proportion of unique words. While individual reports, especially shorter ones, showed high TTR (W), the diversity across reports (TTR (A)) was consistently low for all models and report types, indicating a tendency towards formulaic, template-like language. This suggests that while each report benefits from rich language, consistency across various reports is prioritized for reliability. Notably, synthetic reports had higher TTR (A) than real reports, likely due to the distinct linguistic features of the two types of synthetic reports (“past” and “future”).

We use a financial terminology lexicon[4] to estimate the amount of terms used in the reports, as an average percentage of terms used out of the total number of words. We found it to be a rather low percentage, though higher for TI reports compared to short reports, as expected.

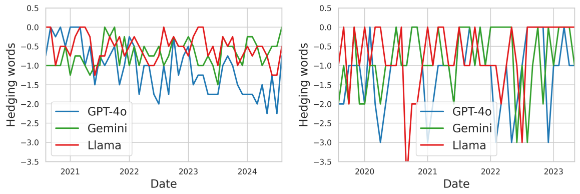
Most models produced neutral sentiment and subjectivity reports. Phi-3 was an exception, generating more positive and subjective reports.

Low readability scores reflect the complex and specialized nature of the content Shorter reports are generally more readable than TI. Short reports by Phi-3 are an exception, achieving a "standard" difficulty rating despite being repetitive and having the lowest quality scores in both human and automated evaluations. This shows the limitations of automated readability assessments. Synthetic reports had slightly higher readability compared to those based on real indices, which may indicate a simplification of the reports based on unseen data.

Most models fail to consistently include all of the TI provided in the reports. We examined the proportion of reports that mention the names of technical indicators provided in the prompt (Table 4, right side). Notably, the Moving Average (MA), Relative Strength Index (RSI), and MACD were consistently mentioned across all TI report types, indicating their widespread recognition and usage. We even observe a small percentage of short reports mentioning MA. Bollinger Bands were frequently referenced in TI reports but were less prevalent in reports with plots, demonstrating the influence of report format on indicator mention. Interestingly, Fibonacci retracement levels were scarcely mentioned, appearing only in Gemini and Llama TI reports with plots, highlighting better instruction following by these models. This distribution may reflect the commonality and perceived importance of these indicators in financial analysis, as well as their positioning in the prompt.

Hedging language increases near the knowledge cutoff date. Figure 5 illustrates the evolution of hedging words in short reports for GPT-4o, Gemini, and Llama3.2 . In subplot (a), based on real indices, we observe an increase in hedging words over time for GPT-4o. As references to external knowledge approach and surpass the knowl-

[4]https://www.iotafinance.com/en/Glossary-of-Financial-Terms.html



(a) Short reports (b) Synthetic short reports

Figure 5: The evolution of hedging words over time for Close Price reports generated by GPT-4o, Gemini and Llama3.2 .

edge cut-off date, uncertainty rises, suggesting the model may be extrapolating or speculating on trends. This pattern is seen in Llama3.2 but not in Gemini. For reports generated from synthetic data (subplot (b)), no clear trend is observed. These trends are consistent across other report types.

4.3. Error Analysis

In this section, we discuss key issues identified during human evaluation of report quality:

Inclusion of Non-Input Data: Aside from including confident references to external real-world events, which are not necessarily undesirable or incorrect, reports occasionally referenced technical indicators, such as the 200-day moving average, which were not part of the input data. This can be explained by the fact that this type of TI is very common for the long-term financial analysis.

Temporal Inconsistencies: Some reports described events in a sequence, which contradicts the natural flow of time. For example, it might mention a peak in mid-December 2023, followed by a downward trend in late September and early October 2023, which is temporally impossible. We observed occasional mentions of months not included in the input time series. The time series was sometimes split into intervals that were not meaningful, at points that were not significant changes.

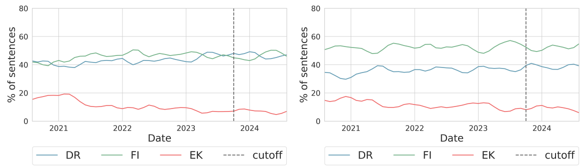
Inconsistent Report Content: The content varied across reports, suggesting a need for more prescriptive prompt design to ensure consistency.

Bias Towards Positive Reporting: There appeared to be a bias towards reporting positive or above-average trends, avoiding low points, This observation aligns with findings by Mantion et al. (2024), who noted a hawkish bias in ChatGPT when processing FedSpeak. However quantifying the extent of this is left for future work.

Use of Informal Language: Occasional use of slang was noted, such as the term "flirting" in the context of technical indicators: “The Relative Strength Index (RSI) occasionally flirted with overbought levels.”. Additionally, we provide highlighted report examples in Appendix C

	Pred. DR	Pred. FI	Pred. EK	P	R	F1	Supp.
DR	352	85	20	0.95	0.77	0.85	457
EK	12	261	20	0.30	0.46	0.36	37
FI	7	13	17	0.73	0.89	0.80	293
Acc.						0.80	787
Macro				0.66	0.71	0.67	787

Table 5: Segment categorization performance per class for Direct Reference (DR), Financial Interpretation (FI), and External Knowledge (EK): Confusion matrix (left), Precision (P), Recall (R), F1-score (F1) and Support (Supp.) (right), Accuracy and Macro-averages (bottom).



(a) GPT-4o short reports (b) GPT-4o TI (plots) reports

Figure 6: The evolution of information categories (DR, FI, EK) in financial reports generated by the GPT-4 model over time. The vertical dashed line indicates the model’s training data cutoff date.

4.4. Segment Source Classification Analysis and Evaluation

Evaluation Table 5 summarizes the confusion matrix and performance metrics for segment source classification. The method achieves an overall accuracy of 80%, indicating its reliability. EK is the hardest category, likely due to limited support. Most misclassifications occur when DR is predicted as FI, partly because FI is broadly defined and may overlap with DR, especially for segments summarizing trends or volatility. Segments hypothesizing about events with hedging language are often confused between FI and EK. Another frequent error is misclassifying Technical Indicator descriptions as FI, as the model does not consistently recognize TIs as part of the input “time series" referenced in the prompt. Only 12 out of 787 annotated segments required splitting into two categories. **Temporal Analysis** Figure 6 shows how segment category proportions in GPT-4o-generated reports change over time for two report types. EK segments are the least common, with only a few direct references to world events (e.g., COVID-19, vaccination rollouts, US presidential election, policy changes, and positive earnings reports). Notably, EK proportions drop sharply near and after the model’s training cutoff, suggesting GPT-4o relies more on time series data and internal financial reasoning when external knowledge is unavailable. The corresponding rise in DR and FI supports this shift to a more data-driven approach. In short reports, FI and DR levels are similar, and technical indicator reports show more financial interpretation than direct references.

5. Conclusion

We present a systematic evaluation of LLMs for generating financial reports from time series data, addressing a novel task without ground truth benchmarks. Using both real and synthetic data, and assessing multiple report formats, we show that LLMs can produce coherent and informative analyses, though performance varies by model and report type. Our automated highlighting system aids in understanding model reasoning and factual grounding. The methodology and resources introduced here provide baselines for future research in financial time series reporting and related domains.

6. Limitations

This study highlights several limitations that suggest areas for future research.

One of the main limitations is the absence of definitive ground truth in the evaluation process, and the potential bias in the automated evaluations towards models from the same family as the LLM generated the report. It is mitigated by verifying and aligning scores using two different LLMs as evaluators and human annotations, but it still remains an open area for future research.

LLMs used in this study still carry the risk of generating plausible but incorrect information—a critical concern in the financial domain requiring stringent human oversight to maintain report integrity. Moreover, the effectiveness of large language models (LLMs) is heavily dependent on the precision of prompt engineering and model selection, which may not fully capture the nuances of financial data, leading to discrepancies in outputs.

7. Ethics Statement

This research adheres to ethical guidelines for NLP research. All data used in this study is publicly available, ensuring transparency and reproducibility. Annotation tasks were conducted by the paper's authors in consultation with financial analysts, who possess the necessary qualifications and expertise for the labeling process, and all annotators received appropriate compensation for their contributions.

The dataset and reports contain no personally identifiable information (PII) or intentionally offensive content. However, we acknowledge that language models may occasionally produce hallucinated or unintended outputs that could be potentially problematic. We recommend that practitioners implementing systems based on this research incorporate appropriate safeguards and human oversight to mitigate such risks.

To promote reproducibility and further research in this area, we intend to make our resources available upon request for legitimate research purposes. Researchers interested in accessing these materials should contact the corresponding author with details of their intended use.

8. Disclaimer

9. Bibliographical References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu

Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Toyin D Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145.

Gerald Appel. 2005. *Technical analysis: power tools for active investors*, first edition. FT Press.

Gabriel Assis, Daniela Vianna, Gisele L Pappa, Alexandre Plastino, Wagner Meira Jr, Altigran Soares da Silva, and Aline Paes. 2024. Analysis of material facts on financial assets: a generative ai approach. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024*, pages 103–118.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. 2025. Llms are biased evaluators but not biased for fact-centric retrieval augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26669–26684.

Shang-Hsuan Chiang, Lin-Wei Chao, Kuang-Da Wang, Chih-Chuan Wang, and Wen-Chih Peng. 2024. Badge: Badminton report generation and evaluation with llm. *arXiv preprint arXiv:2406.18116*.

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*.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. 2024. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21598–21634.

Federal Reserve Bank of St. Louis FRED. 2024. S&p 500 index (sp500), nasdaq composite index (nasdaqcom), dow jones industrial average (djia), nikkei 225 (nikkei225). Data retrieved from FRED, Federal Reserve Bank of St. Louis. Accessed on 16 Oct. 2024.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Masayuki Kawarada, Tatsuya Ishigaki, and Hiroya Takamura. 2024. Prompting for numerical sequences: A case study on market comment generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13190–13200.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Burton. G. Malkiel. 1973. *A Random Walk Down Wall Street*. Norton, New York.

Amaury Mantion, Melvin Kianmanesh Rad, Christophe Morel, Romain Faure, and Zachary Schillaci. 2024. Analysis and mitigation of chatgpt's dovish bias on classifying fedspeak. *Available at SSRN 4769112*.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

John J. Murphy and John J. Murphy. 1999. *Technical analysis of the financial markets*. New York Institute of Finance, Fishkill, N.Y.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Shunsuke Nishida, Yuki Zenimoto, Xiaotian Wang, Takuya Tamura, and Takehito Utsuro. 2023. Headline generation for stock price fluctuation articles. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 22–30.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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*.

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM at IJCAI*.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355.

A. Uncertainty Lexicon

We used the following words as lexicon of words indicating uncertainty:

'unclear', 'unknown', 'doubtful', 'uncertain', 'unconfident', 'tentative', 'tentatively', 'unsettled', 'undecided', 'unresolved', 'ambivalent', 'skeptical, 'questionable', 'questionably', 'unconvinced', 'might', 'maybe', 'possibly', 'could', 'may', 'could','potentially', 'conceivably', 'perhaps', 'perchance', 'probably', 'likely', 'presumably', 'apparently', 'seem', 'appears', 'feasibly', 'reportedly', 'allegedly', 'purportedly', 'plausibly', 'plausible'.

B. Prompts

We present the prompts used for the following tasks below.

B.1. Report generation prompts

Task: Short report generation

"You are a financial analyst tasked with writing a short report analyzing the main aspects of the `{index_name}` index based on its time series data. The report should be concise, focusing on key trends, volatility, and any notable price patterns observed in the data. Your report should be one or two paragraphs long, summarizing the overall performance and recent movements."

Closing Prices:

```
Date       Close
2020-04-28 8607.7
2020-04-29 8914.7
2020-04-30 8889.6
2020-05-01 8605.0
```

Task: Long report generation with numerical technical indicators

"You are a financial analyst tasked with writing a short report analyzing the main aspects of the `{index_name}` index based on its time series data and technical indicators. Focus on key trends, volatility, notable price patterns, and significant changes in the technical indicators such as moving averages or RSI. Summarize the overall performance and recent movements in two or three paragraphs."
Time Series Data with Technical Indicators:

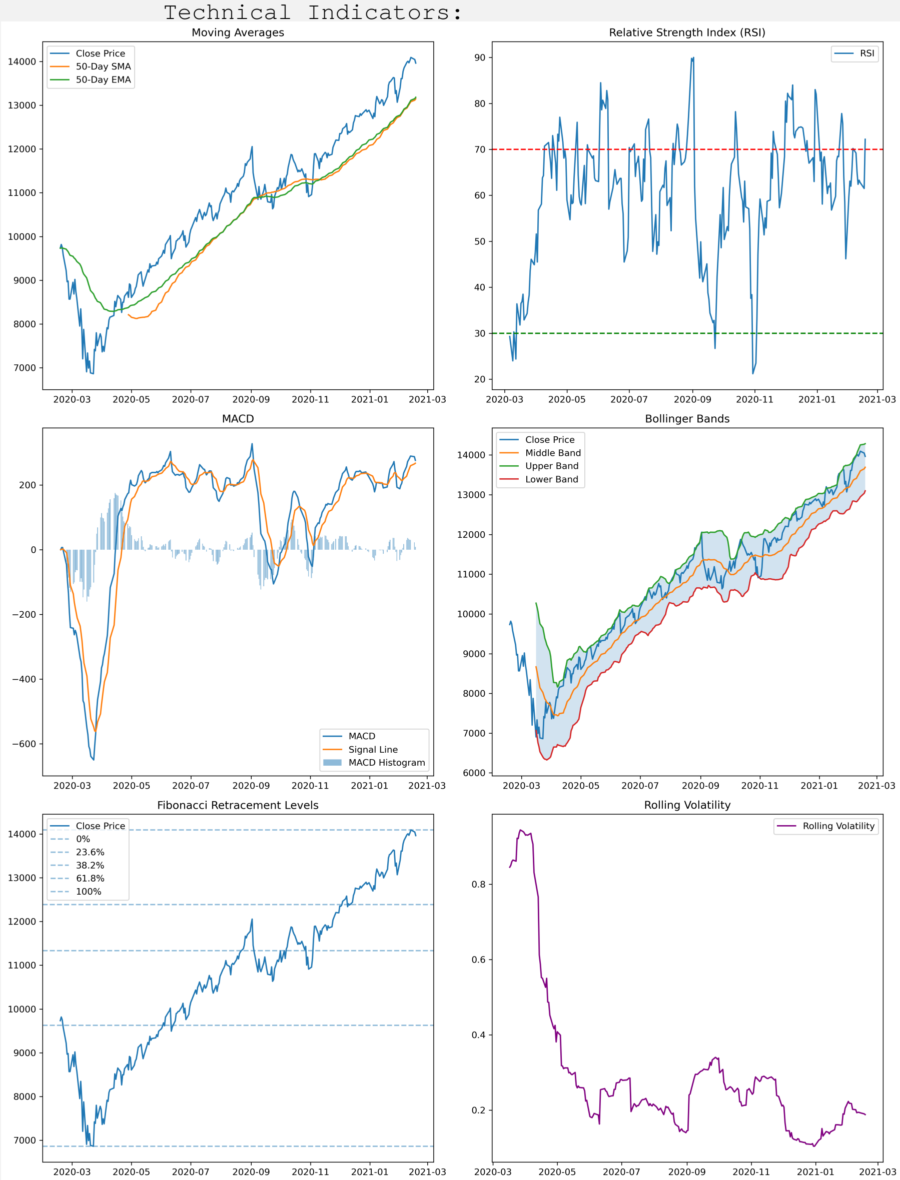
```
Date Close SMA_50 RSI MACD Volatility
2020-04-28 8607.7 8210.1 68.1 175.7 0.416
2020-04-29 8914.7 8193.7 69.7 200.0 0.425
2020-04-30 8889.6 8175.2 67.9 214.8 0.381
2020-05-01 8605.0 8152.2 59.0 201.2 0.408
```

Task: Long report generation with plots of technical indicators

"You are a financial analyst tasked with writing a short report analyzing the main aspects of the `{index_name}` index based on its time series data and technical indicator plots. The report should focus on key trends, volatility, and any notable price patterns observed in the data and the indicator plots. Your report should be two or three paragraphs long, summarizing the overall performance and recent movements."
The plots show the main technical indicators and this is the Time Series Data:

Date	Close
2020-04-28	8607.7
2020-04-29	8914.7
2020-04-30	8889.6
2020-05-01	8605.0

Technical Indicators:



B.2. G-eval

G-Eval: Consistency Evaluation Prompt

Task:
Your task is to rate the report on one metric.

Evaluation Criteria:
Consistency (1-5) - The factual alignment between the financial report and the time series data. A factually consistent report accurately reflects the trends, values, and key events present in the time series without introducing information not supported by it. Reports that contain hallucinated facts (i.e., statements that introduce or infer information not present in the time series) should be penalized.

Evaluation Steps:

1. Read the Time Series: Examine the time series data to understand the key facts, trends, and details it presents.
2. Read the Financial Report: Review the report and compare its content to the time series data. Identify any statements that do not align with the data or introduce unsupported information.
3. Assign a score for consistency based on the Evaluation Criteria.

Input:
Time series data:

```
    Date       Close
    2020-04-28 8607.7
    2020-04-29 8914.7
    2020-04-30 8889.6
    2020-05-01 8605.0
```

Technical Indicators **(if analyzing reports with technical indicators)**

```
  Date SMA_50 RSI MACD Volatility

2020-04-28 8210.1 68.1 175.7 0.416
2020-04-29 8193.7 69.7 200.0 0.425
2020-04-30 8175.2 67.9 214.8 0.381
2020-05-01 8152.2 59.0 201.2 0.408
```

Financial report:
{{Report}}
Evaluation Form (Scores ONLY):

- Consistency:

G-Eval: Coherence Evaluation Prompt

Task:
Your task is to rate the report on one metric.

Evaluation Criteria:
Coherence (1-5) - The degree to which the report is logically organized and well-structured. The report should clearly present the insights from both the time series data and the technical indicators in a way that builds sentence by sentence into a coherent body of information. The report should not feel like a disjointed collection of statements but should present a logical progression of ideas and insights, where each sentence and paragraph naturally follows from the previous ones.

Evaluation Steps:

1. Examine the Time Series and Technical Indicators: Carefully review both the time series data and the technical indicators. Identify the main trends, signals, and key points in the data.
2. Read the Financial Report: Read the financial report and assess its logical flow and structure. Check if the report covers the key trends and points from the time series and technical indicators in a clear, organized, and logical manner. Look for a smooth progression of information, where each insight follows naturally from the previous one.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Input:
Time series data:

```
    Date       Close
    2020-04-28 8607.7
    2020-04-29 8914.7
    2020-04-30 8889.6
    2020-05-01 8605.0
```

Technical Indicators **(if analyzing reports with technical indicators)**

```
  Date SMA_50 RSI MACD Volatility

2020-04-28 8210.1 68.1 175.7 0.416
2020-04-29 8193.7 69.7 200.0 0.425
2020-04-30 8175.2 67.9 214.8 0.381
2020-05-01 8152.2 59.0 201.2 0.408
```

Financial report:
{{Report}}
Evaluation Form (Scores ONLY):

- Coherence:

G-Eval: Fluency Evaluation Prompt

Task:
Your task is to evaluate the report on one metric.

Evaluation Criteria:
Fluency (1–5) – The readability and naturalness of the language used in the report. A fluent report should be free from grammatical errors, awkward phrasing, and unnatural language. It should read smoothly and be easy to understand.

Score Breakdown:
- 1 = The report is highly unnatural with significant grammar and phrasing issues.
- 2 = The report has major fluency problems, with noticeable awkwardness and errors.
- 3 = The report is somewhat fluent, but with some noticeable issues.
- 4 = The report is mostly fluent, with only a few minor issues.
- 5 = The report is fully fluent, with natural and smooth language.

Evaluation Steps:

1. Read the Report Carefully: Pay close attention to the language used, including grammar, phrasing, and overall readability.
2. Identify Language Issues: Look for any grammatical errors, awkward sentences, or unnatural phrasing that may hinder the readability of the report.
3. Assign a score for Fluency on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Input:
Financial report:
{{Summary}}

Evaluation Form (Scores ONLY):
- Fluency (1–5):

B.3. Highlights

Task: Source Segment Classification

Please categorize the following text into three categories:

1. Direct reference: Segments that directly mention numerical values or trends from the input time series data.
2. Financial interpretation: Segments that infer or conclude based on the observed data without external knowledge.
3. External knowledge: Segments that provide context or explanations using knowledge outside the observed time series data.

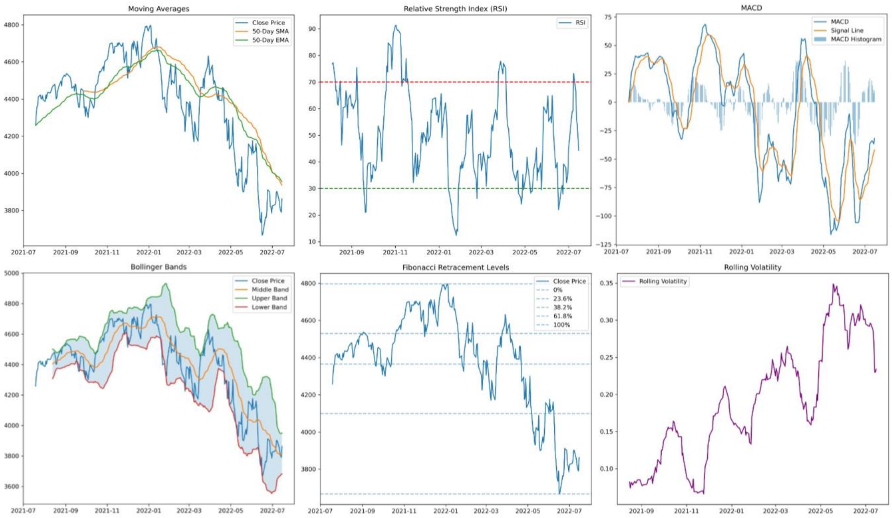
Segments should be non-overlapping.

Return the categorized segments in the following JSON format:

```
{'direct_reference': ['segment1', 'segment2', ...],
'financial_interpretation': ['segment1', 'segment2', ...],
'external_knowledge': ['segment1', 'segment2', ...]}
```

Text: {{sentence}}

C. Report Examples



GPT-4o Short

The analysis of the S&P 500 index from July 19, 2021, to July 15, 2022, reveals notable trends and volatility reflective of broader market conditions. Initially, the index exhibited a positive trajectory, reaching a peak of 4796.6 on January 3, 2022. This period demonstrated sustained growth momentum, driven by post-pandemic recovery expectations and strong corporate earnings. However, the index entered a volatile phase marked by significant fluctuations, influenced by concerns over inflation, interest rate adjustments by the Federal Reserve, and geopolitical tensions, notably the conflict surrounding Ukraine.

From the start of 2022, the S&P 500 displayed increased volatility, culminating in a marked downtrend becoming evident by May and June 2022. Key support levels around 3900 were tested repeatedly, with lows reaching 3666.8 on June 16, 2022, reflecting a more bearish sentiment as market participants grappled with economic uncertainties. Despite these challenges, occasional recoveries occurred, indicating investor resilience. By mid-July 2022, the index remained subdued compared to its prior peaks, underscoring the prevailing cautious outlook amidst persistent economic headwinds. Overall, the period encapsulates a shift from optimism to risk aversion, characteristic of a market recalibrating to evolving global challenges.

GPT-4o-mini Short

The analysis of the S&P index's closing prices from mid-2021 to mid-2022 indicates a primarily upward trajectory during the latter half of 2021, peaking at a closing price of 4796.6 on January 3, 2022. This bullish trend was characterized by consistent gains with only minor pullbacks until early January 2022. However, starting in January 2022, the index began to experience increased volatility, with significant declines, particularly observed in May and June 2022, where prices dropped to a low of 3749.6 on June 13, 2022.

Throughout this period, the index demonstrated pronounced fluctuations, primarily driven by broader economic concerns, including inflation and geopolitical tensions, contributing to heightened market volatility. The price pattern reflects a transition from a strong upward trend to a correction phase, with a notable drop of nearly 22% from its January high to mid-June lows. This report highlights the S&P index's susceptibility to external market pressures, resulting in a notable shift from growth to a risk-off sentiment in early 2022.

Llama Short

S&P Index Review
Based on the provided closing prices, the S&P index exhibited a moderate trend throughout the analyzed period, with some notable oscillations. From July 2021 to December 2021, the index experienced a general upward trajectory, increasing from approximately 4250 to 4800. However, the pace of this growth decelerated in the final quarter of 2021, coinciding with an increase in volatility. A significant drop in January 2022 brought the index to around 4500 before it began to recover. The first half of 2022 saw the index fluctuate between 4200 and 4700, with a slightly upward bias.

Key Trends and Price Patterns
A notable large downfall occurred in March 2022, where the index plummeted from approximately 4600 to 4000, resulting in an overall volatility increase. Partially offsetting this, the S&P index rebounded by November 2022 to reach levels around 4600-4700 once again. In the last few months of 2022, the index exhibited a mixed trend, with strong prices seen in the autumn months (September to November) being wiped out by large downturns in spring. The 50-day moving average and the 200-day moving average demonstrate varying degrees of divergence throughout 2022, indicating wavering market sentiment. Overall, volatility and price movements showcased year-over-year uncertainty since the first reading in July 2021.

Issues: hallucination of numbers and months; mentioned data outside of the observed interval

Gemini Short

The S&P 500 index displayed a strong upward trend from July 2021 through October 2021, reaching a peak above 4,600. The index then experienced a correction, dropping below 4,300 by the end of November. In December, the index rebounded significantly, closing above 4,700 for the first time since early October. This upward momentum continued into early January 2022, but the index then entered a downward trend lasting for several months, ultimately reaching a low below 3,700 in June. Despite this volatility, the index showed signs of recovery in July and August, closing above 4,000, indicating a possible reversal in trend. However, the index has continued to be volatile, with further corrections and periods of growth, demonstrating a mixed performance over the analyzed period.

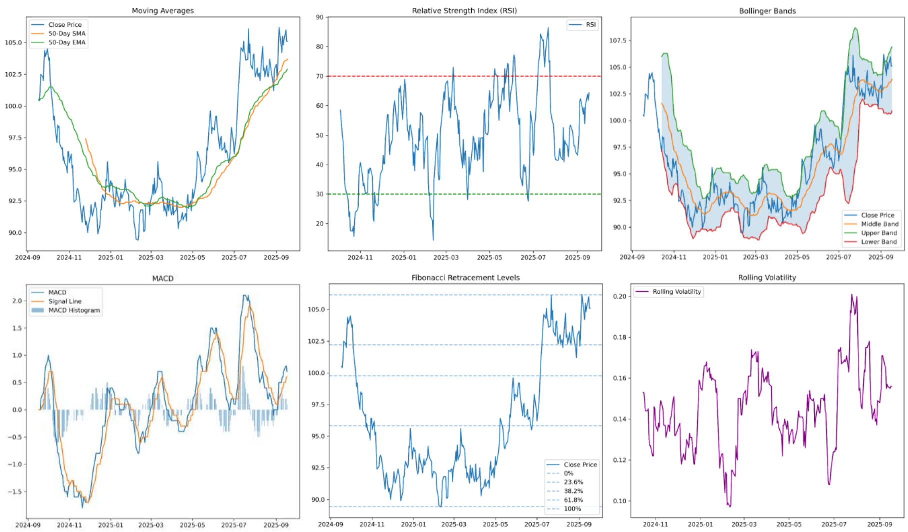
Issues: inaccurate description of upward trend from July 2021 through October 2021 where there is a dip; mentioned data outside of the observed interval

Phi3 Short

The S&P index experienced a general upward trend from July 19 to August 26, with closing prices ranging from 4258.5 to 4486.5. The index saw a slight increase in volatility during this period, with minor fluctuations in the closing prices. Notable price patterns include a steady rise from July 19 to August 26, followed by a slight decline from August 27 to August 30. The index then experienced a more significant drop from August 30 to September 1, followed by a recovery and a gradual increase from September 1 to September 10. The index then experienced a sharp decline from September 10 to September 13, followed by a recovery and a gradual increase from September 13 to September 23. The index then experienced a sharp decline from September 23 to September 27, followed by a recovery and a gradual increase from September 27 to October 1. The index then experienced a sharp decline from October 1 to October 4, followed by a recovery and a gradual increase from October 4 to October 11. The index then experienced a sharp decline from October 11 to October 13, followed by a recovery and a gradual increase from October 13 to October 21. The index then experienced a sharp

Issues: hallucination of numbers; no holistic view of the time series, plain descriptions rather than analysis; sentence not finished

Figure 7: Example reports generated on real data: S&P Index during 2021-07-17 and 2022-07-17. Reports in green boxes are of good quality; reports in orange boxes are problematic with identified issues. Colored highlights indicate **Direct Reference (blue)**, **Financial Interpretation (green)**, and **External Knowledge (red)** using our proposed highlighting system. LLM-generated example outputs not constituting assertions by JPMC authors.



GPT-4o Short

The Global Market Index (GMI) has exhibited notable fluctuations and trends over the recent months. Initially, the period from mid-September to the end of October 2024 showed a modest decline, with the index dropping from 102.5 to approximately 95.5. This period was characterized by heightened volatility, marked by brief recoveries followed by quicker declines, possibly indicating underlying market uncertainties. Subsequently, as the year progressed into early 2025, the index displayed further downward pressure, reaching as low as 89.4 in February 2025, highlighting a prolonged bearish sentiment.

However, March 2025 marked the beginning of a recovery phase, with the GMI increasing to approximately 98.3 by the end of June. Noteworthy is the strong rally observed in July 2025, where the index surged to 105.1, indicating a potential reversal or correction in the bearish trend witnessed earlier. The latter part of the data, particularly around August and September 2025, shows the GMI stabilizing in a higher range around 105. This resurgence signifies renewed investor confidence and a potential shift towards a bullish market sentiment. Overall, while the GMI faced a challenging period characterized by declines and volatility in late 2024 and early 2025, the recent upward trajectory suggests an optimistic outlook as the index approaches previous highs.

GPT-4o-mini Short

Global Market Index (GMI) Financial Report:
The Global Market Index (GMI) has exhibited notable volatility over the analyzed period, particularly between late September 2024 and early November 2024. Starting at a high of 104.5 on September 30, 2024, the index faced a significant downturn, dropping to a low of 90.0 by late November 2024. This pattern reflects a broader market correction, indicating investor caution amid macroeconomic uncertainties. However, a modest recovery began in early December, as the index rebounded to close at 94.2 by December 25, 2024.

In the first half of 2025, the GMI's performance demonstrated resilience, nearing its pre-correction levels and peaking at 106.2 by early September. Nevertheless, the volatility persisted, with fluctuations observed in closing prices within the upper 100 range. The subsequent weeks showed a stable trend, suggesting a potential consolidation phase as market participants assess economic fundamentals. Overall, the GMI's trajectory indicates a market characterized by reactive movements to external economic factors, underlining the importance of ongoing monitoring and analysis for investment strategies.

Issues: mentioned broader market and external economic factors not present in the time series; noting volatility at the beginning but missing higher volatility later on

Llama Short

Financial Analysis Report: Global Market Index (GMI)
The Global Market Index (GMI) experienced a mixed performance over the past year, fluctuating between periods of growth and decline. The closing prices indicate that the index started the year 2024 at 98.8, decreased to a low of 87.8 in October 2024, before rebounding to finish the year at 96.7. In 2025, the index continued to trend upwards, reaching a peak of 106.2 in September 2025. Overall, the GMI showed a slight increase of 8.5% in 2025, demonstrating a moderate level of growth.

Volatility was present throughout the period, with several notable fluctuations. Specifically, the index dropped by 10.5% in October 2024, only to rebound by 15.5% in December 2024. In addition, the rapid increase in the index prices from August to September 2025, which reached a 12.5% rise, highlights the volatility in the market. Despite this, the GMI remains relatively stable, exhibiting a volatility coefficient of 1.21, indicating moderate price fluctuations. These trends and patterns provide valuable insights into the performance and dynamics of the Global Market Index.

Issues: hallucination of numbers; values of drops are not self-consistent within the report

Gemini Short

The GMI-second index has shown a generally positive trend over the analyzed period, with an overall upward trajectory. The index started the period around 100.5 and reached a high of 106.2 by the end of the data. This upward trend was interrupted by periods of volatility, with some notable fluctuations in the price. The most significant decline occurred in October, when the index dropped from its high of over 104 to below 97, before recovering partially.

The data suggests that the GMI-second index experiences periods of both upward and downward movements, suggesting that the index could be susceptible to market fluctuations. The overall positive trend suggests a potential for further growth in the future, although the volatile nature of the index requires careful monitoring to assess potential risks and opportunities.

Phi3 Short

The GMI-second index experienced a general upward trend in its closing prices from September 17 to October 26, with a peak of 104.6 on October 26. The index then saw a significant drop in closing prices from October 27 to November 14, reaching a low of 91.8. After this period, the index experienced a slight recovery, with closing prices ranging between 92.0 and 94.2 from November 15 to December 2. The index then saw a sharp decline from December 2 to December 11, reaching a low of 89.9. From December 12 to December 24, the index fluctuated between 90.0 and 94.2, with a notable increase in closing prices from December 24 to December 25. The index then experienced a sharp decline from December 26 to December 31, reaching a low of 90.0. From January 1 to January 10, the index saw a gradual increase in closing prices, reaching a peak of 94.1. From January 11 to January 18, the index experienced a sharp decline, reaching a low of 90.0. From January 19 to January 26, the index saw a gradual increase in closing prices,

Issues: hallucination of numbers; no holistic view of the time series, plain descriptions rather than analysis; sentence not finished

Figure 8: Example reports generated on synthetic data: GMI Index during 2024-09-17 and 2025-09-17. Reports in green boxes are of good quality; reports in orange boxes are problematic with identified issues. Colored highlights indicate **Direct Reference (blue)**, **Financial Interpretation (green)**, and **External Knowledge (red)** using our proposed highlighting system. LLM-generated example outputs not constituting assertions by JPMC authors.