FinCall-Surprise: A Large Scale Multi-modal Benchmark for Earning Surprise Prediction

Anonymous ACL submission

Abstract

Predicting corporate earnings surprises is a profitable yet challenging task, as accurate forecasts can inform significant investment decisions. However, progress in this domain has been constrained by a reliance on expensive, proprietary, and text-only data, limiting the development of advanced models. To address this gap, we introduce **FinCall-Surprise** (Financial Conference Call for Earning Surprise Prediction), the first large-scale, open-source, and multi-modal dataset for earnings surprise prediction. Comprising 2,688 unique corporate conference calls from 2019 to 2021, our dataset features word-to-word conference call textual transcripts, full audio recordings, and corresponding presentation slides. We establish a comprehensive benchmark by evaluating 26 state-of-the-art unimodal and multimodal LLMs. Our findings reveal that (1) while many models achieve high accuracy, this performance is often an illusion caused by significant class imbalance in the real-world data. (2) Some specialized financial models demonstrate unexpected weaknesses in instruction-following and language generation. (3) Although incorporating audio and visual modalities provides some performance gains, current models still struggle to leverage these signals effectively. These results highlight critical limitations in the financial reasoning capabilities of existing LLMs and establish a challenging new baseline for future research. The FinCall-Surprise dataset is available at https://anonymous.4open.science/r/FinCall-Surprise-F212.

1 Introduction

In the finance domain, a conference call, often referred to as an earnings call, serves as a critical communication channel between the management of a public company and its stakeholders, including analysts, investors, and the media (Kimbrough, 2005). During these calls, executives present the firm's financial results, discuss recent performance, and provide forward-looking guidance. Analysts and investors closely analyze this information to predict their expectations of the company's earnings per share (EPS) (Patell, 1976). An *earnings surprise* occurs when the company's reported EPS differs from market expectations (Latane and Jones, 1979). Historically, a positive earnings surprise, where actual earnings exceed the reported EPS, often correlates with a subsequent increase in the company's stock price, while a negative surprise typically leads to a decline. Therefore, the ability to accurately predict an earnings surprise ahead of the official announcement is a significant challenge and an opportunity for investors seeking to inform their trading decisions (Skinner and Sloan, 2002).

Given the potential financial rewards, researchers and investors have long sought to systematically analyze conference calls to predict earnings surprises. This pursuit began with traditional machine learning models (Dhar and Chou, 2001; Doyle et al., 2006; Bissessur and Veenman, 2016), and has recently shifted attention toward large language models (LLMs) (Zhang et al., 2025; Zhang and He, 2025; Zhu et al., 2025; Liang and Carrasco Kind, 2025). However, while the performance of these models is often impressive, their capabilities are inherently constrained by a fundamental limitation, which is the unimodal and text-only nature of the datasets they are trained on. In reality, human analysts do not just analyze what executives say, but also how they say it. The vocal tone and rhythm of a speaker can reveal confidence or uncertainty that is lost in a plain transcript. Similarly, visual information, such as presentation slides, provides an essential context that is often referenced during the discussion.

To address this critical gap, we introduce **FinCall-Surprise**, a novel, large-scale, multimodal dataset specifically designed for the task of earnings surprise prediction. The dataset is com-

posed of 2,688 unique corporate conference calls spanning from 2019 to 2021, with 919 calls from 2019, 704 from 2020, and 1,065 from 2021. Most importantly, each call in the dataset is complete, containing three synchronized modalities: 1) the word-to-word textual transcript, 2) the full audio recording of the call, and 3) the corresponding presentation slides referenced by executives. With the introduction of this dataset, our primary contributions are as follows:

- We present FinCall-Surprise, the first large-scale, multi-modality dataset focused on real-world corporate earnings calls, providing a rich resource for developing and evaluating earning surprise prediction models.
- We establish a comprehensive benchmark by evaluating more than 20 state-of-the-art unimodal and multi-modal LLMs.
- Our benchmarking findings reveal that current models struggle to effectively leverage the multimodal signals for this task, establishing a challenging baseline and highlighting the need for more sophisticated architectures.

2 Related Work

2.1 Earning Surprise Prediction Models

The use of AI in predicting earnings surprises has a long history that began with traditional statistical and econometric models. In early foundational research, researchers like Numbers (1968) relied on simple models such as linear regressions. A pivotal shift occurred when Ou and Penman (1989) used a broader set of financial statement data with a logit model to predict the direction of earnings, foreshadowing the machine learning era. From approximately 2000 to 2015, classic machine learning models like SVMs, Random Forests, and Gradient Boosting took center stage (Dhar and Chou, 2001; Skinner and Sloan, 2002; Zolotoy, 2012). Their key advantage was the ability to analyze hundreds of financial variables simultaneously, capturing complex, non-linear relationships that eluded older methods. The period after 2015 saw the rise of deep learning, particularly RNNs and LSTMs, which introduced unstructured text as a new and powerful data source (Vargas et al., 2017; Meursault et al., 2023; Zhu, 2020; Sawhney et al., 2020).

The contemporary approach to earnings surprise prediction is dominated by LLMs, which represent a revolutionary leap in analytical capability. Researchers began feeding models text from conference call transcripts and financial reports to extract sentiment and forward-looking statements (Koval et al., 2023). Since LLMs can possess a deep and nuanced understanding of financial jargon and context, they can dissect conference call transcripts, press releases, and news with unprecedented accuracy (Araci, 2019; Kim et al., 2024; Lee et al., 2025). However, due to the lack of high-quality multimodal conference call datasets, models for earnings surprise prediction have not yet reached the multimodal level.

2.2 Earning Surprise Prediction Datasets

Due to the high value and sensitivity of corporate financial data, most resources are stored within major commercial platforms such as WRDS, Bloomberg, and Refinitiv, which require paid subscriptions or API access. Despite this barrier, there has been steady progress in developing open-source financial datasets, each typically targeting narrower and more specialized tasks in the finance domain to advance AI research in this field. For instance, datasets such as (Chen et al., 2021, 2024; Zhu et al., 2021; Lai et al., 2024; Reddy et al., 2024; Yuan et al., 2024) focus on question answering over financial reports. Other datasets have been developed for stock market prediction (Dong et al., 2024; Rao, 2021; Patel, 2021; Qin and Yang, 2019), fraud detection (Feng et al., 2023), sentiment analysis (Borhani, 2024; Cortis et al., 2017), and misinformation classification (Rangapur et al., 2025). More recent efforts have extended financial datasets into the multimodal space, motivated by both the limitations of single-modality data and the growing capabilities of LLMs to understand information beyond text (Li et al., 2020; Shu et al., 2025; Luo et al., 2025).

Earnings surprise prediction has recently gained increasing attention in the context of LLMs, with several studies proposing new methods to improve performance on this task (Zhu et al., 2025; Zhang et al., 2025). However, relatively few works have focused on dataset construction. Early efforts such as Koval et al. (2023, 2024) introduced conference call transcript datasets for this purpose, but these resources are not fully open source, limiting their accessibility and impact. Consequently, many current studies continue to depend on expensive commercial data providers, often accessing conference call content through paid APIs (Heater et al., 2025;

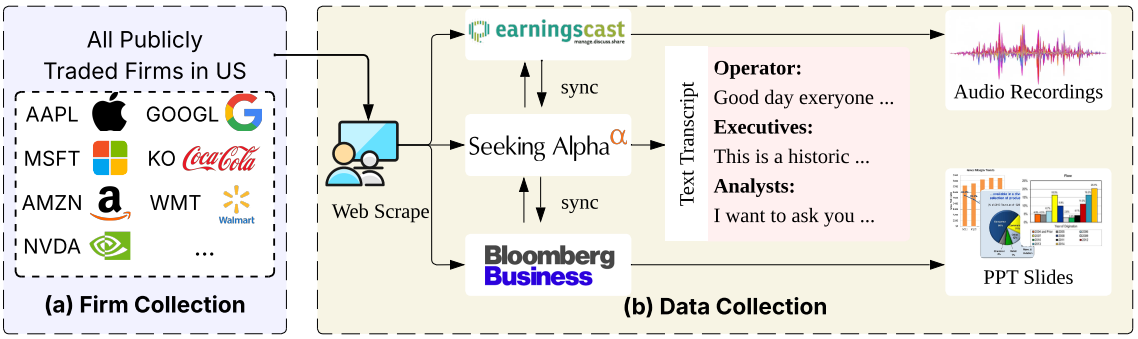

Figure 1: Overview of our data construction pipeline, which consists of two stages: **(a) Firm Collection (left):** We select large, publicly traded US firms based on market capitalization ($> \$1$B) and daily trading volume ($> \$50$M). **(b) Data Collection (right):** For each firm, we gather and synchronize three modalities for each quarterly earnings call: textual transcripts from Seeking Alpha, audio recordings from EarningsCast, and presentation slides from sources like Bloomberg Business.

Alsabah, 2025; Zhang and He, 2025).

Our proposed FinCall-Surprise dataset addresses this challenge by introducing the first fully open-source, multi-modal conference call dataset specifically designed for earnings surprise prediction, covering the period from 2019 to 2021. Each data in our dataset provides synchronized modalities: the complete text transcript of the conference call, the full audio recording, and the accompanying presentation slides referenced by executives. By releasing this dataset, we aim to remove the reliance on commercial APIs and enable the research community to explore earnings surprise prediction at a multimodal level, opening new directions for both financial NLP and multimodal learning.

3 FinCall-Surprise Construction

In this section, we detail the creation process for the FinCall-Surprise dataset.

3.1 Data Construction Pipeline

The construction of FinCall-Surprise follows a comprehensive two-stage pipeline designed to collect, synchronize, and annotate data from multiple sources. Our objective is to build a robust benchmark that integrates the textual, auditory, and visual dimensions of corporate conference calls. As illustrated in Figure 1, the first stage involves assembling the universe of all publicly traded firms in the United States with substantial size and liquidity. To focus on economically significant firms, we restrict the sample to companies with market capitalizations exceeding $1 billion and average daily trading volumes above $50 million. This initial firm collection yielded an initial pool of more than 4,000 unique companies.

In the second stage, we acquire the three distinct modalities for each firm's earnings conference calls. For textual transcripts, we systematically web scrape content from established financial platforms, primarily Seeking Alpha[1]. These transcripts are well structured, containing both speaker identifiers (operators, executives and analysts) and the corresponding speech content. For the auditory component, we web scrape the associated audio recordings from EarningsCast[2]. Finally, we construct a dataset of corporate presentation slides by collecting files from multiple sources, including Bloomberg News and company websites. Most publicly traded firms host one conference call per fiscal quarter. To ensure temporal consistency across modalities, we align all data sources by quarterly reporting periods. In addition, we use the conference call titles to cross-check and verify that the transcript, audio recording, and presentation slides correspond to the same event, ensuring that our multi-modal dataset is accurately matched at the event level.

3.2 Earning Surprise Label Preliminaries

Before classifying each conference call label as positive or negative, we first collected the reported Earnings Per Share (EPS) and the analyst consensus EPS forecasts from the IBES database. Following Latane and Jones (1979), we measure the earnings surprise (ES) using the Standardized Unexpected Earnings (SUE), defined as the difference between the reported EPS and the analyst consen-

[1] https://seekingalpha.com/
[2] https://earningscast.com/

	Transcript (words)			Slide (pages)			Audio (sec)			Label (Percentage)		
Year	Mini	Max	Avg.	Mini	Max	Avg.	Mini	Max	Avg.	Positive	Negative	Total
2019	905	57,485	8,634.66	1	156	26.05	1096.07	8870.97	3642.66	0.79	0.21	919
2020	865	46,623	8,925.10	1	170	27.80	1404.06	7585.31	3859.25	0.86	0.14	704
2021	752	45,632	8,646.77	1	113	27.46	1197.95	11938.90	3720.29	0.89	0.11	1065

Table 1: Data statistics of the FinCall-Surprise. The dataset spans three years (2019–2021), with each conference call containing three synchronized modalities: text transcripts, presentation slides, and audio recordings. We report the minimum, maximum, and average values for each modality. Transcript length is measured in words, slides in pages, and audio in seconds. For labels, we present the percentage distribution of positive and negative classes. The last column shows how much data we have for each year.

sus estimate, scaled by the standard deviation of analyst forecasts. The consensus estimate is calculated as the mean of the most recent valid analyst forecasts issued within one month after the conference call, allowing analysts to revise their expectations based on the call content and recent financial disclosures. This design provides a forward-looking measure of market expectations and yields a more realistic, yet challenging, prediction task. The average time span between the input transcript and the target earnings event is about three months, further highlighting the difficulty of the task.

$$\text{ES} = \frac{\text{EPS}_{\text{reported}} - \text{Avg}(\text{EPS}_{\text{estimated}})}{\text{Std}(\text{EPS}_{\text{estimated}})} \quad (1)$$

$$y = \begin{cases} 0, & \text{ES} \leq -\delta \\ 1, & \text{ES} \geq \delta \end{cases} \quad (2)$$

We convert the continuous earnings surprise (ES) into a binary classification task by assigning a label of Positive $(+1)$ when ES $> \delta$ and Negative (0) when ES $< -\delta$, where $\delta = 0.50$. This threshold follows prior studies on standardized unexpected earnings (Eli Bartov, 1992) and price momentum (Luo et al., 2022), which classify earnings surprises as large when $|\text{SUE}| \geq 0.5$. The chosen cutoff balances sample size with event significance. Observations with immaterial surprises (i.e., ES $\in [-0.50, 0.50]$) are excluded, as these near-zero values typically elicit weak market responses and may reflect earnings management. Although ES is continuous, market reactions are largely binary, responding more to the direction than to the magnitude of the surprise. We therefore focus on material surprises that are more likely to influence investor behavior and asset pricing.

3.3 Data Analysis

A detailed statistical analysis of our dataset is presented in Table 1. The statistics highlight the substantial scale and complexity of the data across all three modalities. On average, the textual transcripts for each conference call contain approximately 8,600 words. The accompanying presentation slides average 27 pages in length. Finally, the audio recordings have an average duration of approximately 3,700 seconds. Each data point in our dataset is annotated with a binary label indicating either a positive or negative earnings surprise. As shown in the last column, the distribution of labels is highly imbalanced, with the proportion of positive labels significantly exceeding that of negative labels. This imbalance is unavoidable, since our dataset is constructed from real-world financial data, where positive earnings surprises occur more frequently. The content of the conference calls is consistently structured around three primary speaker roles: Operators, Executives, and Analysts. An example is provided in Figure 2. The Operator typically opens and closes the call. The Executives, including chief members such as the CEO and CFO, present the company's financial results and strategic outlook. Finally, the Analysts, representing investment firms and financial institutions, pose questions to the executives to gain deeper insights into the company's performance.

4 Benchmark Setting

4.1 Baseline LLMs

To establish a comprehensive performance benchmark for FinCall-Surprise, we selected a diverse set of 26 state-of-the-art, open- and close-source models. For a structured analysis, these models were categorized into four distinct groups.

The first group consists of general purpose unimodal models, which process text-only inputs, and includes GPT-oss-20B (OpenAI, 2025), Qwen-2.5-14B (Team, 2024), Mistral-7B (Jiang et al., 2023), LLaMa-3.2-3B (Grattafiori et al., 2024), Gemma-3-1B (Team, 2025a). The second group con-

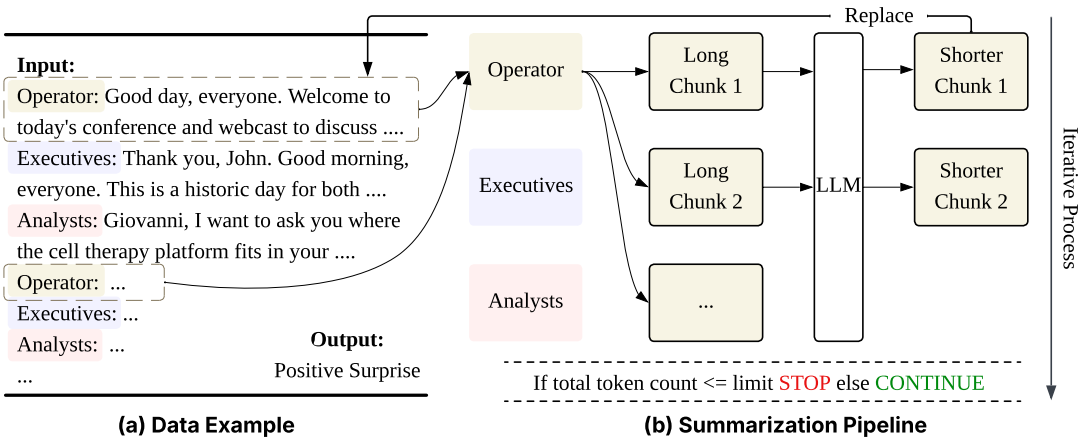

Figure 2: Illustration of our summarization pipeline. (a) A data example showing a conference call transcript with speaker turns (Operator, Executives, Analysts) and its corresponding earning surprise label. (b) The summarization pipeline, where transcripts are grouped by speaker and iteratively summarized by the LLM until the total token count falls below the predefined limit.

sists of finance-specialized unimodal models fine-tuned for the financial domain, including Finance-LLaMa3-8B (Cheng et al., 2024), Finance-LLaMa-8B (tarun7r, 2025), Finance-R1 (KhazarAI, 2024), LLaMa-RAG (Liu et al., 2024). The third category features Vision-Language Models (VLMs), designed to process both images and text. For this group, we evaluated GPT-5, GPT-5-mini, GPT-5-nano, GPT-4.1, Mistral-3.1-24B (Jiang et al., 2023), Gemma-3-12B (Team, 2025a), LLaMa-3.2-11B (Grattafiori et al., 2024), Sa2VA-8B (Yuan et al., 2025), Qwen-2.5-VL-7B (Team, 2025b), Qwen-2-VL-7B (Wang et al., 2024), Cosmos-7B (Azzolini et al., 2025), and LLaVa-1.6-7B (Liu et al., 2023). Finally, we assessed audio-language models, which handle audio and text inputs, including Voxtral-24B (Liu et al., 2025), DeSTA-2.5-Audio-8B (Lu et al., 2025), Qwen-2.5-Omni-7B (Xu et al., 2025), Qwen-2-Audio-7B (Chu et al., 2024), Gemma-3n-E4B (Team, 2025a).

4.2 Benchmark Input Design

To systematically evaluate the performance of different models, we designed a customized input and prompting strategy for each modality.

Text-Only Modality. To ensure a fair evaluation across all baseline models, we standardized the input length to accommodate the smallest context window of 32K tokens. We set the threshold to 31K, because we need to reserve some space for the instruction prompt. As approximately 20% of the transcripts in our dataset exceed this limit, we employed a targeted summarization strategy for these longer texts. Transcripts already under the 31K token threshold were left unmodified. To minimize information loss, we adopted a conservative, iterative summarization process as shown in Figure 2. First, each transcript is segmented into chunks based on speaker type (e.g., Operator, Executive, Analyst). We then incrementally summarize the chunks that are least critical to the financial context, beginning with those from the "Operator". These sections typically contain procedural dialogue, such as call introductions and closings, rather than useful financial discussion. Every time a chunk is summarized, we recalculate the total token count, and the iterative process stops as soon as the transcript length falls below the 31K token threshold. If summarizing all "Operator" chunks is insufficient, the iterative process continues with chunks from "Executives" and then "Analysts". In practice, summarization usually stops midway through this process, leaving the majority of the transcript untouched and thus preserving as much original information as possible. Finally, the processed text was formatted into a single concatenated sequence for model input: "{Instruction} {Text Transcript} {Answer Format}". This structure provides the model with the complete verbal context of the conference call.

For this summarization task, we utilized the BART-Large-CNN model (Lewis et al., 2019), because it is lightweight, reliable, and does not require a specific instruction prompt to function effectively. For token number checking, we used "tiktoken" python library. While summarization was a necessary step for our experiments, we will release the original,

full-length transcripts to support future research that may leverage models with larger context windows or more advanced summarization techniques.

Image-Text Modality. Evaluating multi-modal models that accept text and images required a preprocessing step for the presentation slides, which were originally in PDF format. As most models do not support PDF inputs, we converted each page of a presentation into a separate image. To manage the image input limits common to these models, we developed a random sampling strategy. For each presentation, we randomly selected three images from the slide range, excluding the first and last pages, which typically contain non-substantive content like title or closing. In cases where a presentation had three or fewer content slides, all available slides were used. The final input for this modality was structured as: "{Images} {Instruction} {Text Transcript} {Answer Format}". Note that although our benchmark experiment uses only three images, we still release the full PDF in our dataset.

Audio-Text Modality. For models capable of processing audio, we intentionally omitted the text transcript from the input. Since the audio recording contains the same verbal information as the transcript, this approach isolates the model's ability to comprehend and reason based on auditory signals alone, such as speaker tone and inflection. The input was therefore constructed as: "{Audio} {Instruction} {Answer Format}".

Across all modalities, the prompt structure was designed to test both the model's reasoning capabilities and its ability to follow instructions. We required the final answer to be placed in a specific, designated location within the output, which enabled reliable parsing for our automated evaluation pipeline. Detailed examples of the prompts for each modality are provided in the Appendix A.

4.3 Benchmark Metrics

Since our benchmark involves only two possible outputs, accuracy naturally serves as a primary evaluation metric. However, accuracy alone can be misleading, particularly under class imbalance. To provide a more balanced evaluation, we report macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. These metrics treat both positive and negative classes equally, regardless of their frequency in the dataset. *By incorporating macro-averaged metrics alongside accuracy, we account for the dataset's class imbalance and obtain a more reliable assessment of model performance across both positive and negative cases.*

4.4 Implementation Details

All experiments were conducted on a single NVIDIA A100 SXM4 GPU with 80GB of memory. For all baseline models, we maintained their official repository configurations and loaded them with bfloat16 precision to optimize computational efficiency. Beyond this precision setting, no other modifications were made to the models' default parameters, ensuring a fair and reproducible comparison across all baseline evaluations.

5 Benchmark Results

We have listed all 26 models' result in Table 2 including 5 general purpose text modality models, 4 financial finetuned models, 12 image-text modalities models, 5 audio-text modalities models. We have the following observations.

5.1 High Performance Illusion

As shown in Table 2, many models appear to achieve strong results when evaluated only by accuracy. In fact, 15 out of 26 models surpass 70% accuracy. However, this impression of high performance is misleading and illusional. Our dataset is heavily imbalanced toward positive samples, and when we examine precision, recall, and F1 score, most of the models fall below 55%. This indicates that most models are biased toward predicting the majority (positive) class and fail to generalize to negative cases. A plausible interpretation for this biased behavior is that the models have simply learned the statistical distribution of the financial domain, where positive outcomes are more frequent. This aligns with our own findings during the dataset construction phase, which confirmed that real-world financial data often exhibits significant imbalance.

A closer look also reveals that model accuracy tends to increase on the 2021 subset compared to earlier years. For instance, Qwen-2.5 in the General-Text category achieves accuracies of 0.78 and 0.77 in 2019 and 2020, respectively, but rises to 0.84 in 2021. However, this gain does not reflect genuine model improvement. Rather, it stems from the fact that the 2021 data is more imbalanced, with 89% positive and only 11% negative samples. When models default to predicting the majority class, their accuracy naturally rises under such conditions. This improvement is therefore illusory, as

	2019				2020				2021				Overall			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
General-Text																
GPT-oss (20B)	0.73	**0.57**	**0.59**	**0.58**	0.63	0.49	0.49	0.48	0.75	**0.52**	**0.52**	**0.52**	0.71	**0.53**	**0.54**	**0.53**
Qwen-2.5 (14B)	**0.78**	0.51	0.50	0.49	**0.77**	**0.52**	0.52	**0.52**	**0.84**	0.48	0.49	0.48	**0.80**	0.51	0.50	0.50
Mistral (7B)	0.73	0.50	0.50	0.50	0.68	**0.52**	**0.53**	0.51	0.80	0.51	0.51	0.51	0.74	0.51	0.51	0.51
LLaMa-3.2 (3B)	0.74	0.50	0.50	0.50	0.75	0.51	0.51	0.51	0.83	**0.52**	**0.52**	**0.52**	0.78	0.51	0.51	0.51
Gemma-3 (1B)	0.70	0.43	0.45	0.44	0.73	0.45	0.45	0.45	0.77	0.45	0.45	0.45	0.74	0.44	0.45	0.45
Finance-FT																
Fin-LLaMa3 (8B)	0.10	0.17	0.08	0.09	0.12	0.23	0.07	0.10	0.14	0.27	0.08	0.12	0.12	0.22	0.08	0.11
Fin-LLaMa (8B)	**0.70**	**0.46**	**0.47**	**0.46**	**0.77**	**0.47**	**0.47**	**0.47**	**0.77**	**0.46**	**0.46**	**0.46**	**0.74**	**0.47**	**0.47**	**0.47**
Finance-R1 (1.7B)	0.49	0.37	0.32	0.34	0.48	0.42	0.32	0.35	0.57	0.43	0.35	0.38	0.52	0.40	0.33	0.36
LLaMa-RAG (8B)	0.08	0.14	0.05	0.07	0.10	0.21	0.06	0.09	0.12	0.26	0.08	0.11	0.10	0.20	0.06	0.09
Image-Text																
GPT-5	0.57	**0.54**	**0.56**	0.51	0.52	**0.54**	**0.59**	0.46	0.71	0.54	**0.58**	0.53	0.62	**0.55**	**0.58**	0.52
GPT-5-mini	0.60	0.53	0.54	0.52	0.53	0.52	0.54	0.45	0.73	0.53	0.55	0.52	0.64	0.53	0.55	0.51
GPT-5-nano	0.61	0.53	0.54	0.52	0.48	0.52	0.54	0.43	0.73	0.52	0.53	0.52	0.62	0.52	0.54	0.50
GPT-4.1	0.66	0.53	0.54	**0.53**	0.57	0.53	0.56	0.48	0.79	0.54	0.55	**0.54**	0.69	0.53	0.55	**0.53**
Mistral-3.1 (24B)	0.72	0.49	0.50	0.49	0.72	0.51	0.52	**0.51**	0.86	0.52	0.51	0.50	0.77	0.51	0.51	0.51
Gemma-3 (12B)	0.66	0.50	0.50	0.50	0.65	0.52	0.54	0.50	0.83	**0.55**	0.54	**0.54**	0.73	0.52	0.53	0.52
LLaMa-3.2 (11B)	0.68	0.52	0.54	0.50	0.49	0.52	0.54	0.43	0.52	0.51	0.51	0.47	0.58	0.52	0.54	0.48
Sa2VA (8B)	0.72	0.43	0.46	0.44	0.77	0.45	0.46	0.45	0.80	0.44	0.45	0.45	0.76	0.44	0.46	0.45
Qwen-2.5 (7B)	**0.78**	0.52	0.51	0.48	**0.82**	0.47	0.49	0.47	**0.88**	0.45	0.49	0.47	**0.82**	0.49	0.50	0.48
Qwen-2 (7B)	0.70	0.43	0.45	0.44	0.71	0.49	0.48	0.48	0.81	0.49	0.49	0.49	0.74	0.47	0.47	0.47
Cosmos (7B)	0.75	0.44	0.48	0.45	0.79	0.49	0.49	0.49	0.86	0.47	0.49	0.48	0.81	0.47	0.49	0.47
LLaVa-1.6 (7B)	0.68	0.51	0.51	0.51	0.66	0.49	0.48	0.47	0.84	0.54	0.53	0.53	0.74	0.52	0.52	0.52
Audio-Text																
Voxtral (24B)	0.71	0.51	0.51	**0.51**	0.69	0.51	**0.52**	**0.50**	0.80	**0.53**	**0.53**	**0.53**	0.74	**0.52**	**0.52**	**0.52**
DeSTA-2.5 (8B)	0.54	0.50	0.49	0.47	0.52	0.50	0.49	0.43	0.55	0.50	0.49	0.43	0.54	0.50	0.49	0.45
Qwen-2.5 (7B)	**0.78**	0.46	0.50	0.45	**0.84**	**0.53**	0.51	**0.50**	**0.88**	0.51	0.50	0.48	**0.84**	0.51	0.50	0.47
Qwen-2 (7B)	0.62	0.43	0.42	0.43	0.69	0.46	0.44	0.45	0.67	0.45	0.42	0.43	0.66	0.44	0.43	0.43
Gemma-3n (8B)	0.35	**0.53**	**0.53**	0.35	0.32	0.50	0.50	0.31	0.33	0.48	0.46	0.31	0.34	0.50	0.51	0.33

Table 2: Model comparison on the FinCall-Surprise benchmark. The table is divided into four sections: General-Text refers to unimodal LLMs trained for general purposes; Finance-FT refers to unimodal LLMs finetuned on financial datasets; Image-Text refers to vision-language models that accept both images and text; and Audio-Text refers to audio-language models that accept both audio and text. For each metric and year, the highest score within each category is highlighted in **bold** . (Prec, Rec, and F1 represent Macro Average Precision, Macro Average Recall, and Macro Average F1-score, respectively.)

confirmed by its precision, recall, and F1 scores. For example, although Qwen-2.5 achieves 0.84 accuracy in 2021, its precision, recall, and F1 scores show no corresponding increase.

5.2 Balanced Predictions and Robustness of Closed-Source Models

Surprisingly, the four closed-source models evaluated, GPT-5, GPT-5-mini, GPT-5-nano, and GPT-4.1, achieved lower overall accuracy scores in the Image-Text category compared to smaller, open-source models. For instance, the highest accuracy among these closed-source models was 0.69, considerably lower than the 0.82 achieved by Qwen-2.5. However, we argue that this accuracy metric is misleading (discussed in section 5.1). A deeper analysis reveals that the GPT family, particularly GPT-5, consistently achieved the highest precision and recall scores nearly every year. This discrepancy arises from the models' prediction behavior. Manual inspection of the outputs showed that unlike many open-source models which predominantly default to a "Positive" label, the closed-source models generated a much more balanced distribution of "Positive" and "Negative" predictions. This avoidance of majority-class bias is what lowers their raw accuracy score. We interpret this as a positive sign, as it suggests the GPT models possess a more robust financial reasoning capability, allowing them to avoid the majority-class bias inherent in the dataset.

5.3 Failure Analysis of Finetuned Financial Models

While most models demonstrated reasonable performance on accuracy, a subset of models in the Finance-FT category performed exceptionally poorly. Notably, Fin-LLaMa3 and LLaMa-RAG

achieved overall accuracy scores of only 12% and 10%, respectively. A qualitative analysis of their outputs revealed three primary failure reasons: response bias, poor instruction following, and degraded language generation.

First, unlike most other models, these two models show a strong tendency to predict "negative" labels. The reason for this bias is unclear, but it may stem from the nature of their finetuning data. Second, the models exhibit difficulty following instructions. To evaluate model performance systematically, our benchmark requires answers in a specific format: `Result = [[POSITIVE or NEGATIVE]]`. However, the finetuned models often failed to comply, producing outputs in inconsistent formats. Although we attempted to accommodate these variations, some responses could not be parsed reliably. We treated such cases as incorrect, since the ability to follow task instructions should be considered an essential component of performance. Third, in several instances the models failed to generate grammatically correct or meaningful sentences. We suspect that both the instruction-following failures and the degradation in basic language generation stem from over-finetuning on highly specialized financial datasets. In pursuit of domain-specific performance, finetuning may inadvertently overpenalize general language capabilities, diminishing the model's broader usability. These findings highlight that when developing task-specialized models, it is important to ensure that this does not come at the expense of fundamental language competence and instruction-following ability.

5.4 Do Image and Audio Inputs Improve Model Performance?

As shown in Table 2, model performance varies significantly across categories. Interestingly, we observe both modest improvements and notable degradations in accuracy when additional modalities are introduced. For instance, within the Qwen-2.5 family, the General-Text model with 14B parameters achieves an overall accuracy of 0.80. When image data is incorporated, the 7B Image-Text model improves to 0.82. With audio data, the 7B Audio-Text model increases the accuracy to 0.84. While these gains may appear modest, the fact that smaller multimodal models outperform their larger text-only counterpart suggests that visual and auditory inputs can provide complementary signals beyond the raw conference call transcripts.

However, in other cases, the addition of new modalities leads to performance degradation. For example, the text-only LLaMa-3.2 model with 3B parameters achieves a respectable accuracy of 0.78, yet its larger Image-Text variant with 11B parameters performs significantly worse, dropping to 0.58. Audio-Text models also exhibit highly inconsistent performance. While Voxtral achieves 0.74 accuracy and Qwen-2.5 achieves 0.84 accuracy, other models like DeSTA-2.5 and Gemma-3n reach only 0.54 and 0.34, respectively. These negative results suggest that, for many state-of-the-art models, supplementary signals from presentation slides and audio may introduce noise that cannot be effectively understood with textual content, ultimately hindering predictive capabilities. More broadly, current multimodal models often lack the robustness and consistency that text-only LLMs demonstrate when applied to single-modal financial data.

Taken together, our results highlight both the potential and the limitations of multimodal learning in this domain. On the one hand, certain models clearly benefit from the additional information that pure text alone cannot provide. On the other hand, many existing multimodal models cannot interpret complex financial signals across modalities. We anticipate that future research leveraging our multimodal benchmark will unlock deeper insights and enable models to capture information that plain transcripts alone cannot convey.

6 Conclusions

In this work, we introduced FinCall-Surprise, the first large-scale, open-source, multimodal benchmark for earnings surprise prediction, comprising synchronized word-level transcripts, full audio recordings, and presentation slides from more than 2,600 corporate conference calls between 2019 and 2021. Our evaluation of 26 state-of-the-art models highlights both the potential and the limitations of current approaches. We find that many open-source models achieve deceptively high accuracy due to class imbalance, while closed-source models generally provide more balanced predictions. Financial models fine-tuned on narrow datasets often lose basic instruction-following and language generation abilities, and existing multimodal models continue to struggle with effectively integrating image and audio in the financial domain. We release FinCall-Surprise to reduce reliance on proprietary data and provide an open, challenging benchmark for advancing models in earnings surprise prediction.

Limitations

A potential limitation of FinCall-Surprise lies in the natural class imbalance between positive and negative earnings surprises. Because positive surprises occur more frequently in real-world financial markets, our dataset inevitably reflects this biased distribution. While this imbalance makes the benchmark realistic, it also poses challenges for model evaluation. Future research that employs FinCall-Surprise for training could consider techniques such as upsampling, downsampling, or other modern data-balancing strategies to mitigate this limitation and better assess model robustness across both positive and negative cases.

References

Khaled Alsabah. 2025. Love me do: Twitter likes and earnings surprise. *Journal of Behavioral Finance*, 26(3):283–302.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, and 1 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.

Sanjay W Bissessur and David Veenman. 2016. Analyst information precision and small earnings surprises. *Review of Accounting Studies*, 21(4):1327–1360.

Taha Borhani. 2024. Twitter financial news sentiment dataset. Kaggle. Accessed: 2025-09-13.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, Miami, Florida, USA. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Vasant Dhar and Dashin Chou. 2001. A comparison of nonlinear methods for predicting earnings surprises and returns. *IEEE Transactions on Neural networks*, 12(4):907–921.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4918–4927.

Jeffrey T Doyle, Russell J Lundholm, and Mark T Soliman. 2006. The extreme future stock returns following i/b/e/s earnings surprises. *Journal of Accounting Research*, 44(5):849–887.

Eli Bartov. 1992. Patterns in Unexpected Earnings as an Explanation for Post-Announcement Drift. *The Accounting Review*, 67(3):610–622. Publisher: American Accounting Association.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. 2023. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John C Heater, Ye Liu, Qin Tan, and Frank Zhang. 2025. Winning is not enough: Changing landscapes of earnings surprises and the market reaction. *Contemporary Accounting Research*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. 2023. Diego de las casas. *Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed*, pages 50–72.

KhazarAI. 2024. Personal-finance-r1. `https://huggingface.co/khazarai/Personal-Finance-R1`. Accessed: 2025-09-14.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Michael D Kimbrough. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1):189–219.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the association for computational linguistics: ACL 2023*, pages 8197–8209.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2024. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300.

Viet Dac Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. 2024. Sec-qa: A systematic evaluation corpus for financial qa. *arXiv preprint arXiv:2406.14394*.

Henry A Latane and Charles P Jones. 1979. Standardized unexpected earnings–1971-77. *The Journal of Finance*, 34(3):717–724.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*, pages 1–15.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070.

Qingwen Liang and Matias Carrasco Kind. 2025. How does managers' willingness to disclose affect analysts' earning forecasts-a measurement by llms. *Available at SSRN 5199752*.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Wanlong Liu, Junying Chen, Ke Ji, Li Zhou, Wenyu Chen, and Benyou Wang. 2024. Rag-instruct: Boosting llms with diverse retrieval-augmented instructions. *Preprint*, arXiv:2501.00353.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2025. Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*.

Cheng Luo, Enrichetta Ravina, Marco Sammon, and Luis M. Viceira. 2022. Retail investors' contrarian behavior around news, attention, and the momentum effect. Technical report, Social Science Research Network. Posted: 5 Apr 2020; Last revised: 26 May 2022.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, and 1 others. 2025. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*.

Vitaly Meursault, Pierre Jinghong Liang, Bryan R. Routledge, and Madeline Marco Scanlon. 2023. Pead.txt: Post-earnings-announcement drift using text. *Journal of Financial and Quantitative Analysis*, 58(6):2299–2326.

Income Numbers. 1968. An empirical evaluation of accounting. *Journal of Accounting Research*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Jane A Ou and Stephen H Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics*, 11(4):295–329.

Varpit Patel. 2021. Google stock data. Kaggle. Accessed: 2025-09-13.

James M Patell. 1976. Corporate forecasts of earnings per share and stock price behavior: Empirical test. *Journal of accounting research*, pages 246–276.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 785–788, New York, NY, USA. Association for Computing Machinery.

Rohan Rao. 2021. Nifty 50 stock market data (2000 - 2021). Kaggle. Accessed: 2025-09-13.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, Online. Association for Computational Linguistics.

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. 2025. Finchart-bench: Benchmarking financial chart comprehension in vision-language models. *arXiv preprint arXiv:2507.14823*.

Douglas J Skinner and Richard G Sloan. 2002. Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio. *Review of accounting studies*, 7(2):289–312.

tarun7r. 2025. tarun7r/finance-llama-8b: A llama 3.1 8b model fine-tuned on josephgflowers/finance-instruct-500k. https://huggingface.co/tarun7r/Finance-Llama-8B.

Gemma Team. 2025a. Gemma 3.

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

Qwen Team. 2025b. Qwen2.5-vl.

Manuel R Vargas, Beatriz SLP De Lima, and Alexandre G Evsukoff. 2017. Deep learning for stock market prediction from financial news articles. In *2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA)*, pages 60–65. IEEE.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. 2025. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint*.

Ziqiang Yuan, Kaiyuan Wang, Shoutai Zhu, Ye Yuan, Jingya Zhou, Yanlin Zhu, and Wenqi Wei. 2024. Finllms: A framework for financial reasoning dataset generation with large language models. *IEEE Transactions on Big Data*.

Cong Zhang and Zhenzhi He. 2025. Cross-sectional spillovers of earnings surprises and asset price anomalies. *Available at SSRN 5415255*.

Huopu Zhang, Yanguang Liu, and Mengnan Du. 2025. Sae-fire: Enhancing earnings surprise predictions through sparse autoencoder feature selection. *arXiv preprint arXiv:2505.14420*.

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

Yongqiong Zhu. 2020. Stock price prediction using the rnn model. *Journal of Physics: Conference Series*, 1650:032103.

Yu Zhu, Xiao Liu, and Olivia R Liu Sheng. 2025. Post-earnings-announcement drift prediction: Leveraging postevent investor responses with multitask learning. *Information Systems Research*.

Leon Zolotoy. 2012. Earnings surprise implicit in stock prices: which earnings forecasting models are investors using and what determines their choice? *Journal of Business Finance & Accounting*, 39(9-10):1161–1179.

A Prompt Used in Our Experiments

A.1 Prompt Used in Text-Only Modality

Text-only Modality

You are an expert equity research analyst. Your task is to read the following earnings-conference call transcript and decide whether next quarter's earnings per share (EPS) will **beat** or **miss** consensus estimates.

Instructions

- **Label only**: respond with exactly one of:
  - `POSITIVE` (indicating an expected positive EPS surprise)
  - `NEGATIVE` (indicating an expected negative EPS surprise)
- **Brief rationale** (1–2 sentences) explaining the key driver
- **No other text**

Definitions

- **POSITIVE**: indications of stronger-than-expected earnings or revenue growth, upbeat guidance, new partnerships, successful product launches, cost savings, market expansion, etc.
- **NEGATIVE**: indications of missed earnings or revenue declines, weak or withdrawn guidance, regulatory/legal setbacks, unexpected costs, competitive pressures, etc.

Transcript:
{transcript}

Answer format
Rationale = [[your rationale]]
Result = [[POSITIVE or NEGATIVE]]

A.2 Prompt Used in Image-Text Modality

Multi-modal Modality

[image_1, image_2, image_3]
You are an expert equity research analyst. Your task is to read the following earnings-conference call transcript and the accompanying slide deck images to decide whether next quarter's earnings per share (EPS) will **beat** or **miss** consensus estimates.

Instructions

1. **Label only**: respond with exactly one of:
   - `POSITIVE` (indicating an expected positive EPS surprise)
   - `NEGATIVE` (indicating an expected negative EPS surprise)
2. **Brief rationale** (1–2 sentences) explaining the key driver
3. **No other text**

Definitions

- **POSITIVE**: indications of stronger-than-expected earnings or revenue growth, upbeat guidance, new partnerships, successful product launches, cost savings, market expansion, etc.
- **NEGATIVE**: indications of missed earnings or revenue declines, weak or withdrawn guidance, regulatory/legal setbacks, unexpected costs, competitive pressures, etc.

Transcript:
{transcript}

Answer format
Rationale = [[your rationale]]
Result = [[POSITIVE or NEGATIVE]]

A.3 Prompt Used in Audio-Text Modality

Audio-only Modality

{audio}
You are an expert equity research analyst. Your task is to analyze the provided audio from an earnings conference call and predict whether next quarter's earnings per share (EPS) will **beat** or **miss** consensus estimates.

Instructions

1. **Label only**: respond with exactly one of:
   - `POSITIVE` (indicating an expected positive EPS surprise)
   - `NEGATIVE` (indicating an expected negative EPS surprise)
2. **Brief rationale** (1–2 sentences) explaining the key driver
3. **No other text**

Definitions

- **POSITIVE**: indications of stronger-than-expected earnings or revenue growth, upbeat guidance, new partnerships, successful product launches, cost savings, market expansion, etc.
- **NEGATIVE**: indications of missed earnings or revenue declines, weak or withdrawn guidance, regulatory/legal setbacks, unexpected costs, competitive pressures, etc.

Answer format
Rationale = [[your rationale]]
Result = [[POSITIVE or NEGATIVE]]

B Examples of Model Responses

GPT OSS 20B

<|channel|>analysis<|message|> We need to decide if next quarter EPS will beat or miss consensus. Transcript is about acquisition of Celgene by BMS and EPS accretion announced, many positive statements, expected near-term launches, synergy. But we need to see if EPS guidance they gave earlier? They said "EPS guidance we issued this morning." But the details not present. However they mention strong EPS accretion day one, and near-term product launches. So likely positive. There's no mention of negative factors or earnings decline. Therefore predict positive EPS beat. Provide rationale: acquisition and near-term launches with high revenue potential and EPS accretion. Thus final answer: POSITIVE. Provide rationale in 1-2 sentences: e.g., acquisition creates immediate EPS accretion and multiple near 2011 term launches generating >$15B potential, likely beating consensus.<|end|><|start|> assistant <|channel|> final <|message|> Rationale = [[The acquisition of Celgene is positioned to deliver immediate EPS accretion and the company expects six near 2011 term launches with over $15B of peak sales, creating strong upside for earnings beyond current guidance.]] Result = [[POSITIVE]]<|return|>

Qwen 2.5 14B

Rationale = [[Alexion's earnings conference call indicates strong momentum with the launch of MG, successful pipeline advancements, including early approval for Ultomiris and promising data for NMO, and strategic acquisitions like Wilson Therapeutics and Syntimmune. The company plans to convert PNH patients to Ultomiris, expand into additional indications like ALS and PPMS, and maintain a focus on disciplined business development. These factors suggest robust growth prospects. Additionally, Alexion's emphasis on long-term sustainability and volume growth through strategic pricing aligns positively with investor expectations.]] Result = [[POSITIVE]]

Mistral 7B

Rationale = Stronger-than-expected earnings growth due to the creation of a leading biopharma company with a broad and deep pipeline, leading marketed products portfolio, and multiple near-term product launches. Result = POSITIVE

LLaMa 3.2 3B

Result = POSITIVE Rationale = The executives provided strong guidance on the upcoming launches of the combined company, including 6 near-term product launches with significant revenue potential. They also highlighted the pipeline's broad and deep nature, with a focus on long-term growth opportunities and the potential for value creation in the medium and long term.

Gemma 3 1B

Rationale = [[The transaction represents a significant strategic realignment for Bristol Myers Squibb, positioning the company as a leader in oncology and immunology, leveraging the strengths of Celgene2019s pipeline and expanding its access to novel therapies. The combination offers substantial growth potential, with the combined company projected to deliver strong returns for shareholders.]] Result = [[POSITIVE]]

Finance-LLaMa3 8B

We're how we see the company we see the product we see the product we see the company

Finance-LLaMa 8B

Assistant: POSITIVE. Explanation: next quarter 2019s earnings per share (EPS) will beat consensus estimates. The combination of Bristol-Myers Squibb and Celgene will create a leading biopharma company with strong potential for growth. The combined portfolio will have leading franchises in oncology, immunology and inflammation, and cardiovascular disease. The combined company will have a robust late-stage pipeline and early-stage pipeline across disease areas, as well as cutting-edge technologies and discovery platforms. The transaction is financially compelling on day one, with strong returns and immediate EPS accretion. The combined company will have significant growth opportunities, including 6 near-term product launches with more than $15 billion in revenue potential. The combined company will have an industry-leading late-stage pipeline, with 6 near-term product launches, and a robust early-stage pipeline across disease areas. The combined company will also have cutting-edge technologies and discovery platforms that will enable the acceleration of new medicines for patients. The combined company will have strong potential for growth, with leading franchises in oncology, immunology and inflammation, and cardiovascular disease. The combined company will have significant growth opportunities, including 6 near-term

product launches and a robust early-stage pipeline. The combination of Bristol-Myers

Finance-R1

<think> Okay, let's tackle this. The user wants to know if next quarter's EPS will beat or miss consensus estimates based on the earnings conference call transcript provided. First, I need to scan through the transcript for any mentions of EPS, revenue growth, cost savings, market expansion, or other factors that would indicate a positive or negative surprise. Looking at the transcript, the main points seem to be about the merger between Bristol-Myers Squibb and Celgene. They mention creating a leading biopharma company, combining their assets, and the strategic benefits. There are talks about expanding their portfolio, new product launches (like Revlimid, Pomalyst, and others), and the potential for growth in oncology and immunology areas. The executives talk about the financial implications of the deal, stating that they'll deliver strong returns with immediate EPS accretion. They also mention the combined pipeline with potential revenue from 6 near-term product launches. There's a lot of emphasis on the synergy and the long-term growth opportunities. However, the transcript doesn't mention any actual EPS numbers or revenue figures for the current quarter. The focus is more on the strategic moves and future outlook rather than past or current EPS performance. The key drivers here are the strategic alignment, combined pipeline potential, and the expected revenue growth from new products. The definition of POSITIVE is when earnings or revenue growth is stronger than expected, or there's new partnerships, successful product launches, cost savings, etc. The transcript mentions new product launches, expanded portfolio, and the potential for growth in key areas. Since these are positive factors, it's likely that the EPS will beat expectations. The NEGATIVE would be if earnings or revenue declines are expected, but the transcript doesn't mention any declines. The focus is on growth and positive outcomes. So the key driver here is the strategic expansion and the potential for strong revenue from the new product launches and expanded portfolio, which would lead to a positive EPS surprise. </think> Rationale = [The transcript emphasizes strategic alignment, expanded portfolio, and high-growth potential from combined pipelines, including 6 near-term product launches. These factors suggest strong revenue and EPS growth, indicating a positive surprise.] Result = [POSITIVE]

LLaMa-RAG

Good a good an the the the and **Fees to have a good a positive to and to ** to the (and to come the positive (best (1 the negative The. We the ** to have an overall to come and **to the positive the 2018 to come the **the same to come the ** to come the best of the 2018 and ** to come the best of the ** (Positive the positive the next year we the to a good a good (the negative and the negative the 1 ** the positive the **to the **the. We **the negative the higher the positive the positive (the (negative. We the 1. **the positive (Best of a positive the we, **the positive (1 **the positive and **the same the **the negative to come the positive the positive (1 **the negative the negative the positive (negative Following the The The We the negative the best (the positive (The positive, positive (negative the positive, the negative The positive to come the Positive ** to come the positive the positive, and Positive the positive, positive (negative the negative the following an overall the

Example of Response Bias

Transcript:

We expect the transaction will bring significant financial benefits to shareholders of both companies from day one. We believe we'll generate returns in excess of both companies cost of capital and deliver more than 40% accretion in the first full year. From a balance sheet perspective, we will remain in a very strong position. We project substantial free cash flow in excess of $45 billion in the first 3 years...

Finance-R1 Output:

... The company faces potential risks in the global market and slight uncertainty in supply chains, which could impact future margins ... Result = [NEGATIVE]

Example of Poor Instruction Following

Finance-R1 Output:

... The executives seem very confident about the new product launch. The sentiment is bullish and I expect them to beat the estimates. Therefore, the next quarter earning surprise is positive.