Decoding the Market's Pulse: Context-Enriched Agentic Retrieval Augmented Generation for Predicting Post-Earnings Price Shocks

Chenhui Li
Millennium Management LLC
Chenhui.Li2@mlp.com

Weihai Lu[*]
Peking University
luweihai@pku.edu.cn

Abstract

Accurately forecasting large stock price movements after corporate earnings announcements is a longstanding challenge. Existing methods–sentiment lexicons, fine-tuned encoders, and standalone LLMs–often **lack temporal-causal reasoning** and are prone to **narrative bias**, echoing overly optimistic managerial tone. We introduce **Context-Enriched Agentic RAG (CARAG)**, a retrieval-augmented framework that deploys a team of cooperative LLM agents, each specializing in a distinct analytical task: evaluating historical performance, assessing the credibility of guidance, or benchmarking against peers. Agents retrieve structured evidence from a Causal-Temporal Knowledge Graph (CTKG) built from financial statements and earnings calls, enabling grounded, context-rich reasoning. This design mitigates LLM hallucinations and produces more objective predictions. Without task-specific training, our system achieves state-of-the-art zero-shot performance across NASDAQ, NYSE, and MAEC datasets, outperforming both larger LLMs and fine-tuned models in macro-F1, MCC, and Sharpe, beating market benchmarks (S&P 500 and Nasdaq) for the same forecasting horizon. The complete code, datasets and prompts are available at https://github.com/luweihai/CARAG.

1 Introduction

Corporate earnings calls release a deluge of information that profoundly influences financial markets (Price et al., 2012). These events, featuring management's results and guidance, trigger significant single-day stock price movements. The rich, unstructured data in these calls–including remarks, Q&A, and forward-looking statements–presents a formidable predictive modeling challenge (Chen et al., 2022; Wu et al., 2023).

With the remarkable success of deep learning in various other tasks (Lu et al., 2025; Xie et al., 2026, 2025a; Lu and Yin, 2025; Cui et al., 2025), an increasing body of research has integrated deep learning architectures into the domain of stock price prediction based on earnings calls. As Transformer models (Zhu et al., 2026) and pre-training techniques achieve significant breakthroughs across diverse fields (Tong et al., 2025; Zhang et al., 2025), existing methods in forecasting these price shocks have evolved from lexicon-based sentiment analysis (Loughran and McDonald, 2011) to domain-specific transformer models like FinBERT (Araci, 2019), and hierarchical transformers (Koval et al., 2023). Recent literature shows that graph-based encoders (Medya et al., 2022; Liu et al., 2024) excel in mapping hidden relationships between stocks networks. While these models improved contextual understanding, they cannot perform sequential logical deduction or fuse insights from heterogeneous sources. More recent approaches using Large Language Models (LLMs) have shown promise to synthesize information from disparate document types (Chang et al., 2023; Ni et al., 2024). However, when processing an earnings call transcript in isolation, existing methods, including standalone LLMs, face two fundamental challenges:

- **Deficiencies in Temporal and Causal Reasoning.** Standalone models are unable to place new information in its proper historical context. For instance, they cannot contextualize a firm's reported 16% sales growth against its 55% growth in prior quarters, as was the case for Five Below in Q4 2021. Furthermore, they lack the capacity to independently verify management's claimed drivers for performance, often accepting narratives that are intentionally managed to influence perception (Li, 2010). This absence of structured reasoning (Cao et al., 2023), renders the model

[*] Corresponding author.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3055–3073
March 24-29, 2026 ©2026 Association for Computational Linguistics

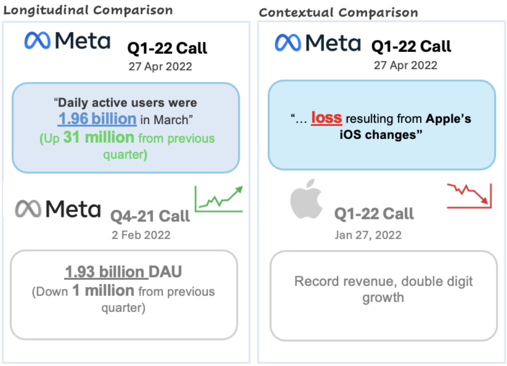

Longitudinal Comparison
Contextual Comparison
Meta Q1-22 Call
27 Apr 2022
"Daily active users were 1.96 billion in March"
(Up 31 million from previous quarter)
Meta Q4-21 Call
2 Feb 2022
1.93 billion DAU
(Down 1 million from previous quarter)
Meta Q1-22 Call
27 Apr 2022
"... loss resulting from Apple's iOS changes"
Q1-22 Call
Jan 27, 2022
Record revenue, double digit growth

Figure 1: **Left: Longitudinal comparison** contrasts Meta's Q1-22 statement with prior quarters **Right: Contextual comparison** links Meta's reported losses to Apple's Q1-22 call, grounding causal explanations in external disclosures.

incapable of distinguishing genuine corporate strength from a mere deceleration in growth.

- **Monolithic Reasoning and Narrative Bias.** A single model processes information through a monolithic lens, making it highly susceptible to polished narratives crafted by management and inducing optimistic bias (Cui and Zhai, 2023). It lacks the dynamic, critical evaluation of human experts, who synthesize insights from multiple angles–such as historical performance and peer comparison–to form a robust, balanced judgment.

To overcome these limitations, we introduce **Context-Enriched Agentic RAG (CARAG)**, a framework that mirrors the division of labor in fundamental investing—specialists collaborate on intrinsic valuation and discount-rate work, peer/relative valuation, and sector-level synthesis—by coupling multi-agent orchestration with a structured knowledge base (Kadan et al., 2012; Bradley et al., 2017; Groysberg et al., 2013).

We first construct a **Causal-Temporal Knowledge Graph (CTKG)** from historical financial statements and past earnings calls to serve as the long-term memory. Then, a swarm of specialist LLM agents collaborates to analyze new earnings calls. Each agent is assigned a distinct directive, which evaluates historical performance, assesses peer-level benchmarks, or reviews past guidance, querying CTKG to ground their analysis. A final synthesizer agent integrates these diverse analytical memos to produce a single, context-aware prediction.

The primary contributions of this work are as follows:

1. We propose the **Context-Enriched Agentic RAG (CARAG)** framework. To our knowledge, this is the first study in this field that integrates causal reasoning and temporal inference to enhance the accuracy of stock price prediction.

2. We propose the **Causal-Temporal Knowledge Graph (CTKG)**, which encodes financial data with temporal and causal structure. We also introduce a **Multi-Agent Reasoning Framework** that analyzes earnings calls from diverse perspectives to reduce mono-model reasoning bias.

3. We release a curated dataset linking thousands of earnings call transcripts with matched financial statements and multi-horizon price movements, **forming a reproducible benchmark for NLP-finance research.**

4. On NASDAQ, NYSE, and MAEC datasets, our zero-shot framework achieves state-of-the-art results, outperforming large LLMs, fine-tuned encoder models, and audio-text alignment models in predicting price shocks. Furthermore, our 2-year backtest yields an annualized return of **36 percent** and a **Sharpe of 1.64** on a sample of 3158 calls.

2 Related Work

Earnings Call Price Prediction Predicting post-earnings announcement drift (PEAD)–a market anomaly where stock prices drift following an earnings surprise (Ball and Brown, 1968)–have evolved from early methods using sentiment dictionaries (Stone et al., 1966; Loughran and McDonald, 2011) and engineered transcript features (Chin and Fan, 2023) to more sophisticated machine-learning architectures (Peng, 2025; Wang et al., 2025).

Transformer-based encoders like FinBERT advanced this analysis but were constrained by input length, necessitating the truncation of long earnings call transcripts (Araci, 2019; Huang et al., 2024). Hierarchical models overcame this by segmenting transcripts (Koval et al., 2023; Zaheer et al., 2020). Other modern approaches leverage multitask learning to predict post-announcement drift (Zhu et al., 2025) and integrate multimodal data sources for intra-call predictions (Ghosh et al., 2025).

More recently, graph-based models such as StockGNN and ECHO-GL have been developed to capture inter-firm relationships by creating concept networks from call transcripts (Medya et al., 2022; Liu et al., 2024).

Nevertheless, existing methods struggle with multi-step reasoning and cross-source synthesis. Our framework fuses earnings-call narratives with structured financial data to enable temporal and causal inference beyond monolithic models.

LLM Agents Parallel to the evolution of Large Language Models (LLMs), multi-agent frameworks have become increasingly prominent, demonstrating significant efficacy across a diverse range of applications (Zhang et al., 2025; Chen et al., 2024; Wei et al., 2025a; Zhao et al., 2025; Wei et al., 2025b; Xie et al., 2025b). LLM-driven earnings-call prediction is at its nascent stage. Early research (Chang et al., 2023) shows a raw ChatGPT sentiment score in the [–10, 10] range yields predictive power in forecasting the drift. Further research showed improvement in accuracy when LLMs in wide context jointly read transcripts, ratios, and prices (Ni et al., 2024), but these studies exclusively use single shot prompts - without applying retrieval, iteration, or agent cooperation.

Extensive research has explored the benefits of cooperative agents equipped with shared memory and feedback mechanisms, both between and within agents (Talebirad and Nadiri, 2023; Park et al., 2023; Han et al., 2024; Guo et al., 2024; Zhang et al., 2024; Gao et al., 2024; Xu et al., 2025; Ji et al., 2025). Recent studies in other domains, such as misinformation, have shown that recreation-augmented, role-specialized multi-agent frameworks consistently outperform standalone LLMs (Li et al., 2024). Systems such as OpenAGI divide claims into subtasks, share the retrieved context, and achieve state-of-the-art fact checking without supervision (Zhang and Gao, 2023; Ge et al., 2024). Multi-hop retrievers (MUSER) and graph-aware GET further improve evidence grounding (Liao et al., 2023; Xu et al., 2022). In the financial domain, GraphRAG (Barry et al., 2025) shows gains in leveraging knowledge graphs with classical textual retrieval to reduce hallucinations and improve efficiency in finance-centric QA tasks.

3 Methodology

3.1 Data Preparation of Inputs and Targets

Data Ingestion Our pipeline ingests two data streams for each firm–quarter pair: first, we ingest **full transcripts** (prepared remarks *and* Q&A) for NYSE and NASDAQ firms, scraped from *Seeking Alpha*, we also ingest **structured data – financial statements** : recent *income, balance-sheet* and *cash-flow* statements spanning multiple quarters, parsed from publicly available 10-K and 10-Q filings on the SEC's EDGAR database. The transcripts and statement data are strictly sorted by their publication dates. When the agent analyzes quarter t, it has access *only* to transcripts and fundamentals dated $\leq t$. This ensures every prediction is made with a realistic information set, eliminating look-ahead bias. Statistics and details of the fundamentals data can be found in I.

Prediction Target For an earnings call on trading day T, we define the k-day forward return (for $k \in \{1, 3, 7, 15, 30\}$) as $r_{T-1:T+k} = (P^{\text{close}}_{T+k}/P^{\text{close}}_{T-1}) - 1$. The prediction target is the sign of this return, $y = \text{sgn}(r_{T-1:T+k}) \in \{-1, 1\}$. We task the Synthesizer Agent with making a direct, zero-shot prediction for enhanced explainability. It outputs a confidence score $s \in [0, 10]$. This score is then converted to a binary prediction: 1 for scores greater than 5, and -1 for scores of 5 or less. We evaluate accuracy by comparing $\hat{y}$ with y.

3.2 Mitigation of LLM Look-Ahead Bias

LLMs risk *hindsight bias*, where later knowledge colors earlier interpretations—for instance, labeling NVIDIA's 2019 call as bullish due to Ampere's eventual success, an outcome unknowable at the time.

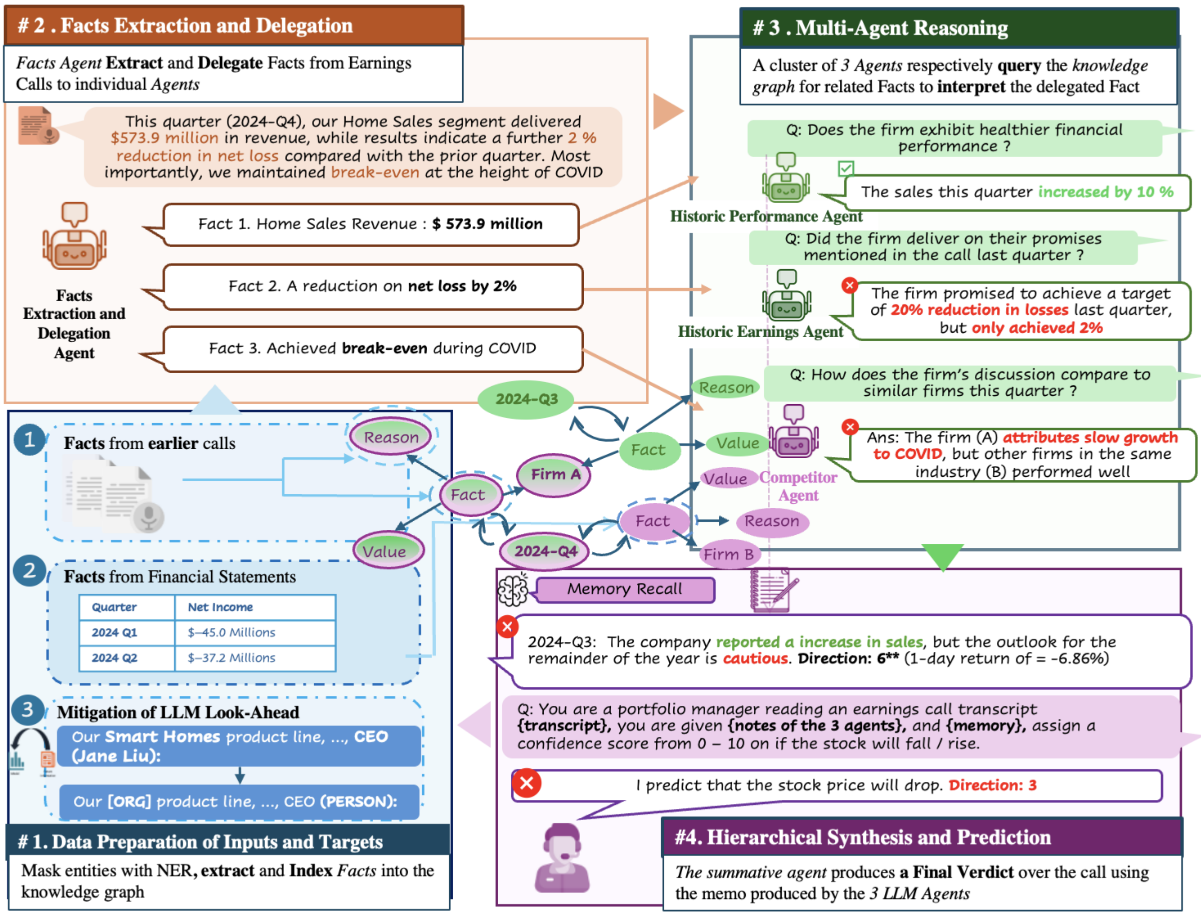
We mitigate leakage via **named entity masking**, replacing tokens for firms and products (e.g., *NVIDIA*, *Ampere*) with placeholders so the model relies on fundamentals rather than brand or future-informed cues (see Figure 2).

To address temporal leakage, we also test on 2,259 post-cutoff transcripts (Aug 2023–Dec 2024), ensuring results are not biased by knowledge beyond the LLM's training window (Appendix A).

3.3 Facts Extraction and Delegation

3.3.1 Facts Extraction

Traditional knowledge graph pipelines that produce static triples (for example, *CompanyA*, `has_CEO`,

# 2 . Facts Extraction and Delegation

Facts Agent **Extract** and **Delegate** Facts from Earnings Calls to individual *Agents*

This quarter (2024-Q4), our Home Sales segment delivered $573.9 million in revenue, while results indicate a further 2 % reduction in net loss compared with the prior quarter. Most importantly, we maintained break-even at the height of COVID

Fact 1. Home Sales Revenue : **$ 573.9 million**

Fact 2. A reduction on **net loss by 2%**

Fact 3. Achieved **break-even** during COVID

3 . Multi-Agent Reasoning

A cluster of *3 Agents* respectively **query** the *knowledge graph* for related Facts to **interpret** the delegated Fact

Q: Does the firm exhibit healthier financial performance ?

The sales this quarter increased by 10 %

Q: Did the firm deliver on their promises mentioned in the call last quarter ?

The firm promised to achieve a target of 20% reduction in losses last quarter, but only achieved 2%

Q: How does the firm's discussion compare to similar firms this quarter ?

Ans: The firm (A) attributes slow growth to COVID, but other firms in the same industry (B) performed well

1 Facts from **earlier** calls

2 Facts from Financial Statements

Quarter	Net Income
2024 Q1	$-45.0 Millions
2024 Q2	$-37.2 Millions

3 Mitigation of LLM Look-Ahead

Our **Smart Homes** product line, ..., **CEO (Jane Liu):**

Our **[ORG]** product line, ..., CEO **(PERSON):**

1. Data Preparation of Inputs and Targets

Mask entities with NER, **extract** and **Index** *Facts* into the knowledge graph

Memory Recall

2024-Q3: The company reported a increase in sales, but the outlook for the remainder of the year is **cautious**. **Direction: 6**** (1-day return of = -6.86%)

Q: You are a portfolio manager reading an earnings call transcript **{transcript}**, you are given **{notes of the 3 agents}**, and **{memory}**, assign a confidence score from 0 – 10 on if the stock will fall / rise.

I predict that the stock price will drop. **Direction: 3**

#4. Hierarchical Synthesis and Prediction

The summative agent produces **a Final Verdict** over the call using the memo produced by the *3 LLM Agents*

Figure 2: **CARAG**. Modules cover retrieval, alignment, multi-agent reasoning, and prediction

John Doe) are ill-suited for the financial domain, as they overlook the two most critical categories of information in earnings calls: time-varying dynamics and causal claims. Statements like “this quarter’s profit margin declined sequentially” or “revenue increased due to our new product launch” are exactly the evidence analysts use to assess the value of a firm. To capture this information, we develop a **Causal-Temporal Knowledge Graph (CTKG)** populated by a dedicated large-language model.

Compared to a simple event index or a time-series database, the unique advantage of the CTKG is that it not only stores events and metrics but explicitly captures *causal relationships and temporal dependencies* between events. This structured semantics allows agents to perform deeper multi-hop reasoning. In contrast, traditional time-series databases merely store facts and lack the relational structure necessary for reasoning.

A specialized **Facts Extraction and Delegation Agent** orchestrates the creation of facts by parsing unstructured text into structured output. A Fact is a 5-tuple: $f = (s, q, \tau, \mu, \nu)$, whereby s stands for ticker , q stands for quarter, τ is the fact’s type, μ is the financial metric and ν is its value. Each Fact is categorized by its nature–*Result*, *Forward-Looking*, *Risk Disclosure*, *Sentiment*, or *Macro*–and contains the specific `metric` (e.g., “Revenue”) and its `Value`. Specifically, *Result* facts capture a concrete metric stated in an earnings-call transcript or financial statement. We use financial statements to supplement *Result* facts with quarter-over-quarter (QoQ) or year-over-year (YoY) change.

Each `Fact` node is linked to its corresponding `Ticker` and `Quarter` nodes. As shown in the knowledge graph in the center of fig. 2, the Historic Performance and Historic Earnings Agents anchor on `Ticker A` and traverse the graph **longitudinally** backward along the green-colored timeline of successive `Quarter` nodes (from 2024-Q4 to 2024-Q3) to construct a complete performance history. The **Competitor Agent** starts instead from a focal `Quarter` node, (`2024-Q4`), and fans out laterally through the purple peer links to all tickers that share that quarter (`Firm B`). The schema for the facts and the corresponding knowledge graph can be found in the Appendix fig. 7. More examples of facts can be found in J.

To model causality, a `Fact` is connected to a separate `Reason` node via an `HAS_REASON` relationship. This `Reason` node serves a crucial dual function. First, the node may store the exact justification provided by the company's management during an earnings call. For instance, if management states there is a "Positive outlook for continued growth", that exact phrase is captured as a Reason. Second, the node can store the LLM's own analytical interpretation. For example, the LLM may paraphrase "Phenomenal week learning from customers" in the earnings call to a structured `Reason`: 'Indicates a positive sentiment towards customer relationships and feedback.'

Each Fact node is stored in Neo4j, and its 1,536-dimensional embedding is indexed for semantic similarity searches. By separating the **what** (the `Fact` node) from the **why** (the `Reason` node), we encourage diversity in analysis. It allows agents to query for all firms reporting a certain metric, thereby enabling a meta-analysis of how different management teams frame similar results.

3.3.2 Facts Delegation

Let each extracted fact be represented as $f_i = (\tau_i, c_i)$. Define a routing function: $\phi : \mathcal{F} \longrightarrow \mathcal{P}(\mathcal{A})$, $\mathcal{A} = \{A_{\text{HP}}, A_{\text{HE}}, A_{\text{COMP}}\}$ that maps each fact to one or more downstream agents. Each downstream agent works with a pre-filtered, highly relevant stream of information. For example, a fact of type "Result" (e.g., a reported revenue figure) would be delegated to the **Historic Performance Agent**, which compares it to the firm's own past results. In contrast, a "Forward-Looking" fact, such as a management projection, would be routed to the **Historic Earnings Agent**, tasked with assessing the credibility and groundedness of managerial claims, and such facts are prime candidates for its scrutiny. Similarly, a "Risk Disclosure" fact is sent to the **Competitor Agent** to determine if the risk is firm-specific or sector-wide, and to the Historic Earnings Agent to evaluate the specificity and transparency of the disclosure.

3.4 Multi-Agent Reasoning

Operating in parallel, a cluster of three specialist LLM agents query the **CTKG** to produce analytical memoranda (*Memos*), which serve as inputs for a final synthesis stage.

3.4.1 Historic Performance Agent.

The design of this agent is motivated by the principle that the intrinsic value of a firm, as reflected in its financial statements, serves as a long-term anchor for its stock price (Ou and Penman, 1989).

For an extracted fact as the 5-tuple:$f = (s, q, \tau, \mu, \nu)$, with semantic embedding is $v \in \mathbb{R}^d$. Let $\mathcal{H}(s, \mu) = \{\, f' = (s, q', \tau', \mu, \nu') : q' < q \}$, be the history of the same ticker–metric pair. The agent defines cosine similarity $\sigma(v, v_\tau) = \langle v, v_\tau \rangle / (\|v\| \|v_\tau\|)$. It then retrieves k most similar historical records

$$\mathcal{N}_k(f) = \underset{h \in \mathcal{H}(s,m)}{\arg \overset{k}{\text{top}}} \ \sigma(v, v_h).$$

For a new fact $f = (s, q, \textsc{Result}, \mu, \nu)$ let $f^{-1} = (s, q{-}1, \textsc{Result}, \mu, \nu^{-1})$ be the recent prior fact having Type = Result with the same ticker s and metric μ.

Using the historical facts, the agent then produces a structured comparison that flags trend shifts, contextualizing a firm's current reported results against its own historical performance. The agent computes the period-on-period change with value ν: $\Delta\nu = \nu - \nu^{-1}$, $\delta = \frac{\Delta\nu}{\nu^{-1}}$, then, the agent scrutinizes reasons behind the change. Define the mapping $rho : \mathcal{F} \longrightarrow \mathcal{R}$ which assigns to each extracted fact $f \in \mathcal{F}$ its corresponding Reason node $r = \rho(f) \in \mathcal{R}$. The agent assesses tuple $(\delta_t,\ r)$, by inspecting the reason r behind performance, it flags instances where management's commentary fails to reconcile the current figure with its historical benchmark. Citing an evidence in our corpus, when *Codexis* (CDXS) reported a FY-2019-Q4 gross margin of 47 %, the agent surfaced the prior-year figure of 51 % and correctly highlighted a profitability erosion despite double-digit top-line growth.

3.4.2 Historic Earnings Agent.

Grounded in classic attribution theory (Heider, 1958; Kelley, 1967), this agent operates on the premise that management claims backed by verifiable, quantitative evidence are perceived as more credible, and grounded statements elicit stronger market reactions and analyst revisions than unsupported rhetoric (Mayew et al., 2015).

Similar to the Historic Performance agent, the Historic Earnings agent retrieves the k most similar facts but differently, it retrieves all types of facts, not limited to Type = Result. The Historic Earnings

Agent is tasked with evaluating whether forward-looking guidance from previous calls was beaten, or missed. Citing another example, in *Newmark* (NMRK) Q1-2021, management guided to $20\,\%$ to $25\,\%$ FY revenue growth yet delivered only $4.1\,\%$ in the 1st quarter; the agent flagged a high risk of under-delivery.

3.4.3 Competitor Analyst Agent.

A substantial body of empirical finance literature establishes that industry and sector affiliations are primary drivers of return comovement. Studies show that factors in the global industry can explain more cross-sectional variance than geography (Cavaglia et al., 2000; Baca et al., 2000; Heston and Rouwenhorst, 1994; Moskowitz and Grinblatt, 1999).

Drawing on this principle, the Competitor Analyst Agent is designed to evaluate a firm's performance not in isolation, but in direct comparison to its peers. Let $\mathcal{U}$ be the universe of tradable firms and $\mathcal{S} = \{s_1, \ldots, s_{11}\}$ the GICS sectors, with each firm $i \in \mathcal{U}$ tagged by $\sigma(i) \in \mathcal{S}$. For a focal firm $i^\star$, its *sector peer set* is $\mathcal{P}(i^\star) = \{\, j \in \mathcal{U} \setminus \{i^\star\} \;:\; \sigma(j) = \sigma(i^\star)\}$. Consider the focal fact $f^\star = (i^\star, q^\star, \mu, v^\star)$ with metric μ. Its *predicate class* collects all peer-firm facts for the same quarter,

$$p(f^\star) = \bigl\{ f' = (j, q^\star, \mu, v_j) \in \mathcal{F} \;:\; j \in \mathcal{P}(i^\star) \bigr\}.$$

From this class the agent retrieves the k most cosine-similar facts. It then forms the sector benchmark and computes the firm's performance deviation from the sector benchmark. The agent also analogously derives information from verbal descriptions and sentiment based facts.

3.5 Hierarchical Synthesis and Prediction

To integrate the specialist agents' outputs, we introduce a higher-level *synthesizer agent* that performs meta-level reasoning. The synthesizer ingests the analytical memos produced by the 3 agents and maintains a rolling memory

$$\mathcal{H}_t \;=\; \bigl\{\, (\hat{y}_{t-k},\, r_{t-k},\, u_{t-k}) \bigr\}_{k=1}^{K},$$

where $\hat{y}_{t-k} \in \mathbb{R}$ is the forecast it issued k quarters earlier, r_{t-k} is the realised excess return, and u_{t-k} is the accompanying textual verdict. From this information the synthesizer produces two sentences: (i) a **summary** that concisely states supporting evidence; and (ii) a **direction** together with a confidence score $s_t \in [0, 10]$, where $s_t = 0$ denotes the strongest conviction of a price decline and $s_t = 10$ denotes the strongest conviction of a price increase.

4 Experiments

4.1 Experimental Design

Dataset Curation To support full reproducibility, we curate and release two new corpora of earnings calls for companies listed on the **NASDAQ** and **NYSE**, each paired with their corresponding financial statements. We use these datasets to benchmark our approach against **text-only** baselines. They are provided in the supplementary material.

To evaluate the predictive power of financial fundamentals against audio signals, we benchmark our framework against state-of-the-art audio-text alignment models on the MAEC dataset (Li et al., 2020). We replace the dataset's audio tracks with our quarter-matched financial statements, holding all other settings constant. This isolates the performance contribution of **financial-statement grounding** relative to **acoustic alignment**. Descriptive statistics for all datasets are provided in table 1.

Baseline Model Families We benchmark against two **text-only** baseline families across the NASDAQ and NYSE datasets: **(1) Zero-shot LLM baselines with transcript-only input**: These include (a) the Loughran–McDonald dictionary, (b) GPT-4o-mini, (c) GPT-3.5-turbo, (d) a two-agent LangGraph pipeline (Wang and Duan, 2024) that simulates analyst–critic interaction, and (e) a Graph-of-Thoughts (GoT) variant (Zhou et al., 2023) that expands reasoning by generating diverse hypotheses and selecting the top-ranked output. These baselines use no structured financial context and operate solely on call transcripts. Lastly, we ablate against a (f) wide-context ablation baseline: a strong baseline that provides GPT-4o-mini with both transcripts and financial statements as input.

(2) Supervised text encoder baselines: These include (a) FinBERT (Huang et al., 2024), a 110M-parameter BERT variant trained on financial filings and transcripts, and (b) Hierarchical FinBERT (Koval et al., 2023), which segments long transcripts into sentence-aligned chunks (with $L \in 32, 64, 128$), encodes them via FinBERT, and aggregates using a lightweight Transformer.

(3) Audio-text alignment models: The third family of baselines comprises multimodal systems designed to integrate acoustic and textual signals

Table 1: Descriptive statistics of our curated NASDAQ and NYSE corpora and the MAEC corpora

Exchange	**Date span**	**# Transcripts**	**# Tickers**	$T_{\min}$	T_{med}	$T_{\max}$
NASDAQ	2019-05-23 – 2023-02-09	1,772	507	1,416	7,059	8,061
NYSE	2019-05-21 – 2023-02-22	1,386	470	771	7,006	7,924
MAEC	2015-02-25 – 2018-06-21	2,725	963	420	1,994	9,081

Table 2: Multi-horizon Macro-F1 and MCC over 10 runs. Improvement is the differential between our model and the best baseline; significance via paired t-test at 5%.

	NASDAQ										**NYSE**									
	F1 (%)					**MCC (%)**					**F1 (%)**					**MCC (%)**				
Method	+1bd	+3bd	+7bd	+15bd	+30bd	+1bd	+3bd	+7bd	+15bd	+30bd	+1bd	+3bd	+7bd	+15bd	+30bd	+1bd	+3bd	+7bd	+15bd	+30bd
LM	40.9	41.7	40.5	39.5	37.0	2.9	5.0	3.8	2.6	-0.4	44.4	44.4	44.5	44.7	42.4	8.3	7.7	6.5	5.5	3.4
4o-mini	47.3	47.4	46.5	45.8	43.8	11.9	11.7	10.8	9.8	8.8	49.0	49.1	50.0	51.0	49.3	15.4	14.0	14.4	14.8	13.4
3.5-turbo	43.9	44.1	43.0	42.0	40.8	8.0	8.3	7.2	6.0	7.4	48.0	47.7	48.0	48.2	45.6	15.2	14.0	12.8	11.7	9.2
LangGraph	35.6	35.7	35.3	34.9	33.0	3.6	3.4	5.0	6.6	6.4	36.1	36.2	35.8	35.4	33.5	4.1	3.9	5.5	7.1	6.9
GoT	41.3	41.7	42.5	43.2	41.7	4.3	4.2	7.0	10.0	6.0	44.0	42.8	43.5	44.0	44.8	9.9	6.0	6.0	6.0	9.2
WideCtx	48.1	48.2	47.2	46.4	44.5	12.4	12.0	11.1	10.2	9.1	49.6	49.7	50.5	51.6	49.9	15.7	14.4	14.8	15.1	13.7
FinBERT	53.4	53.3	53.2	53.1	53.0	11.8	11.3	10.7	10.0	9.6	53.6	53.4	53.0	52.7	52.6	13.5	12.8	12.7	12.6	12.1
HierFinBERT	53.5	52.1	52.2	52.3	52.0	11.4	10.9	10.3	9.7	9.2	53.7	53.5	52.8	51.9	51.8	13.0	12.5	12.3	12.0	11.5
Ours	**55.7**	**55.1**	**54.2**	**53.5**	**53.2**	**13.7**	**12.6**	**11.5**	**10.6**	**10.1**	**57.4**	**57.0**	**57.2**	**57.5**	**56.3**	**16.2**	**15.2**	**15.4**	**15.6**	**13.9**
Improv. (%)	+4.1	+3.4	+1.9	+0.8	+0.4	+10.5	+5.0	+3.6	+3.9	+5.2	+6.9	+6.5	+7.9	+9.1	+7.0	+3.2	+5.6	+4.1	+3.3	+1.5
(p)-value (%)	3.0	3.5	3.8	4.2	4.2	3.5	3.9	3.9	3.9	3.9	2.3	2.3	2.2	2.2	2.3	3.9	3.5	3.7	3.9	3.9

Panel B – MAEC benchmark – Comparison of our model against models that use audio data

	F1 (%)					**MCC (%)**				
Method	+1 bd	+3 bd	+7 bd	+15 bd	+30 bd	+1 bd	+3 bd	+7 bd	+15 bd	+30 bd
MDRM	48.7	52.0	51.0	53.7	55.9	1.9	2.8	2.7	2.8	6.3
StockGNN	53.0	51.1	52.0	52.6	55.2	4.0	4.3	5.0	6.4	6.0
ECHO-GL	54.4	54.3	54.4	54.9	57.3	6.7	6.3	6.0	6.7	6.9
Ours	**58.7**	**57.9**	**57.0**	**55.2**	**57.6**	**17.6**	**16.0**	**14.0**	**8.8**	**10.0**
Improv.	4.3	3.6	2.6	0.3	0.3	10.9	9.7	8.0	2.1	3.1
(p)-value (%)	2.2	2.3	3.0	4.2	4.2	3.7	3.7	3.7	3.0	2.3

from earnings calls. For this comparison, evaluated on the MAEC dataset, we include: (a) MDRM (Qin and Yang, 2019), a deep regression model that fuses audio and transcript inputs; (b) StockGNN (Medya et al., 2022), which employs a gated graph neural network to model intra-transcript structure via co-occurrence graphs; and (c) ECHO-GL (Liu et al., 2024), a dynamic heterogeneous graph learning framework that links entities such as stocks, sentences, and topics through attention-based aggregation.

Model Configuration and Evaluation Each specialist agent in our framework is powered by GPT-4o-mini. We demonstrate that a coordinated system of these lightweight models achieve superior predictive accuracy compared to larger, monolithic baselines. All model interactions employ deterministic decoding by setting $temperature = 0$ and $top\_p = 1$. During the retrieval step, each of the three agents queries the five most similar facts ($k = 5$) from the knowledge graph. For performance benchmarking, we follow the methodology of prior work (Liu et al., 2024) and use the macro-F1 score and Matthews Correlation Coefficient (MCC) as our primary evaluation criteria, assessed across multiple forward-looking horizons.

Overall Performance Against Baselines Performance comparison with text-only baseline families for the NASDAQ and NYSE datasets are shown in Panel A of table 2.

(O1) Our model substantially outperforms other zero-shot LLMs in macro-F1, achieving over a 17% improvement against a transcript-only GPT-4o-mini on both NASDAQ and NYSE. Crucially, it also surpasses a strong wide-context baseline that was given the same raw transcripts and financial statements. This demonstrates that the performance gains stem not just from more data, but from our framework's structured reasoning: the explicit temporal-causal extraction and multi-agent synthesis via the CTKG provide complementary predictive power beyond what is achievable with a monolithic model and a large context window.

(O2) Our model outperforms text-based encoder models. At the +1-day horizon, our zero-shot agent surpass fine-tuned Hierarchical FinBERT by +2.2 absolute F1 points (+4.1%) and +2.3 MCC (+20.2%) on NASDAQ and +3.7 absolute F1 points (+6.9%) and +3.2 MCC (+24.6%) on NYSE, despite using no task-specific training.

Third, we benchmark our approach against audio-text-alignment models, as shown in Panel B of table 2.

(O3) Financial fundamentals is a far more potent predictor than acoustic cues As shown in Panel-B of table 2, our zero-shot framework now tops the strongest multimodal competitor, ECHO-GL, at the +1-day horizon by +4.3 absolute F1 points (+7.9%) and +10.9 MCC points (+162.7%). This margin is achieved even though ECHO-GL–and every other baseline–is supervised and specifically retrained for each forecast window, whereas our model remains zero-shot across all horizons. The result highlights how grounding language models in fundamental financial evidence yields substantially richer predictive insight than augmenting transcripts with acoustic features alone.

Token Usage, Graph-Query Latency, and LLM Latency The four-agent swarm incurs a *median* inference cost of just $0.0040 per filing. Processing for a complete filing completes in 157 seconds on average. This speed profile supports real-time usage. Full token usage and latency breakdown can be found in Appendix 8.

Tackling Class Imbalance in Zero-Shot LLM Forecasts When we only feed the GPT-4o-mini model with static transcript information, it attains a respectable **54%** balanced-accuracy on NASDAQ and **56%** on NYSE, its **macro-F1** drops to **47%** and **49%**, respectively. A closer look at fig. 3 reveals that the transcript-only model over-predicts positives (red) as management emphasizes positives and downplays negatives. Meanwhile, our model (right) produces a more symmetric distribution. To illustrate the root cause of the class imbalance, we provide an example to contrast the verdicts of a transcript-only baseline with those of our agentic system in the Appendix B.

Variant	NASDAQ	NYSE	MAEC
Full Model (Ours, ZS)	**55.7**	**57.4**	**58.7**
w/o HistoricPerformance	54.2 (↓ 2.7%)	55.6 (↓ 3.1%)	57.1 (↓ 2.7%)
w/o CompetitorAnalyst	55.3 (↓ 0.7%)	56.7 (↓ 1.2%)	58.1 (↓ 1.0%)
w/o HistoricEarnings	54.8 (↓ 1.6%)	56.3 (↓ 1.9%)	57.7 (↓ 1.7%)
w/o KnowledgeGraph	54.3 (↓ 2.5%)	55.3 (↓ 3.7%)	56.8 (↓ 3.2%)
w/o Memory	55.2 (↓ 0.9%)	57.3 (↓ 0.2%)	58.4 (↓ 0.5%)

Table 3: Ablation study: Macro-F1 scores. Declines are in red; improvements (if any) would appear in blue.

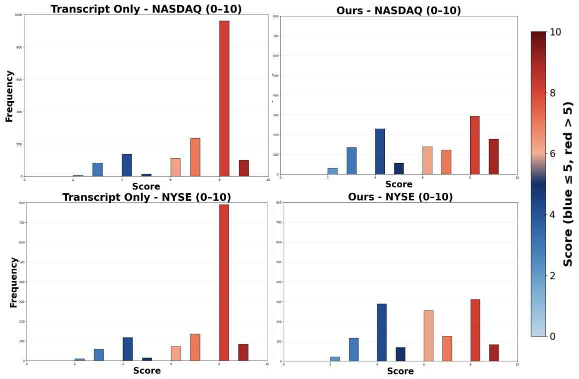


Figure 3: Raw predictions, transcript-only against our model

Ablation Study Table 3 reports the macro-F1 scores obtained after disabling each context module in turn while keeping every other component intact. The steepest decline occurs when the *Knowledge Graph* is removed. Both *Historic_Performance* and *Historic_Earnings*–also provide substantial lift. In contrast, ablating either the short-term *Memory* buffer or the *Competitor_Analyst* module causes only marginal erosion. These results show that graph-aligned context is essential, and temporally grounded agents outperform cross-sectional competitor cues.

To rigorously validate the necessity of the multi-agent architecture, we expand our ablation study to compare the full framework against monolithic and weighted variants. Specifically, we evaluate: (1) **Single-Agent CTKG**, where the synthesizer's prompt and all three specialist prompts are merged into a unified instruction for a single model with access to the graph; and (2) **Weighted Synthesis**, where the final decision is a weighted average of directional scores (0–10) from the three agents rather than a synthesized reasoning step.

Results in table 4 demonstrate that the **CTKG-only single-agent model** consistently underperforms the full multi-agent framework, degrading

Macro-F1 by 2.1% on NASDAQ and 2.4% on MAEC. This suggests that role separation is crucial; forcing a single context window to handle historical baselining, peer comparison, and attribution simultaneously dilutes the reasoning quality. Furthermore, the standalone performance of individual agents (e.g., Only HistoricPerformance) lags significantly behind the collective system, confirming that no single view is sufficient for market prediction. Finally, the Weighted Synthesis strategy yields inferior results compared to our hierarchical textual synthesis, indicating that the synthesizer agent adds value by resolving conflicts logically rather than merely averaging numerical scores.

Table 4: Expanded ablation analysis comparing the Full CARAG model against single-agent monolithic variants, individual agent performance, and heuristic synthesis strategies. Values denote Macro-F1 scores with relative percentage decline in parentheses.

Variant	NASDAQ	NYSE	MAEC
Full Model (Ours)	**55.7**	**57.4**	**58.7**
Architectural Variants			
CTKG-only	54.5 ((↓ 2.1%))	56.2 ((↓ 2.1%))	57.3 ((↓ 2.4%))
Weighted Synthesis	55.3 ((↓ 0.7%))	57.1 ((↓ 0.5%))	57.6 ((↓ 1.9%))
Standalone Agent			
Only HistoricPerf.	54.4 ((↓ 2.3%))	55.1 ((↓ 4.0%))	57.2 ((↓ 2.6%))
Only Competitor	53.2 ((↓ 4.5%))	54.2 ((↓ 5.6%))	56.7 ((↓ 3.4%))
Only HistoricEarn.	54.1 ((↓ 2.9%))	54.9 ((↓ 4.4%))	57.1 ((↓ 2.7%))

Backtest Results We run a backtest assuming execution of signal computed using data on day T at $T + 1$. We find our strategy (blue) to yield **an annualized return of 35.83%** and **Sharpe of 1.64**, beating transcript-only gpt-3.5-turbo (annualized return of 13.64% and Sharpe of 0.64) and LM dictionary (annualized return of -16.69% and Sharpe of -0.47). Over the same period, the S&P 500 posted an annualized return of 1.80% with a Sharpe ratio of 0.19, while Nasdaq recorded an annualized return of -8.94% and a Sharpe ratio of -0.24, beating market benchmarks (S&P 500 and Nasdaq). The optimization algorithm used by the backtest can be found in the Appendix .

5 Conclusion

In summary, we present the CARAG architecture that couples specialist LLMs with a Causal–Temporal Knowledge Graph to forecast post-earnings price shocks. Experiments on NASDAQ, NYSE, and MAEC corpora deliver consistent macro-F1 and MCC improvements across multiple prediction horizons *without* task-specific fine-tuning, achieving a consistent and superior Sharpe in 2-year backtest. Future work will scale the agent swarm with larger open-source pretrained language models and examine CTKG-aware pre-training to further deepen temporal-causal reasoning. One may also test appending a lightweight numerical regressor to the Synthesizer's output, producing ensemble scores that combine symbolic and statistical signals.

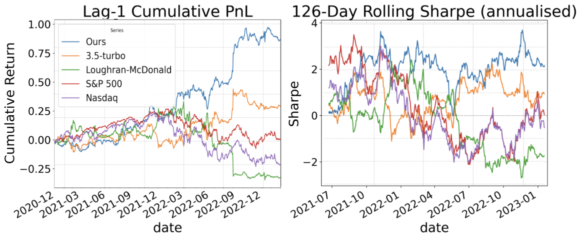


Figure 4: 2-year backtest on our dataset of 1772 NASDAQ transcripts and 1386 NYSE transcripts

6 Limitations

While our proposed framework demonstrates strong empirical performance, several limitations remain.

First, our system was implemented under budget constraints, relying exclusively on GPT-4o-mini as the backbone model. Although this choice highlights the efficiency and cost-effectiveness of our approach, it precludes direct comparison against open-source large language models, which may facilitate broader reproducibility and transparency. Future work should evaluate the framework with open-source backbones to better isolate the benefits of the agentic architecture itself.

Second, our fact extraction pipeline assumes management statements can be grounded against structured fundamentals. In practice, validation of extracted facts against ground truth at the time of the call would require access to proprietary vendor data (e.g., Compustat, IBES) that provide benchmarked metrics. These datasets are not immediately available at call release time, limiting their utility for real-time prediction. While our Causal–Temporal Knowledge Graph partially addresses this gap, a more rigorous fact-checking pipeline remains an open challenge.

References

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

Susan P. Baca, Brian L. Garbe, and Richard A. Weiss. 2000. The rise of sector effects in global equity markets. *Financial Analysts Journal*, 56(5):34–40.

Ray Ball and Philip Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2):159–178.

M. Barry, S. Sanyal, A. K. Maji, and S. K. Adduru. 2025. Graphrag: Leveraging graph-based efficiency to minimize hallucinations in llm-driven rag for finance data. In *Proceedings of the 1st Workshop on Generative AI for Knowledge-Intensive Tasks (GenAIK)*, pages 55–70, Vienna, Austria. Association for Computational Linguistics.

Daniel Bradley, Süleyin Gokkaya, Xi Liu, and Fei Xie. 2017. Are all analysts created equal? industry expertise and monitoring effectiveness of financial analysts. *Journal of Accounting and Economics*, 63(2–3):179–206.

Jingyuan Cao, Hongzhan Chen, Zirui CHEN, Yujie Ding, Guanqun Hou, Zhiyuan Liu, Zhixu Li, Jiaxing Liu, Xiao-Yang Liu, Yang Liu, Maosong Sun, Hengrui Zhang, Yifei Zhang, and Peng Zhou. 2023. A survey on financial large language models. *arXiv preprint arXiv:2311.13788*.

Stefano Cavaglia, Chris Brightman, and Michael Aked. 2000. The increasing importance of industry factors in global equity returns. *Financial Analysts Journal*, 56(5):41–54.

Anne Chang, Xi Dong, Xiumin Martin, and Changyun Zhou. 2023. Ai (chatgpt) democratization and trading inequality. Technical Report 2023-019, Olin Business School Center for Finance & Accounting Research. Working Paper.

Chaochao Chen, Xiaohua Feng, Yuyuan Li, Lingjuan Lyu, Jun Zhou, Xiaolin Zheng, and Jianwei Yin. 2024. Integration of large language models and federated learning. *Patterns*, 5(12).

Yancheng Chen, Zirui Liu, and Feng Lu. 2022. A text-based machine learning approach to forecasting stock returns with conference call transcripts. *The North American Journal of Economics and Finance*, 61:101673.

Andrew Chin and Yuyu Fan. 2023. Leveraging text mining to extract insights from earnings call transcripts. *Journal of Investment Management*, 21(1):81–102.

Ruidi Cui and Jingjing Zhai. 2023. Can chatgpt reduce human financial analysts' optimistic biases? *China Finance Review International*.

Xiaoxi Cui, Weihai Lu, Yu Tong, Yiheng Li, and Zhejun Zhao. 2025. Diffusion-based multi-modal synergy interest network for click-through rate prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 581–591.

J. Gao, W. Xu, T. Li, and Z. Liu. 2024. Memory sharing for multi-agent large language models. *arXiv preprint arXiv:2404.09982*.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, and 1 others. 2024. Openagi: When llm meets domain experts. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2024)*.

S. Ghosh, A. Maji, and SK. Naskar. 2025. Mimic: Multi-modal indian earnings calls dataset to predict stock prices. *arXiv preprint arXiv:2504.09257*.

Boris Groysberg, Paul M. Healy, George Serafeim, and Devin M. Shanthikumar. 2013. The stock selection and performance of buy-side analysts. *Management Science*, 59(5):1062–1075.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 8048–8057.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM multi-agent systems: Challenges and open problems. arXiv:2402.03578. Preprint.

Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. Wiley, New York.

Steven L. Heston and K. Geert Rouwenhorst. 1994. Does industrial structure explain the benefits of international diversification? *Journal of Financial Economics*, 36(1):3–27.

Allen H. Huang, Hui Wang, and Yi Yang. 2024. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

H. Ji, S. Wang, Y. Zhao, and B. Xu. 2025. Llm-based multi-agent systems as graph generative models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1045–1058, Vienna, Austria. Association for Computational Linguistics.

Ohad Kadan, Leonardo Madureira, Rong Wang, and Tzachi Zach. 2012. Analysts' industry expertise. Working paper version accessed.

Harold H. Kelley. 1967. Attribution theory in social psychology. *Nebraska Symposium on Motivation*, 15:192–238.

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

Feng Li. 2010. The information content of forward-looking statements in corporate filings–a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

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 and Knowledge Management (CIKM '20)*, pages 3063–3070, New York, NY, USA. Association for Computing Machinery.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*, pages 4461–4472.

Mengpu Liu, Mengying Zhu, Xiuyuan Wang, Guofang Ma, Jianwei Yin, and Xiaolin Zheng. 2024. Echo–gl: Earnings calls–driven heterogeneous graph learning for stock movement prediction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, pages 13972–13980, Vancouver, Canada. AAAI Press.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Weihai Lu, Yu Tong, and Zhiqiu Ye. 2025. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 559–567.

Weihai Lu and Li Yin. 2025. Dmmd4sr: Diffusion model-based multi-level multimodal denoising for sequential recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6363–6372.

William J. Mayew, Mani Sethuraman, and Mohan Venkatachalam. 2015. Md&a disclosure and the firm's ability to continue as a going concern. *Journal of Accounting Research*, 53(5):1385–1419.

Sourav Medya, Mohammad Rasoolinejad, Yang Yang, and Brian Uzzi. 2022. An exploratory study of stock price movements from earnings calls. In *Companion Proc. WWW 2022*, pages 1–11.

Tobias J. Moskowitz and Mark Grinblatt. 1999. Do industries explain momentum? *The Journal of Finance*, 54(4):1249–1290.

Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, Panfeng Li, Shiyao Zhang, Qifu Yin, Yuanqing Wang, and Yuxi Chan. 2024. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. *arXiv preprint*.

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.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM.

Y. Peng. 2025. Earnings prediction using machine learning: A survey. *Osaka Economic Papers*.

S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011.

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.

Philip J Stone, Dexter C Dunphy, Marshall S Smith, and Daniel M Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, MA.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. arXiv:2306.03314. Preprint.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

Jialin Wang and Zhihua Duan. 2024. Agent ai with langgraph: A modular framework for enhancing machine translation using large language models. `https://arxiv.org/abs/2412.03801`. Code available at `https://github.com/langchain-ai/langgraph`, accessed July 27, 2025.

Z. Wang, TK. Trinh, W. Liu, and C. Zhu. 2025. Temporal evolution of sentiment in earnings calls and its relationship with financial performance. *Applied and Computational Engineering*, 141:195–206.

Xiaolong Wei, Yuehu Dong, Xingliang Wang, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025a. Beyond react: A planner-centric framework for complex tool-augmented llm reasoning. *arXiv preprint arXiv:2511.10037*.

Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025b. Igniting creative writing in small language models: Llm-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17171–17197.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sarthak Soral, Subhabrata Mukherjee, Gricha Loudon, Jianwen Wu, Dan Simonson, Zhang-Xun Wu, Hangyu Pan, Lianhui Qin, Jin Ma, Yichu Zhou, Yangsheng Pang, Yikun Gong, Beibei Shi, Ruirui Li, and 8 others. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zequn Xie, Haoming Ji, Chengxuan Li, and Lingwei Meng. 2025a. Dynamic uncertainty learning with noisy correspondence for text-based person search. *arXiv preprint arXiv:2505.06566*.

Zequn Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025b. Chat-driven text generation and interaction for person retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5259–5270.

Zequn Xie, Boyun Zhang, Yuxiao Lin, and Tao Jin. 2026. Delving deeper: Hierarchical visual perception for robust video-text retrieval. *Preprint*, arXiv:2601.12768.

K. Xu, R. Zhang, Y. Wang, and J. Chen. 2025. A-mem: Agentic memory for large language model agents. *arXiv preprint arXiv:2502.12110*.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022 (TheWebConf 2022)*, pages 2501–2510.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, and et al. 2020. Bigbird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33*.

Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. 2025. Mitigating propensity bias of large language models for recommender systems. *ACM Transactions on Information Systems*, 43(6):1–26.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 14544–14607.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Zhejun Zhao, Yuehu Dong, Alley Liu, Lixue Zheng, Pingsheng Liu, Dongdong Shen, Long Xia, Jiashu Zhao, and Dawei Yin. 2025. Tura: Tool-augmented unified retrieval agent for ai search. *arXiv preprint arXiv:2508.04604*.

Shinn Zhou, Kai Zhang, Andrey Kurenkov, Ling Fan, Kevin Zhou, Yilun Lu, and Stefano Ermon. 2023. Graph-of-thought: Solving elaborate problems with large language models. *arXiv preprint arXiv:2305.10601*.

Jingyuan Zhu, Anbang Chen, Bowen Wang, Sining Huang, Yukun Song, and Yixiao Kang. 2026. Comparing neural architectures for english-spanish machine translation: From lstm to transformer.

Y. Zhu, X. Liu, and ORL. Sheng. 2025. Post-earnings-announcement drift prediction: Leveraging postevent investor responses with multitask learning. *Information Systems Research*.

A Results after model's knowledge cut-off

To provide a more robust evaluation that completely mitigates LLM lookahead bias, we repeat the experiments after `gpt-4o-mini`'s knowledge cut-off (August 2023–December 2024, 2259 transcripts). Results are reported separately for macro-F1 (table 5) and MCC (table 6).

Table 5: Panel A1 – Macro-F1 (%) after GPT-4o-mini knowledge cut-off (August 2023–December 2024, 2259 transcripts).

Method	**+1bd**	**+3bd**	**+7bd**	**+15bd**	**+30bd**
LM	40.5	39.8	39.5	40.5	39.7
3.5-turbo	34.2	31.8	30.8	32.7	31.7
WideCtx	60.4	59.7	58.7	59.1	57.5
FinBERT	58.9	58.5	58.2	58.3	58.1
HierFinBERT	59.7	59.3	58.9	59.0	58.7
Ours	**63.6**	**63.1**	**62.4**	**61.6**	**60.9**
Improv. (%)	+5.3	+5.7	+5.9	+4.2	+3.7

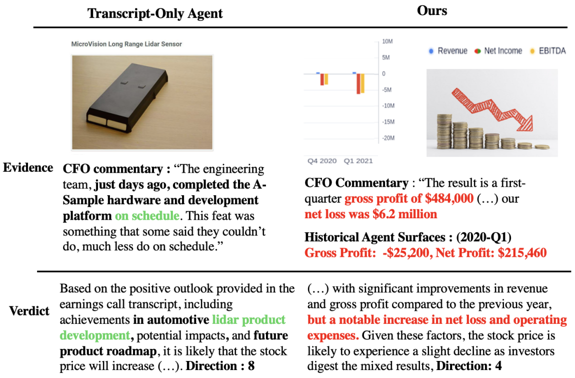
Table 6: Panel A2 – MCC (%) after GPT-4o-mini knowledge cut-off (August 2023–December 2024, 2259 transcripts).

Method	**+1bd**	**+3bd**	**+7bd**	**+15bd**	**+30bd**
LM	9.7	8.5	7.8	8.7	5.4
3.5-turbo	8.5	7.0	6.4	6.4	5.2
WideCtx	27.5	25.9	20.0	20.3	17.5
FinBERT	22.5	22.0	21.5	21.7	20.9
HierFinBERT	23.6	23.0	22.6	22.8	21.9
Ours	**28.6**	**27.7**	**26.4**	**24.0**	**22.6**
Improv. (%)	+4.0	+6.9	+16.8	+5.3	+3.2

The results after the `gpt-4o-mini` knowledge cut-off confirm that our framework's improvements are not an artifact of latent hindsight information. Across both Macro-F1 (table 5) and MCC (table 6), our method consistently outperforms all baselines by a non-trivial margin. Even under the stricter, leakage-free evaluation regime, the CARAG maintains its advantage. The model's performance edge is attributable to its temporal–causal reasoning and structured fact grounding, rather than to knowledge embedded in the pretraining corpus.

B Explainability

To illustrate the root cause of class-imbalance, we contrast the verdicts of the transcript-only baseline with those of our agentic system for the **MicroVision** call (fig. 5), whereby the stock fell $-29\,\%$ the next trading day. The example shows that transcript-only models tend to echo the upbeat language of senior management and therefore overpredict positive outcomes. In contrast, our framework surfaces a widening net loss and increasing operating expenses, yielding a prediction that aligns with the actual negative return (Direction score 4 vs. 8).


MicroVision Long Range Lidar Sensor
10M
5M
0
-5M
-10M
-15M
-20M
Q4 2020
Q1 2021
Revenue
Net Income
EBITDA

	Transcript-Only Agent	Ours
Evidence	**CFO commentary :** "The engineering team, **just days ago, completed the A-Sample hardware and development platform** on schedule. This feat was something that some said they couldn't do, much less do on schedule."	**CFO Commentary** : "The result is a first-quarter **gross profit of $484,000** (…) our **net loss was $6.2 million** **Historical Agent Surfaces : (2020-Q1)** **Gross Profit: -$25,200, Net Profit: $215,460**
Verdict	Based on the positive outlook provided in the earnings call transcript, including achievements **in automotive lidar product development**, potential impacts, and **future product roadmap**, it is likely that the stock price will increase (…). **Direction : 8**	(…) with significant improvements in revenue and gross profit compared to the previous year, **but a notable increase in net loss and operating expenses.** Given these factors, the stock price is likely to experience a slight decline as investors digest the mixed results, **Direction: 4**

Figure 5: Case Study on the MicroVision (2021-Q1) Call

C Hyperparameter Sensitivity

The highest macro–F_1 scores occur when each agent retrieves *five* historical facts and *five* peer facts in fig. 6. Increasing k_{hist} from $5 \to 10$ yields small gains on the dense U.S. datasets, whereas increasing k_{comp} beyond 5 slightly degrade performance on MAEC, as shown in fig. 6.

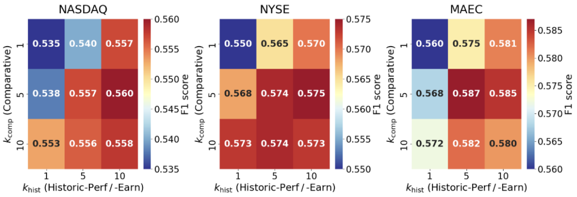


Figure 6: Hyperparameter Sensitivity across all datasets

D Details of the Backtest

We evaluate the predictive signals in a simulated trading strategy that is strictly market neutral. Specifically, we run a backtest assuming execution of signals computed using information available on day T at $T+1$. The optimizer is implemented in MOSEK and solves a second-order cone program (SOCP) that balances expected return against risk under explicit neutrality and exposure constraints.

We estimate a diagonal covariance matrix for individual stock risk over a half-year look-back window, denoted Σ_h. Writing $\Sigma_h = L_h^\top L_h$ for its Cholesky factorization, the daily portfolio $\mathbf{x}_h \in \mathbb{R}^N$ is obtained from:

$$\begin{aligned}
&\max_{\mathbf{x}_h} \ \boldsymbol{\mu}_h^\top \mathbf{x}_h \\
&\text{s.t.} \quad \mathbf{1}^\top \mathbf{x}_h = 0 && \text{(market neutrality)}, \\
&\qquad \|L_h \mathbf{x}_h\|_2 \le \sigma^\star && \text{(ex-ante volatility cap)}, \\
&\qquad \|\mathbf{x}_h\|_1 \le G^\star && \text{(gross-exposure cap)},
\end{aligned}$$

The neutrality constraint $\mathbf{1}^\top \mathbf{x}_h = 0$ ensures the portfolio maintains zero net market exposure on every trading day. This eliminates systematic beta, so portfolio performance reflects only the alpha in the signals rather than coincidental market direction. The gross-exposure cap limits leverage, while the ex-ante volatility cap provides risk control by bounding total portfolio risk ex ante.

We set $\sigma^\star = 10\%$ and $G^\star = 200\%$. This configuration yields a long–short, dollar-neutral, risk-managed strategy. By design, it isolates predictive content in the signals: any excess return or Sharpe ratio achieved cannot be attributed to passive market drift, but must stem from the model's ability to anticipate cross-sectional return differences.

E Facts and Knowledge Graph Schemas

Figure 7: An example of a *Result* Fact for Axon (ticker: AXON) from Q4 2023.

Key	Value
embedding	[-0.0115..., 0.0115...]
metric	"Annual Recurring Revenue (ARR)"
quarter	"2023-Q4"
ticker	"AXON"
type	"Result"
value	"USD 697 million"

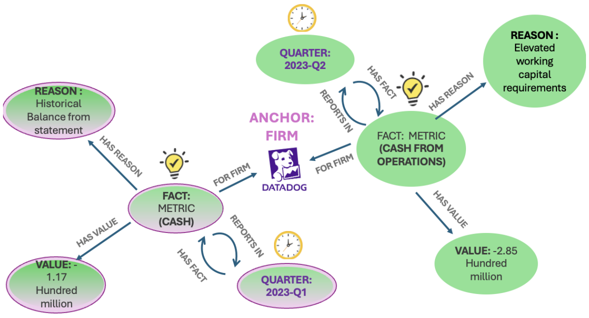

QUARTER: 2023-Q2
REASON : Elevated working capital requirements
REASON : Historical Balance from statement
ANCHOR: FIRM
HAS FACT
REPORTS IN
HAS REASON
FACT: METRIC (CASH FROM OPERATIONS)
FOR FIRM
DATADOG
HAS REASON
FOR FIRM
FACT: METRIC (CASH)
HAS VALUE
REPORTS IN
HAS FACT
HAS VALUE
VALUE: -2.85 Hundred million
VALUE: - 1.17 Hundred million
QUARTER: 2023-Q1

Figure 8: Longitudinal traversal of the agent - agent anchors by the firm and traverses across facts nodes across quarters (2023-Q1, 2023-Q2)

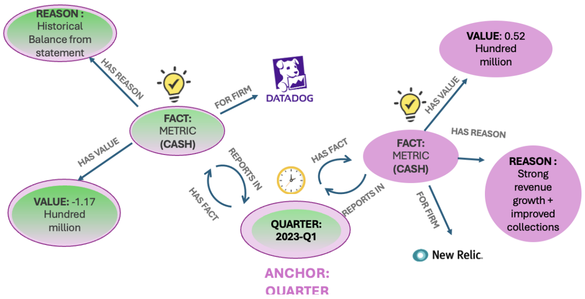


Figure 9: Lateral traversal of the agent - agent anchors by the quarter and traverses across firms (DataDog, New Relic)

Each Fact node is associated with both a Ticker and a Quarter node, ensuring temporal and firm-level alignment. The schema defines six relationship types. The REPORTS_IN edge connects a ticker to a quarter, while HAS_FACT links a quarter to its reported facts. Each fact is connected back to the firm through the FOR_FIRM relationship and to its corresponding value via HAS_VALUE. Explanatory context is preserved through EXPLAINED_BY, which links a fact to its reason node. Finally, QoQ_COMPARES relates facts across consecutive quarters for the same metric, annotated with a percentage change attribute (delta_pct) to capture quarter-over-quarter dynamics.

E.1 CTKG Topology and Construction

To quantify the richness of the structured long-term memory, we analyze the topological properties of the constructed Causal-Temporal Knowledge Graph. The CTKG transforms unstructured transcripts into a dense network of financial logic. As detailed in table 7, the graph encodes hundreds of thousands of nodes across the three datasets. Specifically, the MAEC graph is the most extensive, containing over 400,000 nodes and 440,000 edges.

The graph construction process explicitly models dependencies. *Node-level* features capture discrete entities such as Facts, Reasons, and Metrics, while *Link-level* features (Edges) map the relational structure, such as causal attributions (EXPLAINED_BY) and temporal grounding (REPORTS_IN). The density of HAS_REASON edges (equal to the number of Facts) ensures that every quantitative claim is semantically grounded in qualitative management commentary.

F Latency Benchmarking

Agent	Input (tokens)	Output (tokens)	Graph-Query (sec)	LLM (sec)
CompetitorAnalyst	719	653	38.6	3.8
HistoricPerformance	1,160	715	49.1	11.8
HistoricEarnings	1,012	444	16.3	6.4
Main_Agent	15,580	242	26.6	4.7
Total	18,471	2,054	**130.6**	**26.7**

Table 8: Per-earnings-call tokens and median latencies

G Datasets Contributed

This work contributes several new datasets for earnings call analysis. The full datasets will be open-sourced upon the acceptance of the paper, currently, a selection of them is included in the supplementary material. They include - **Enhanced Earnings Call Datasets**, a curated collection of 1,386 earnings calls from NYSE-listed companies is available, with each record containing the full transcript, associated quarterly financial data, and labels indicating post-earnings price movements. A parallel dataset containing 1,772 earnings calls from NASDAQ-listed companies can also be found.

G.1 Financial Statements

Complementing the earnings-call data is a comprehensive Financial Statements Directory, which contains quarterly financial data for all companies represented in the call datasets, covering over 1,500 unique tickers with as much as 20 years of historical data for each. For every company, three fundamental financial statements are provided, offering a holistic view of financial health. These include the **income statement (_income_statement.csv)**, which details revenues, costs, and profitability; the **balance sheet (_balance_sheet.csv)**, which outlines assets, liabilities, and equity; and the **cash-flow statement (_cash_flow_statement.csv)**,

Table 7: Topological statistics of the constructed Causal-Temporal Knowledge Graph (CTKG). The table details node cardinality (Panel A) and relational edge distributions (Panel B).

Panel A: Node Level Features

Dataset	Transcripts	Tickers	Facts	Reasons	Values	Metrics	Quarters	Total Nodes
NASDAQ	1,772	507	79,775	44,300	79,775	79,775	18	285,922
NYSE	1,386	470	55,738	34,650	55,738	55,738	18	203,738
MAEC	2,725	963	109,417	68,125	109,417	109,417	20	400,084

Panel B: Edge Level Features (Links)

Dataset	FOR_FIRM	HAS_VALUE	EXPLAINED_BY	HAS_FACT	REPORTS_IN	-	Total Links
NASDAQ	79,775	79,775	79,775	79,775	1,772	-	320,872
NYSE	55,738	55,738	55,738	55,738	1,386	-	224,338
MAEC	109,417	109,417	109,417	109,417	2,725	-	440,393

which reports on operating, investing, and financing activities.

H Agent Prompt Templates

The following sections reproduce the exact prompts supplied to each specialist agent. Placeholders in curly braces (e.g., `{self_ticker}`) are populated at runtime.

H.1 Fact Extraction Agent

Fact Extraction Agent

You are a financial analyst tasked with extracting structured facts from a company's earnings call transcript.
Company Ticker: `{ticker}`
Quarter: `{quarter}`
Transcript Text: `{raw_text}`
TASK:

1. Read the text and identify all quantifiable statements, key performance indicators (KPIs), forward-looking guidance, and qualitative assessments of performance.

2. For each identified statement, create a JSON object with the keys: `‘statement‘`, `‘metric‘`, `‘value‘`, and `‘type‘`.

Your output must be a JSON list of these structured fact objects.

H.2 Comparative Peers Agent

Comparative Peers Agent

You are analyzing a company's earnings call transcript alongside statements made by similar firms.
Firm Ticker: `{self_ticker}`
Firm's Facts: `{json.dumps(facts, indent=2)}`
Peer_Facts: `{json.dumps(related_facts, indent=2)}`
Your Task:

- Describe how the firm's reasoning about its own performance differs from other firms, for each fact if possible.

- Cite factual evidence from historical calls.

Keep your analysis concise. Do not discuss areas not mentioned.

H.3 Historical Earnings Agent

Historical Earnings Agent

You are analyzing a company's earnings call transcript against its own past earnings calls.
Current Facts: `{json.dumps(fact, indent=2)}`
Current Quarter: `{current_quarter}`
Historical Facts: `{json.dumps(related_facts, indent=2)}`
Your Task:

1. **Validate Past Guidance**: For every forward-looking statement from past quarters, state if the firm met, beat, or missed that guidance in `{current_quarter}`. Reference concrete numbers (e.g., "Revenue growth was 12% vs. the 10% guided in 2024-Q3").

2. **Compare Results**: Compare the current results being discussed to historical ones, referencing concrete numbers.

3. **Provide Evidence**: Format each evidence line as:
   • *metric*: historical statement → current result.

4. **Highlight Unexpected Outcomes**: Identify where management did not address an important historical comparison or where results diverged sharply from trends. Explain why this matters.

H.4 Financials-Statement Agent

Financials-Statement Agent

You are reviewing a company's earnings call transcript and comparing a key fact to its historical financial statements.
Current Fact from quarter: `{json.dumps(fact, indent=2)}`
Similar Past Facts: `{json.dumps(similar_facts, indent=2)}`
Your Tasks:

- **Direct Comparison**: Compare the current fact to each past fact. Note the quarter, metric, and value. Highlight trends.

- **Supported Outcomes**: Identify where management's comments are confirmed by the financial data.
- **Unexpected Outcomes**: Highlight results that management did not address or that diverge from historical trends. Explain the importance to investors.

Focus on improvements in bottom-line performance (e.g., net income). Note: Figures may be stated in ten-thousands (wan) or hundreds of millions (yi). Account for these scale differences.

H.5 Synthesis & Prediction Agent

Synthesis & Prediction Agent

You are the final arbiter in a multi-agent financial analysis system. Synthesize findings from three specialist agents to predict the stock price reaction.
Analyses Received:

- **Pee Analysis**: `{peer_analysis_summary}`
- **HistoricalAnalysis**: `{historical_analysis_summary}`
- **FinancialsAnalysis**: `{financials_analysis_summary}`

TASK:

1. **Synthesize Findings**: Write a one-paragraph executive summary integrating insights from all three analyses.
2. **Identify Key Drivers**: List the top 3–5 critical factors (positive or negative) that will likely influence investor sentiment.
3. **Predict Price Shock**: Based on the evidence, classify the outcome as: `‘Positive Shock‘`, `‘Negative Shock‘`, or `‘Neutral‘`.
4. **Justify Prediction**: Provide a concise, evidence-based rationale for your prediction, referencing specific agent analyses.

I Key Metrics Extracted from the Financial Statements

To keep the structured input tractable yet economically comprehensive, we distill the ∼15 line-items available in each quarterly balance sheet, income statement, and cash-flow statement down to fifteen core metrics (table 9). These metrics gives the swarm of agents an information-rich view of firm fundamentals that can be compared longitudinally and cross-sectionally within the Causal–Temporal Knowledge Graph. These metrics are embedded as `Fact` nodes in the CTKG and linked to their respective `Quarter` and `Ticker` nodes, allowing agents to reason over temporal trends and peer benchmarks when forming price-move predictions.

Economic theme	**Line-items ingested**
Balance-sheet strength	Cash and cash equivalents; Accounts receivable; Inventory; Property, plant and equipment; Short-term debt; Total current liabilities; Total liabilities; Total shareholders' equity
Profitability & margins	Net profit; Gross profit; Depreciation and amortization
Cash-flow drivers	Net cash flow from operating activities; Net cash flow from investing activities; Net cash flow from financing activities
Per-share performance	Diluted earnings per share (common stock)

Table 9: Unified set of fifteen financial-statement metrics passed to downstream reasoning agents.

J Example Facts Extracted from Earnings Call

Below we provide an example of facts extracted from a **single** earnings call (ALTO, 2023-Q3) from the fact extraction agent. For each fact we report the *type* and the *agents* it is routed to. We provide around 1000 of such calls with associated facts in the supplementary information.

- **Adjusted EBITDA**: $4.7 million. Type: Result. Agents: Historic Performance. Reason: Positive adjusted EBITDA compared to negative $20.6 million in 2022.
- **Gross Profit**: Improved over the same period in 2022. Type: Result. Agents: Historic Performance. Reason: Due to strong crush margins.
- **Production Volumes**: Lower than anticipated. Type: Result. Agents: Historic Performance. Reason: Due to unusually high unscheduled downtime.
- **Specialty Alcohol Sales**: 85 million gallons. Type: Result. Agents: Historic Performance. Reason: Tracking closer to this figure despite lower consumer demand.
- **Corn Basis Levels**: Increased by $0.31 vs. prior period. Type: Result. Agents: Historic Performance. Reason: Illustrates a sharp increase sequentially.

- **Derivative Losses**: $3.7 million. Type: Result. Agents: Historic Performance. Reason: Resulted from adverse impacts on hedging strategy.
- **Annual Adjusted EBITDA Target (mid-2026)**: Over $65 million. Type: Forward-Looking. Agents: Historic Earnings. Reason: Completion of near-term projects.
- **Annual Adjusted EBITDA Target (2027)**: Approximately $125 million. Type: Forward-Looking. Agents: Historic Earnings. Reason: Yeast, CCS, and other long-term initiatives.
- **2024 Contracting Season**: On pace to exceed 2023 delivered gallons. Type: Forward-Looking. Agents: Historic Earnings. Reason: At premiums to fuel-grade ethanol.
- **Specialty Alcohol Sales (2024)**: Targeting 90 million gallons. Type: Forward-Looking. Agents: Historic Earnings. Reason: Based on current contracting progress.
- **Repair and Maintenance Costs**: Significant increases incurred. Type: Risk Disclosure. Agents: Competitor Analyst, Historic Earnings. Reason: Adversely impacted production volumes.
- **Supply Chain Constraints**: Materially impacted installation costs. Type: Risk Disclosure. Agents: Competitor Analyst, Historic Earnings. Reason: Increases of over 70% from original estimates.
- **Capital Market Environment**: Challenging. Type: Risk Disclosure. Agents: Competitor Analyst, Historic Earnings. Reason: Affecting project funding and prioritization.
- **Customer Demand**: Lower than expected. Type: Risk Disclosure. Agents: Competitor Analyst, Historic Earnings. Reason: Impacting sales and profitability.
- **Overall Tone**: Positive. Type: Sentiment. Agents: None (kept for synthesis). Reason: “We remain enthusiastic about our prospects and confident in our long-term growth strategy.”
- **Management’s Confidence**: Optimistic. Type: Sentiment. Agents: None (kept for synthesis). Reason: “We continue to chip away at it and make good progress.”
- **Economic Environment**: Current conditions affecting timelines. Type: Macro. Agents: None (context only). Reason: “Based on the current economic environment.”
- **Market Demand**: Reflective of buyers managing inventory. Type: Macro. Agents: None (context only). Reason: “Customer commitments are not solely based on market demand.”
- **Winter Buying Patterns**: Deep seasonal decline from Q3. Type: Macro. Agents: None (context only). Reason: Seasonal trends affecting sales.
- **Cash Flow from Operations**: $23 million. Type: Result. Agents: Historic Performance. Reason: Reported for the third quarter.
- **Quarterly Capital Spend**: $7 million. Type: Result. Agents: Historic Performance. Reason: Year-to-date plant investment totaled $25 million.
- **Cash Balance**: $26 million. Type: Result. Agents: Historic Performance. Reason: As of September 30, 2023.
- **Total Loan Borrowing Availability**: $118 million. Type: Result. Agents: Historic Performance. Reason: To support business operations and growth.
- **Operating Line of Credit**: $53 million. Type: Result. Agents: Historic Performance. Reason: Part of total loan borrowing availability.
- **Term Loan Facility**: $40 million. Type: Result. Agents: Historic Performance. Reason: Part of total loan borrowing availability.
- **Additional Loan Request Option**: Up to $25 million. Type: Result. Agents: Historic Performance. Reason: Available for operations.
- **Future Uptime and Reliability**: Expected benefits from repairs. Type: Forward-Looking. Agents: Historic Earnings. Reason: Completed during the quarter.
- **Operational Challenges**: Intermittent. Type: Risk Disclosure. Agents: Competitor Analyst,

Historic Earnings. Reason: Subject to various market challenges.

- **Corn Oil Yield Improvements**: Achieved at Pekin dry mill. Type: Result. Agents: Historic Performance. Reason: Through process optimization unique to ICM design.
- **Carbon Capture and Sequestration Project**: Progressing with partner. Type: Forward-Looking. Agents: Historic Earnings. Reason: Designing pipeline and sequestration system.

K Example Analysis from the Synthesis and Prediction Agent

Below we provide an example of input from a **single** earnings call (ALTO, 2023-Q2) ingested into the final synthesis and prediction agent.

Direct Comparison The current facts show an adjusted EBITDA of $16 million, positive operating cash flow, production capacity utilization of 81%, a cash balance of $23 million, and total loan borrowing availability of up to $114 million. These figures reflect operational strength and improved liquidity.

Most Similar Past Facts Past performance provides useful context. Net cash flow from financing activities was negative at -$221.50 thousand, which marks a shift towards better cash flow management. Net profit stood at $22.11 million, higher than the current adjusted EBITDA of $16 million, underscoring profitability as an ongoing challenge. Diluted earnings per share was 0.30, suggesting continued focus on profitability. Operating costs of $353 million highlight the scale of expenses relative to EBITDA. Cash and cash equivalents of $21.20 million were lower than today's $23 million, signaling stronger liquidity. Gross profit was $8.85 million, considerably below the current EBITDA, showing improved efficiency. Net cash flow from operating activities was $9.64 million, while today's positive figure indicates better cash generation. Finally, total current liabilities of $14.94 million contextualize obligations relative to liquidity.

Notable Trends The increase in cash balance from $21.20 million to $23 million points to improved liquidity management. The shift to positive operating cash flow reflects greater efficiency, while adjusted EBITDA growth indicates stronger operations, though still below historical net profit levels.

Supported Outcomes Management's emphasis on achieving positive operating cash flow is validated by these results. Confidence in EBITDA improvements is also justified, though there is still room to enhance profitability.

Validate Past Guidance Forward-looking statements projected over $65 million in incremental annual EBITDA from capital projects by 2025, and approximately $125 million from CCS and other initiatives by 2026. While not directly comparable to current figures, ongoing projects and initiatives suggest progress toward these goals.

Compare Results Discussed Operating cash flow is positive in 2023 compared to $9.64 million in 2022. Net income is also positive, showing an improvement over the $22.11 million recorded in 2022. Capital expenditures reached $8 million, a substantial investment indicating focus on transformation.

Supporting Evidence Operating cash flow improved from $9.64 million to positive in 2023. Net profit also improved from $22.11 million to positive net income. Capital expenditures of $8 million highlight a commitment to long-term transformation.

Analysis vs. Comparable Firms ALTO emphasizes optimism and leadership transition, unlike peers that rely primarily on financial reporting. Regulatory alignment under the Low Carbon Fuel Standard is framed as a long-term advantage, while peers stress immediate financial impacts. Inflationary pressures and rising corn prices are openly acknowledged by ALTO, unlike peers. Market discipline is highlighted as margin support, whereas peers omit it. ALTO communicates with a positive tone, in contrast with peers' results-driven style. CapEx is framed as transformational, with emphasis on year-to-date investments, while peers provide little discussion. ALTO also highlights share repurchases as part of capital strategy, a theme absent in peer reporting.

Conclusion ALTO highlights management sentiment, strategic investments, and market conditions, distinguishing itself from peers who focus narrowly on financial results. The company delivered adjusted EBITDA of $16 million, capacity utilization of 81%, and positive operating cash flow in 2023. Management's optimism, combined with capital projects and favorable regulatory shifts, suggests a positive trajectory for earnings growth.