Fast Numbers, Slow Language: Bridging Quantitative and Qualitative Earnings Signals

Anonymous ACL submission

Abstract

Earnings announcements release two types of information sequentially: **quantitative surprise** (numeric earnings-per-share (EPS)/revenue versus analyst estimate) arrives first in press releases and financial news, processed by algorithmic traders within minutes; **qualitative language** (management tone, guidance, question-and-answer (Q&A) credibility) arrives 30–90 min later in the earnings conference call transcript (ECT), requiring human interpretation overnight. Financial economists have studied quantitative surprise for 50 years; natural language processing (NLP) researchers have studied qualitative ECT signals for a decade. Despite studying the same event, the two communities used incompatible frameworks: different targets (return vs. volatility), trading setups (long top decile and short bottom-decile vs. trade-all), and metrics (return spread between top and bottom 20% (Q5−Q1) vs. mean squared error MSE); making direct comparison and connection challenging.

We bridge these communities with EARNINGSINONE, the first corpus aligning earnings news, ECTs, and intraday and next-day prices across SP 1500 (broad U.S. equity universe, 2022–2025). Applying unified trading and evaluation tools to both signal types, we confirm a clean speed separation: *fast numbers, slow language*: quantitative surprise peaks at announcement and is largely eliminated by next market open; qualitative ECT sentiment peaks on the next trading day, real and tradeable, but hidden under prior transcript-based evaluation that optimised sign-agnostic volatility with pointwise MSE. *Code and dataset will be released upon acceptance.*

1 Introduction

Earnings announcements release two types of information sequentially. **Quantitative**: the numeric surprise (e.g. "Apple beat earnings per share (EPS) by $0.08, reporting $1.52 vs. $1.44 expected")arrives first, actionable within minutes. **Qualitative**: management tone and guidance (e.g. "we are cautiously optimistic heading into third quarter (Q3), though macro headwinds remain a concern")arrives 30–90 min later in the earnings conference call, requiring human interpretation overnight. Despite studying the same event, two research communities, financial economists and natural language processing (NLP) researchers, have examined these signals in isolation using incompatible frameworks.

Two gaps we address. Financial economists have studied quantitative surprise for over 50 years (Ball and Brown, 1968; Bernard and Thomas, 1989; Fink, 2021), but focused primarily on next market open and multi-day horizons. We use newsflow timestamps to anchor the *first tradable realization* of numeric information: without precise event-level timestamps, numeric signal exhausted within minutes is indistinguishable from language signal maturing overnight (Christensen et al., 2025). NLP researchers have studied qualitative earnings conference call transcript (ECT) signals (Qin and Yang, 2019; Yang et al., 2020, 2022) but without a practical trading framework (no buy/sell selection based on signal confidence, all stocks equally weighted and traded) and metrics not designed for directional trading (mean squared error (MSE) over sign-agnostic volatility across all stocks), leaving real trading value unexplored: a signal is only as useful as the positions it generates, while the financial economics community has long provided a mature trading framework: long top-decile, short bottom-decile (buy strongest signals, sell weakest), and quantitative asset management offers the natural evaluation tools for signal ranking and strategy evaluation: information coefficient (IC, Spearman rank correlation between signal and returns) and Sharpe ratio (a widely used measure of risk-

adjusted trading performance: return earned per unit of risk) (Grinold and Kahn, 2000).

Our bridge. Neither community could see how the other's signal performs, as incompatible frameworks made direct cross-community comparison challenging, leaving each community with only part of the picture. We connect these approaches by asking: *what happens when evaluating both quantitative and qualitative earnings information together, at announcement resolution and next market open, with intraday holding periods and unified financial evaluation tools?* To answer this, we introduce EARNINGSINONE, the first corpus aligning earnings news, ECTs, and intraday and next-day prices for the same events across SP 1500. Our answer reveals a clean separation by processing speed: quantitative surprise is *fast*: IC peaks at announcement and decays to near zero by the next market open; qualitative sentiment is *slow*: predictive content peaks on the next trading day. The bridge yields concrete benefits for both communities: for the financial economics community, we evaluate quantitative earnings surprises at announcement-level timestamps, revealing short-horizon alpha that multi-day studies miss (Martineau, 2022); for the NLP community, applying a long top-decile/short bottom-decile trading framework and quantitative asset management metrics (IC, Sharpe) reveals that ECT sentiment is a tradeable signal with real economic value.

Contributions.

(1) **Cross-community bridge:** We propose a unified evaluation framework connecting financial economics and NLP earnings research, sharing a trading convention (long top-decile, short bottom-decile), metrics (IC, Sharpe), and data (earnings events), enabling direct cross-community comparison.

(2) **Dataset:** EARNINGSINONE is the first corpus resolving all three cross-community pain points simultaneously: precise timestamps from news, intraday and next-day prices, and ECT+earnings news alignment across SP 1500.

(3) **Empirical findings:** Applying the unified framework reveals two results. *Fast numbers, slow language*: quantitative surprise peaks at announcement and decays by next market open; qualitative ECT sentiment peaks on the next trading day. *NLP signals have real economic value*: evaluated under a practical trading framework, ECT sentiment produces actionable trading returns previously hidden by misaligned metrics (MSE over sign-agnostic volatility targets).

2 Related Work

Quantitative Earnings Prediction: The PEAD Literature. The PEAD phenomenon—first identified by Ball and Brown (1968) and formalized by Bernard and Thomas (1989)—documents that stock returns drift in the direction of earnings surprise for days to weeks after announcement, measurable via standardized unexpected earnings (SUE) and evaluated with quintile spread (Q5−Q1) and cumulative abnormal return (CAR) (Latane and Jones, 1979). Revenue surprises carry incremental predictive content beyond EPS (Jegadeesh and Livnat, 2006), and management guidance moves prices independently at announcement time (Billings et al., 2015). Critically, Martineau (2022) show that multi-day PEAD has been non-existent for non-microcap stocks since 2006, prices now fully reflect earnings surprises on the announcement date itself, as near-instantaneous price jumps replace gradual drift (Christensen et al., 2025). This compression of the exploitable window motivates our evaluation at short holding periods: quantitative surprise alpha has not disappeared, it has accelerated, and this work recovers it by evaluating with immediate actions rather than waiting for the next market open.

Qualitative Earnings Prediction: NLP and ECT-Based Models. The dominant NLP paradigm applies multimodal learning over ECT audio and text (Qin and Yang, 2019; Yang et al., 2020; Cao et al., 2024, 2025), with extensions for numerical reasoning (Yang et al., 2022; Qin et al., 2025), graph-based inter-stock modeling (Sawhney et al., 2020; Liu et al., 2024), and agentic workflows (Li and Lu, 2026). Recent work questions the informativeness of ECT embeddings as a standalone signal: same-company transcripts are highly formulaic across periods, limiting incremental information beyond boilerplate (Shih et al., 2025; Yu et al., 2025). Meursault et al. (2023) show that text-based ECT patterns produce drift larger than classic PEAD, providing independent support for our fast/slow distinction — yet all these works evaluate at the multi-day horizon where drift has largely dis-

appeared (see above) under misaligned pointwise metrics; §3.2 elaborates. Instead, this work applies LLM sentiment scoring over ECTs and news under a long/short trading framework.

3 Framework: Fast Numbers, Slow Language

3.1 The Earnings Event Timeline

We model an earnings event as four sequential stages (Figure 1):

(1) **Analyst consensus estimate** ($t < t_0$): analyst expectations accumulate over weeks via reports and news (e.g. analyst estimate = \$0.50).

(2) **Earnings press release** ($t = t_0$): reported actuals disclosed immediately, triggering price discovery (e.g. EPS actual = \$0.60).[1]

(3) **Financial news dissemination** ($t = t_0 + \varepsilon$, $\varepsilon \approx$ 1–5 min): journalists publish articles co-reporting actuals and estimates (*e.g. EPS actual \$1.05 vs. analyst estimate \$1.02*).

(4) **Earnings conference call (ECT)** ($t = t_0 + \Delta$, $\Delta \approx$ 30–90 min): management call with analysts, **primary text source of prior NLP work**, typically available after stage (3).

Stages (1)–(3) carry *quantitative* information (numeric actuals, consensus estimates, and surprise magnitude) and are all reflected in financial news articles; while stage (4) carries primarily *qualitative* information recorded as earnings conference call transcripts (ECTs): management tone, credibility signals, and forward guidance language. The two are structurally linked: the ECT is a *linguistic encoding of the numbers*, available only after those numbers have been public for 30–90 minutes, with management exercising *discretionary framing* over how results are presented: a company can beat EPS while signalling caution, or miss while projecting confidence, language and numbers can tell entirely opposite stories.

This creates two consequences: *informational complementarity* (numbers encode **what** was reported; language encodes **how** management chose to present it) and a speed difference rooted in the timeline itself: the 8-K precedes the ECT by 30–90 min, making numeric surprise actionable before qualitative language exists; and numbers are unambiguous and machine-readable; algorithms act simultaneously, driving rapid price discovery, while tone requires interpretation and credibility assessment, so market consensus on qualitative signals forms usually overnight.

3.2 One Phenomenon, Two Communities

Earnings announcements have been studied through two lenses: the *financial economics community* focuses on numeric surprise and directional return; the *NLP community* focuses on ECT-based verbal cues (management tone, guidance language, and Q&A credibility) as incremental predictors beyond the headline numbers. Table 1 summarises how these beliefs translate into five methodological divergences (see Appendix A for full details).

- **Target:** financial economics predicts directional return; NLP predominantly predicts sign-agnostic volatility; we target directional return, the most actionable measure of market reaction to earnings.
- **Data & Signal:** financial economics uses vendor numeric surprise (ΔEPS%, SUE); NLP uses ECT language embeddings; this work combines both via news-based information extraction (IE) and LLM sentiment (§5).
- **Entry & Horizon:** both communities enter at next market open and hold for 3–30 days, missing the intraday window where fast surprise alpha concentrates (Martineau, 2022); we enter at *FirstActual* (5–60 min) and *NextOpen* (5–390 min), each at its natural speed.

The fourth and fifth divergences, **trading framework and evaluation metric**, are causally linked: the absence of a trading framework in NLP work leads directly to a misaligned metric. The financial economics community goes long the top decile, short the bottom decile (Bernard and Thomas, 1989), treating earnings as a cross-sectional ranking problem naturally evaluated by IC and Sharpe. The NLP community evaluates pointwise with MSE over all announcing stocks, implicitly treating every event as equally tradeable. *Consider a session with 10 announcing stocks: a financial economist goes long the top-1 beat and short the worst-1 miss, measuring whether the signal correctly separates them (IC); an NLP model predicts all 10*

[1] Announcements occur either AMC (after-market close, 16:00–20:00 ET) or BMO (before-market open, 06:00–09:30 ET), determining the entry window; AMC trades at next session open unless entered via extended hours, BMO at the regular open. We enter at *FirstActual* in both cases.

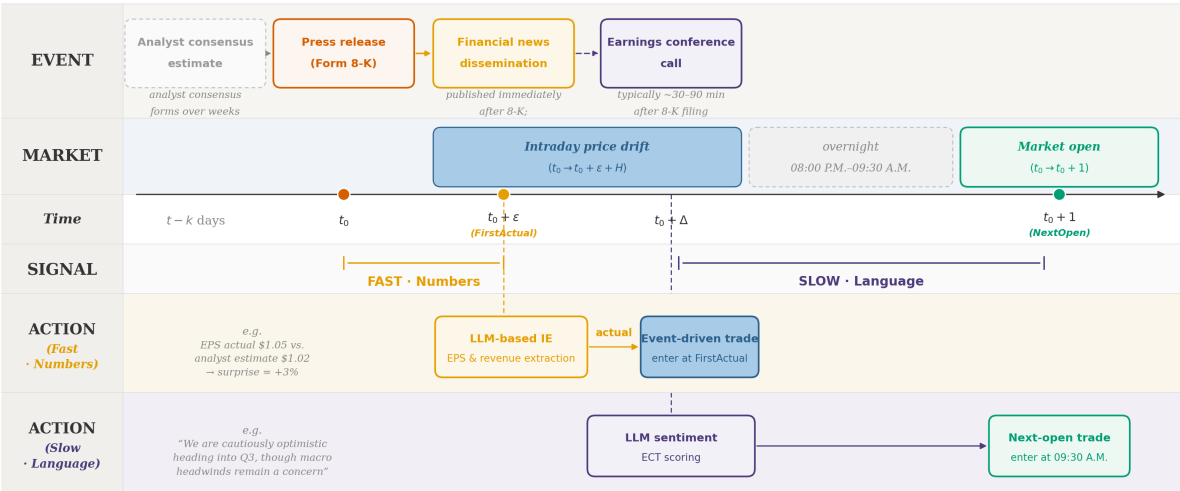

Figure 1: The earnings event timeline and two trading approaches. **EVENT:** Four sequential stages around announcement time t_0: analyst estimate (weeks prior); press release (Form 8-K, t_0) discloses **what** the numbers are; financial news ($t_0+\varepsilon$, $\varepsilon \approx 0$–5 min) contextualises actuals against estimates; earnings conference call ($t_0+\Delta$, $\Delta \approx 30$–90 min) conveys **how** results are framed and provides forward guidance. **FAST (Numbers):** LLM-based information extraction yields EPS/revenue surprise from news; strategies enter at *FirstActual* and hold for up to 60 min (*e.g. EPS actual $1.05 vs. analyst estimate $1.02 → surprise = +3%*). **SLOW (Language):** LLM sentiment scored over the earnings conference call transcript; strategies enter at *NextOpen* and hold for up to 390 min (next-day close) (*e.g. "We are cautiously optimistic heading into Q3, though macro headwinds remain a concern."*).

Dimension	Financial Economics Community	NLP Community	This Work (Bridge)
Target	Post-ann. return[1-5]	Post-ann. volatility[6-11] / Post-ann. return direction [12]	Post-ann. return
Data & Signal	EPS/revenue surprise ΔEPS% + SUE[1-5]	ECT (text / audio); language embeddings[6-12]	Earnings news + ECT; ΔEPS% + Sentiment socre
Entry & Horizon	Next market open; 3–30 days[1-5]	τ-day post call; 3–30 days[6-12]	At announcement (5–60 min); next market open (5–390 min)
Trading Mechanism	Long top-decile short bottom-decile[1-5] (cross-sectional sort)	None (implicitly trade all)[6-12]	Long top-decile short bottom-decile (historical percentile)
Metric	Q5−Q1 / CAR[1-5]	MSE[6-11] / F1[12]	Q5−Q1, IC, SR, Tot. Ret.

Table 1: Methodological comparison across the financial economics and NLP communities and this work. **Financial economics:** [1] Ball and Brown (1968); [2] Bernard and Thomas (1989); [3] Latane and Jones (1979); [4] Martineau (2022); [5] Ben-Rephael et al. (2021). **NLP:** [6] Qin and Yang (2019); [7] Yang et al. (2020); [8] Yang et al. (2022); [9] Sawhney et al. (2020); [10] Cao et al. (2024); [11] Yu et al. (2025); [12] Liu et al. (2024). Post-ann. = post-announcement (i.e. after the earnings release); Post-announcement volatility = log rolling standard deviation of daily returns (Kogan et al., 2009) ($\tau \in \{3, 7, 15, 30\}$ days); SUE = $(\mathrm{EPS}_{\mathrm{act}} - \mathrm{EPS}_{\mathrm{est}})/\sigma(\varepsilon)$ (standardised unexpected earnings), where $\sigma(\varepsilon)$ is the standard deviation of historical analyst forecast errors; τ-day = $\tau \in \{3, 7, 15, 30\}$ days post earnings conference call; MSE = mean squared error; F1 = harmonic mean of precision and recall; IC = information coefficient (Spearman rank); Q5−Q1 = return spread between top and bottom quintile; SR = Sharpe ratio; Tot. Ret. = total return (equivalent to CAR). long top-decile / short bottom-decile = buy stocks with strongest signals, sell stocks with weakest signals (top and bottom 10%); cross-sectional sort = stocks ranked against peers within the same session; historical percentile = stocks ranked against the market-wide historical score distribution.

and optimises MSE, indifferent to whether the high-conviction predictions are correct. We adopt IC and Sharpe as primary metrics and retain Q5−Q1 for comparability with the PEAD literature.

4 Dataset: EARNINGSINONE

4.1 Novelty and Positioning

EARNINGSINONE is the first dataset enabling unified evaluation of fast quantitative and slow qualitative earnings signals. No prior dataset simultaneously covers both dimensions:

- **Intraday and interday price resolution.** Both communities evaluate at next market open; EARNINGSINONE pairs each event with minute-level price bars, enabling intraday evaluation of fast signals and next market open evaluation of slow signals.
- **Earnings news: precise timestamps, real-time actuals, and complementary narrative.** Prior NLP work relies solely on ECTs (Mukherjee et al., 2022; Yu et al., 2025); financial news articles additionally provide precise publication timestamps enabling intraday entry, real-time EPS actuals without vendor lag (Hoechle et al., 2015), and a complementary qualitative narrative that pairs with ECTs.

4.2 Coverage, Split, and Availability

EARNINGSINONE covers the SP 1500 universe, comprising SP 500 (large-cap), SP 400 (mid-cap), and SP 600 (small-cap)[2], spanning 2022–2025. The corpus comprises 5,428 earnings events across 701 unique tickers, covered by 257,815 news articles. News articles, intraday prices, and analyst consensus estimates are all sourced from data vendor EODHD Historical Data[3]. We plan to release verified article URLs, earnings event metadata, and the alignment schema upon publication; full dataset details are in Appendix F.

5 Unified Evaluation Framework

5.1 Signal Extraction

We apply GPT-5-mini[4] to earnings-related news article to extract EPS actuals and revenue actuals; estimates are sourced from EODHD Historical Data pre-release.[5] We denote **(EN)** signals as LLM-extracted actuals with EODHD pre-release estimates; **(EOD)** as vendor-confirmed EODHD actuals and estimates. For qualitative signals, we apply GPT-5-mini to each ECT or news article to produce a scalar sentiment score in $[-1, 1]$ reflecting overall management tone and earnings narrative. Full prompt in Appendix G.

5.2 Trading Mechanisms

Both strategy types go long top 10% and short bottom 10% signals within each session, following canonical trading settings. They differ in how rankings are formed:

- **Canonical PEAD** (Bernard and Thomas, 1989): contemporaneous cross-sectional sort within each session; applicable at *NextOpen* when all announcements are known (e.g. 10 stocks report the same day: the top beat is bought, the worst miss sold).
- **Historical percentile:** ranks each signal against an expanding-window cumulative distribution function (CDF) of the past 12 months. Required for *FirstActual* entry, as stocks announce throughout the session, making contemporaneous sorting infeasible (e.g. a stock scoring in the top 10% of the whole market's past 12-month signal distribution is bought).

Full trading setting in Appendix C.

5.3 Entry, Horizon, and Return Measurement

We define two entry anchors:

- **FirstActual**: earliest article with confirmed EPS/revenue actual; estimate from EODHD; intraday horizons $H \in \{5, 15, 30, 60\}$ min.[6]
- **NextOpen**: 09:30 am open on the following day, standard entry anchor in both communities; intraday horizons $H \in \{5, 15, 30, 60, 120, 180, 390\}$ min.[7]

For each anchor τ and horizon H, log return:

$$r_{i,\tau,H} = \log P^{\text{close}}_{i,\tau^+ + H} - \log P^{\text{open}}_{i,\tau^+}$$

where i = stock, P = price, τ^+ is the first H-min bar starting at least 1 min after τ (1-min information processing lag); e.g. τ=16:00:32 $\rightarrow$ τ^+=16:02 $\rightarrow$ τ^++H=16:07 (if H=5 min).

[2] SP 600 represents only ∼100 events after data cleaning, reflecting lower financial news coverage for smaller firms; results are thus primarily driven by large- and mid-cap stocks.

[3] Data vendor: EODHD Historical Data

[4] GPT-5-mini Model Card. Knowledge cutoff: May 31, 2024, slightly overlapping our test period (2024–2025).

[5] We use vendor estimates (EODHD Historical Data) since news estimates do not always co-appear with actuals and may add delay; EODHD pre-release estimates are available 1–3 days before the press release at negligible cost.

[6] Typically ≈07:00 for BMO announcements and ≈16:00–16:30 for AMC announcements; *FirstActual* horizons are capped at 60 min since AMC occur after market close, leaving no extended intraday window. full details in Appendix B.

[7] 390 min = end of regular trading session.

6 Experiments

6.1 Experimental Setup

Setup. We evaluate on the SP 1500 universe using a strict temporal split: train 2022, validation 2023, test 2024–2025. For each strategy the best horizon is selected on the validation set and held fixed for the test. Full details in Appendix C.

Evaluation Metrics. We report five metrics; full definitions in Appendix D:

- **IC** (t_{NW}): Spearman rank IC (Grinold and Kahn, 2000) with Newey-West t-statistic; measures pure signal quality; position-agnostic.
- **Q5−Q1 bp** (t): Fama-MacBeth session-averaged quintile spread (Bernard and Thomas, 1989) in basis points; adds selectivity to IC, model-free and matching PEAD reporting.
- **Sharpe**: annualised risk-adjusted return; measures overall strategy quality by incorporating signal quality, selectivity, and sizing; the most complete yet implementation-sensitive.
- **Tot. Ret.**: cumulative total return.
- N: number of trades.

Passive, Naïve, and Prior-NLP Baselines. Market benchmarks with no earnings signal: **SPY** and **QQQ**[8] (S&P 500 and Nasdaq-100 index funds); *Random L/S* and *Long-only* (no-signal floors). We additionally include two prior-NLP baselines: *Ticker-Mean* (mean of historical same-ticker post-announcement returns) and *Embedding*[9] (ECT embeddings passed through a 2-layer MLP). Yu et al. (2025) show that ECT embeddings exhibit stronger within-ticker than cross-ticker similarity,[10] such that embedding predictions largely reflect ticker identity rather than earnings content.

Quantitative Conditions. Five variants isolate four factors:

- **Ranking Mechanism** (PEAD vs. historical percentile): *PEAD-EPS(EN)* uses contemporaneous sort (Bernard and Thomas, 1989); *EPS(EN)* uses historical percentile; both at NextOpen, isolating the mechanism effect.
- **Entry anchor** (NextOpen vs. FirstActual): *EPS(EN)* at both anchors isolates entry timing.
- **Source breadth** (EPS vs. Joint): *Joint(EN)* adds revenue surprise (Jegadeesh and Livnat, 2006) to EPS at FirstActual, isolating the revenue contribution.
- **Data quality** (EOD vs. EN):[11] vendor-confirmed EODHD actuals vs. LLM-extracted, providing a data-quality ceiling.

Qualitative Conditions. Text-based signals evaluated at next market open:

- *Sentiment(ECT)* at *NextOpen*: LLM sentiment over full ECT transcript; primary prior-work comparison point. Two variants explicitly examine the role of numeric information: *Sentiment(ECT-mask)* replaces all numbers with `[NUM]` (isolates pure verbal signal) and *Sentiment(ECT-informed)* prepends the EPS/revenue surprise header (grounding in known beat/miss); both in the main table.
- *Sentiment(EN-agg)* at *NextOpen*: sentiment aggregated over up to 10 post-earnings news articles up to market open; direct comparison with ECT on the same timing window.

6.2 Results

Figure 2 characterises signal quality across horizons before we report trading results (Table 2); the full table with all variants is in Appendix H.

Baselines. Naïve baselines yield no reliable alpha: *Random L/S* (Sharpe = −0.88), *Long-only* (Sharpe = 0.11), *Ticker-Mean* (Sharpe = −3.13). *Embedding* sign variant (Sharpe = 0.80) outperforms these but underperforms all signal-based strategies; the magnitude variant is weaker (full results in Appendix H).

Quantitative Signals. Mechanism. *PEAD-EPS(EN)*[12] achieves Sharpe = 3.31 despite IC = 0.03 (t = 0.97); high Sharpe reflects extreme selectivity (top and bottom 10% long/short), not broad ranking quality.

Entry anchor. *EPS(EN)* at *FirstActual* (IC = 0.03, Sharpe = 1.81, Q5−Q1 = 137.7 bp) substantially exceeds the *NextOpen* variant (IC = −0.02, Sharpe = −0.31), confirming intraday alpha is fully arbitraged away by open.

[8] SPY and QQQ are the most widely traded passive ETFs tracking the S&P 500 and Nasdaq-100; SPY, QQQ.

[9] text-embedding-3-large (OpenAI).

[10] We confirm this in EARNINGSINONE: 0.80 vs. 0.67 avg.

[11] EODHD actuals have reporting lag and are not live-deployable; *Joint(EOD)* is a data-quality ceiling.

[12] *PEAD-Joint(EN)* does not improve over *PEAD-EPS(EN)*; full results in Appendix H.

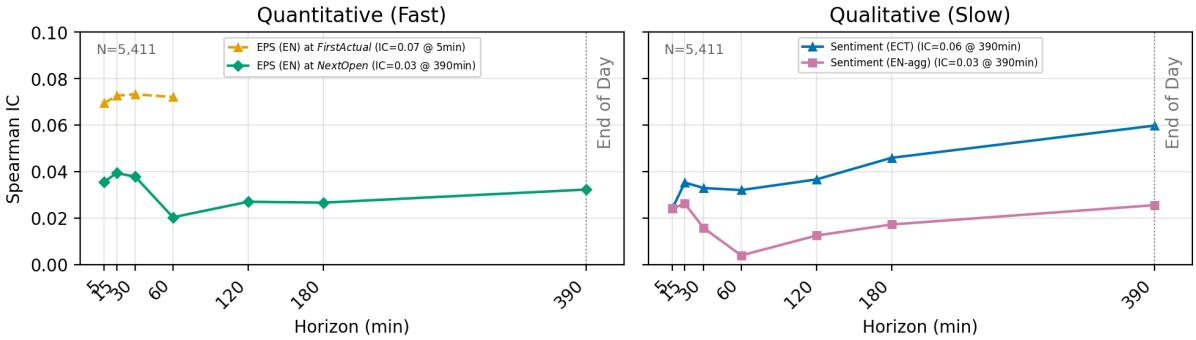

Figure 2: Pooled Spearman IC by horizon, SP 1500 (2022–2025; stable across splits). **Left (Quantitative):** *EPS(EN)* at *FirstActual* (dashed) holds IC ≈ 0.07 stably across 5–60 min; at *NextOpen* (solid) decays from 0.04 to 0.03, short-horizon alpha arbitraged away by open. **Right (Qualitative):** *Sentiment(ECT)* rises to IC = 0.06 at next-day close; *Sentiment(EN-agg)* dips near zero at 60 min before recovering. Horizons beyond 390 min excluded (IC < 0).

Category	Strategy	Entry anchor	IC (t_{NW})	Q5−Q1 bp (t)	Sharpe	Tot. Ret.	N
Naïve	Random L/S	NextOpen	—	—	−0.88	−10.5%	2426
	Long-only	NextOpen	—	—	0.11	5.8%	2077
	Ticker-Mean	NextOpen	—	—	−3.13	−2.2%	2360
	Embedding	NextOpen	—	—	0.80	23.3%	2468
Quantitative (Fast)	*PEAD-EPS(EN)*	NextOpen	0.03 (0.97)	33.9 (1.36)	**3.31**	29.1%	380
	EPS(EN)	NextOpen	−0.02 (−0.56)	−5.4 (−0.12)	−0.31	−4.2%	400
	EPS(EN)	FirstActual	0.03 (0.94)	**137.7 (3.54)**	1.81	50.4%	400
	Joint(EN)	FirstActual	0.05 (1.58)	61.3 (1.94)	1.79	39.9%	296
	Joint(EOD)	FirstActual	**0.08 (2.40)**	118.2 (4.13)	2.43	**64.5%**	343
Qualitative (Slow)	*Sentiment(ECT)*	NextOpen	**0.11 (4.04)**	**153.0 (4.30)**	2.28	63.6%	591
	Sentiment(ECT-mask)	NextOpen	0.07 (2.62)	118.7 (3.96)	0.40	7.0%	412
	Sentiment(ECT-informed)	NextOpen	0.06 (2.10)	94.1 (2.62)	**2.65**	**72.0%**	599
	Sentiment(EN-agg)	NextOpen	0.05 (1.89)	−4.6 (−0.24)	0.42	5.9%	496

Table 2: Out-of-sample trading performance, test set (2024–2025 with 2,468 events), SP 1500. **Entry anchors:** *FirstActual* = first news article with confirmed EPS/revenue actuals; *NextOpen* = 09:30 am on the following trading day. **Metrics:** IC (t_{NW}) = Spearman rank IC (Grinold and Kahn, 2000) computed per event date (t_{NW}: Newey-West t-statistic correcting for serial correlation); Q5−Q1 bp (t) = Fama-MacBeth date-averaged quintile spread (Bernard and Thomas, 1989) (t: Fama-MacBeth standard errors); $|t| > 2$ corresponds to $p < 0.05$ for both; N = number of trades. **Signals:** *Joint* = EPS + revenue surprise combined; EOD: EODHD pre-release estimates; professional-grade vendor actuals with superior data quality, but not available in real time for live deployment; **Metric Interpretation:** IC measures overall signal quality (does the signal rank stocks correctly?); Q5−Q1 measures tail separability (do the extreme quintiles trade well?); Sharpe measures full strategy performance (does it make money risk-adjusted?). High IC with low Sharpe indicates the signal ranks stocks well but individual trade returns are noisy and volatile; high Sharpe with low IC signals selectivity-driven returns (strong performance concentrated in a small number of high-conviction trades). **Bold** = best per column within group. **Passive Market benchmarks:** SPY and QQQ are the most widely traded ETFs tracking the S&P 500 and Nasdaq-100, serving as buy-and-hold baselines: SPY total return 48% (Sharpe 1.02); QQQ total return 54% (Sharpe 0.94).

Source breadth. *Joint(EN)* raises IC to 0.05 over *EPS(EN)* but reduces Q5−Q1 and Sharpe, suggesting revenue surprise improves cross-sectional ranking but reduces tail concentration.

Data quality. *Joint(EOD)* achieves IC = 0.08 ($t = 2.40$), Sharpe = 2.43, outperforming *Joint(EN)*; data quality matters at the margin.[13]

Qualitative Signals. **ECT sentiment (vanilla).** *Sentiment(ECT)* achieves IC = 0.11 ($t = 4.04$), Sharpe = 2.28, Q5−Q1 = 153.0 bp ($t = 4.30$), all three metrics stronger and more significant than any quantitative variant at *NextOpen*, confirming ECT sentiment is a tradeable signal at *NextOpen*.

ECT ablations. *Sentiment(ECT-mask)* achieves IC = 0.07 but Sharpe = 0.40; the masking disrupts score distributions for position sizing, producing noisier returns. *Sentiment(ECT-informed)* reverses this: Sharpe = 2.65 (highest overall) but IC = 0.06; numeric grounding boosts tail concentration without improving average ranking.[14]

[13] Full EN vs. EODHD actuals validation in Appendix E.

[14] All three ECT variants show similar pooled IC and high

News sentiment. *Sentiment(EN-agg)* (IC = 0.05, Sharpe = 0.42) underperforms ECT sentiment across all metrics, suggesting news articles carry less qualitative signal than conference calls.

Fast vs. Slow: The Full Picture. Mirroring Figure 2: numeric surprise peaks at *FirstActual* (IC = 0.03, Sharpe = 1.81) and decays by *NextOpen*; ECT sentiment peaks at *NextOpen* (IC = 0.11, Sharpe = 2.28). Fast numbers is arbitraged away; slow language persists overnight.

6.3 Analysis

Why fast numbers, slow language? Numeric surprise is a single scalar actionable by algorithms within minutes; language requires overnight interpretation before institutional investors can deploy their large capital pools.

What does LLM sentiment capture? Decomposing ECT into 23 fine-grained dimensions (Appendix I), three patterns emerge: backward-looking financial quality dominates (*cash flow* IC = 0.055 vs. *guidance direction* 0.020); analyst reception is informative (*analyst surprise/tone* IC = 0.037–0.038, reflecting real-time expert interpretation); and *surprise novelty* is negatively predictive (IC = −0.029), as novel disclosures create disagreement over direction.

	EPS(EN)	*Sent(ECT)*	*Sent(EN-agg)*
EPS(EN)	1.00	0.21	0.38
Sent(ECT)	0.21	1.00	0.52
Sent(EN-agg)	0.38	0.52	1.00

Table 3: Spearman rank correlations (2022–2025, $N \approx 5{,}000$; stable across splits).

Cross-signal correlations. The signals exhibit a nuanced correlation structure (Table 3). *Sentiment(ECT)* is largely orthogonal to *EPS(EN)* surprise ($\rho = 0.21$), confirming that ECT sentiment captures management tone and framing, not just the headline beat-or-miss. *Sentiment(EN-agg)* correlates more strongly with *EPS(EN)* ($\rho = 0.38$), as expected: post-announcement news articles co-report actuals alongside narrative, partially echoing the numeric outcome. *Sentiment(ECT)* and *Sentiment(EN-agg)* are moderately correlated ($\rho = 0.52$), both covering the same event.

pooled Spearman rank correlations (0.82/0.80), suggesting overlapping information despite performance differences.

Combining numbers and language. The two signals peak at incompatible horizons: *EPS(EN)* peaks at *FirstActual* but decays by *NextOpen*, where *Sentiment(ECT)* peaks. *Sentiment(ECT-informed)* trades off IC for Sharpe (IC = 0.06 vs. 0.11; Sharpe = 2.65 vs. 2.28), suggesting ECT already encodes the beat/miss outcome implicitly. Directly merging *Joint(EN)* or *EPS(EN)* with *Sentiment(ECT)* at *NextOpen* yields further drops in both metrics relative to *Sentiment(ECT)* (Appendix H). Using ECT sentiment to extend or exit intraday positions opened at *FirstActual*, dynamically bridging the fast/slow gap, is left to future work.

Selectivity and Coverage. Table 4 shows Sharpe rises as q decreases;[15] NLP evaluation implicitly uses q=0.50 (trade all events), the worst operating point; we use q=0.10 in our experiments.

q	0.50	0.40	0.30	0.20	0.10
Sent(ECT)	0.49	0.60	0.98	1.42	2.32
EPS(EN)	1.53	1.23	1.57	2.21	2.27

Table 4: Sharpe by quantile threshold q (q = top/bottom fraction traded), test set (2024–2025).

Model robustness. Claude Haiku 4.5 confirms the signal at IC = 0.08 ($t = 2.43$); Gemini 3 Flash confirms the signal at IC = 0.08 ($t = 2.77$), confirming robustness across LLM families [16].

Transaction costs. After 1 bp one-way per trade, *EPS(EN)* and *Sentiment(ECT)* retain substantial total returns of 42.4% and 51.8%, with further reduction possible via execution optimisation.

7 Conclusion

Two communities have studied earnings announcements for decades with incompatible frameworks. We bridge them through EARNINGSINONE and a unified evaluation framework, revealing a clean speed separation: numeric surprise peaks at announcement; ECT sentiment peaks at next-open, largely orthogonal. Both were capturing real but complementary phenomena, qualitative signals yield real trading returns when evaluated correctly. More broadly, our findings suggest that language signals may show better value when evaluated at their natural speed, a challenge shared by data-rich domains mixing structured and textual signals arriving asynchronously, such as healthcare (vital signs vs. clinical notes).

[15] *EPS(EN)* at 5 min; *Sentiment(ECT)* at 390 min.

[16] Claude Haiku 4.5 Model and Gemini 3 Flash Model

Limitations

Limitations. *Data:* licensing restricts public release of raw news and price data; we release metadata, IE annotations, and the alignment schema. *Scope:* experiments cover U.S. equity markets only; generalisation to other markets or asset classes is left to future work. *Backtesting:* results assume no transaction costs and perfect fill at 5-min bar open; strategies use canonical equal-weight sizing without position limits or risk controls, which may understate or overstate live performance. *LLM extraction:* systematic biases on financial jargon or low-coverage tickers are the primary failure mode ($<2\%$ of events, typically unit-scaling mismatches such as per-share vs. aggregate revenue); sentiment scoring uses a single-pass prompt without advanced prompt engineering (e.g. chain-of-thought, self-consistency), leaving room for improvement. *Stock relations:* inter-stock dependencies (e.g. sector contagion, supply-chain linkages) are not modelled; signals are treated independently per event. *Risks:* the trading signals described are intended for research purposes; deployment in live markets without proper risk management could result in financial losses. The intraday strategies may exacerbate short-term price volatility around earnings announcements. LLM extraction errors, if undetected, could generate misleading signals in automated pipelines.

Ethics Statement

All news and price data were obtained through licensed commercial APIs. No material non-public information was used. Trading results are academic backtests and do not constitute investment advice.

References

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

Azi Ben-Rephael, Steffen Hitzemann, and Yuanyuan Xiao. 2021. Mind the gap: The non-fundamental role of earnings days. *Available at SSRN 3966758*.

Victor L. Bernard and Jacob K. Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27(Supplement):1–36.

Mary Brooke Billings, Robert Jennings, and Baruch Lev. 2015. On guidance and volatility. *Journal of Accounting and Economics*, 60(2-3):161–180.

Yupeng Cao, Zhi Chen, Prashant Kumar, Qingyun Pei, Yangyang Yu, Haohang Li, Fabrizio Dimino, Lorenzo Ausiello, KP Subbalakshmi, and Papa Momar Ndiaye. 2025. Risklabs: Predicting financial risk using large language model based on multimodal and multisources data. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 901–909. IEEE.

Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Lee, KP Subbalakshmi, and Papa Momar Ndiaye. 2024. Ecc analyzer: Extracting trading signal from earnings conference calls using large language model for stock volatility prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 257–265.

Kim Christensen, Roel C. A. Oomen, and Roberto Renò. 2025. Warp speed price moves: Jumps after earnings announcements. *Journal of Econometrics*. Forthcoming.

Katharina Fink. 2021. A review of the post-earnings-announcement drift. *Journal of Behavioral and Experimental Finance*, 31:100551.

Richard C. Grinold and Ronald N. Kahn. 2000. *Active Portfolio Management*. McGraw-Hill.

Daniel Hoechle, Nic Schaub, and Markus Schmid. 2015. Time stamp errors and the stock price reaction to analyst recommendation and forecast revisions. *Available at SSRN 2768194*.

Narasimhan Jegadeesh and Joshua Livnat. 2006. Revenue surprises and stock returns. *Journal of Accounting and Economics*, 41(1):147–171.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

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

Chenhui Li and Weihai Lu. 2026. Decoding the market's pulse: Context-enriched agentic retrieval augmented generation for predicting post-earnings price shocks. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3055–3073, Rabat, Morocco. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13972–13980.

Charles Martineau. 2022. Rest in peace post-earnings announcement drift. *Critical Finance Review*, 11(3-4):613–646.

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.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Yu Qin, Chengshang Zhang, and Wei Xu. 2025. *Bridging Numbers and Narratives: Enhancing Financial Market Risk Predictions through Numerical Information from Financial Documents*, pages 307–316.

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.

Yung-Yu Shih, Yun-Nung Chen, and Chung-Chi Chen. 2025. Company-specific knowledge matters: Retrieval-augmented generation for earnings call answer rehearsal. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5243–5247.

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11604–11612.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.

Ding Yu, Zhuo Liu, and Hangfeng He. 2025. Same company, same signal: The role of identity in earnings call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18403–18422, Vienna, Austria. Association for Computational Linguistics.

A Detailed Community Comparison

The following elaborates on each dimension of Table 1 with a paragraph-level comparison across the financial economics community, the NLP community, and this work.

Target.

- **Financial economics community:** post-announcement return (the directional price move following the earnings release).
- **NLP community:** predominantly post-announcement volatility (the log rolling standard deviation of daily returns), which is sign-agnostic by construction; a minority of works target return directly.
- **This work (bridge):** post-announcement return, the most direct market reaction to earnings, fundamental to both trading (direction required for long/short) and volatility itself, which is computed from returns by construction.

Data & Signal.

- **Financial economics community:** third-party vendor (e.g. IBES, EODHD): numeric surprise in two forms: percentage (ΔEPS%) and standardised (SUE).[17]
- **NLP community:** ECTs, sentiment embeddings capturing management tone, credibility, and guidance without explicit numeric magnitude.
- **This work (bridge):** earnings news for IE-based numeric surprise; ECTs and news for LLM sentiment (§4).

Entry anchor.

- **Financial economics community:** enters at next-open, after intraday alpha has been priced in and multi-day PEAD has vanished (Martineau, 2022), missing the intraday window where surprise alpha now concentrates.
- **NLP community:** enters at next-open; qualitative signals peak overnight as investors process the earnings narrative.[18]
- **This work (bridge):** evaluates both intraday and next market open horizons, each signal at its natural speed (§5.3).

Trading.

- **Financial economics community:** dollar-neutral long/short : long top-decile SUE (beats), short bottom-decile SUE (misses).
- **NLP community:** no trading framework; predictions are evaluated over all announcing stocks, implicitly treating every event as equally tradeable.
- **This work (bridge):** dollar-neutral long/short on both signal types, following the top-decile/bottom-decile convention of the financial economics community.

Metric.

- **Financial economics community:** quintile spread (Q5−Q1) and cumulative abnormal return (CAR) (Bernard and Thomas, 1989), model-free measures matching canonical PEAD reporting.
- **NLP community:** MSE and direction accuracy, optimising the low-signal middle of the distribution rather than the high-conviction tails where alpha concentrates. The failure is structural: NLP evaluation is *pointwise* (each stock in isolation) while earnings trading is *cross-sectional* (rank all stocks in a session, long top, short bottom). A model with low MSE can produce a cross-sectional ranking that is pure noise.
- **This work (bridge):** IC and Sharpe : practitioner-standard metrics from quantitative asset management (Grinold and Kahn, 2000), complemented by Q5−Q1 (§6.1).

B Full Timestamp Offset Statistics

Figure 3 shows the empirical distribution of *FirstActual* and *NextOpen* anchor offsets relative to the earnings cutoff t_e, separately for BMO and AMC sessions.

For BMO announcements ($t_e \approx$ 07:00 ET, n=3,314), *FirstActual* arrives near-simultaneously with the earnings release (median offset = 1 min), well before market open; *NextOpen* is fixed at 09:30 ET (151 min after cutoff on average). The

[17] $\Delta\text{EPS}\%_i = (\text{EPS}_{\text{act}} - \text{EPS}_{\text{est}})/|\text{EPS}_{\text{est}}|$; $\text{SUE}_i = (\text{EPS}_{\text{act}} - \text{EPS}_{\text{est}})/\sigma(\varepsilon_i)$.

[18] ECT transcripts lack sentence-level timestamps, making intraday entry infeasible for ECT-based signals.

tight clustering around zero confirms that financial news rapidly co-reports actuals and estimates at announcement time.

For AMC announcements ($t_e \approx$ 16:00 ET, n=2,114), *FirstActual* arrives within minutes of the release (median = 11 min), while *NextOpen* is the following morning's open (median 1,051 min, ≈17.5 hours). The gray shading marks the overnight window where no trading occurs. The large gap between *FirstActual* and *NextOpen* for AMC sessions motivates the intraday entry strategy: alpha is available immediately after the release but is largely absorbed by next open.

Session definitions. A trading session is defined at the (*date*, *before-/after-market*) level: all earnings events sharing the same calendar date and market window (BMO or AMC) belong to the same session. *FirstActual* is anchored to the first news article with confirmed EPS/revenue actuals within that session. *NextOpen* is defined as the next regular-session open (`next_open_dt`): for AMC announcements, this is 09:30 ET on $t_0 + 1$; for BMO announcements on day t_0, *NextOpen* consolidates both the preceding AMC session (date $t_0 - 1$, after-market) and the current BMO session (date t_0, before-market), anchoring all trades at the 09:30 ET open on t_0.

C Trading Pipeline and Implementation Details

C.1 Embedding Baseline: MLP Pipeline

The *Embedding* baseline freezes the text-embedding-3-large (OpenAI, 3072-d) encoder and trains a lightweight MLP on top.

Architecture.

```
Embedding [3072] → StandardScaler
                 → Linear(3072, 256) + ReLU
                 → Linear(256, 128)  + ReLU
                   → Linear(128, 1)
```

Adam optimizer ($\lambda = 10^{-4}$ weight decay), batch size 512, auto-detects CUDA/MPS/CPU. Two tasks are trained per fold:

- **Classification (sign variant):** BCEWithLogitsLoss; output signal $= P(\text{return} > 0) - 0.5 \in [-0.5, 0.5]$.
- **Regression (magnitude variant):** MSELoss; output signal = predicted return.

Rolling-window protocol. For each test month t:

1. **Split:** train $= [t - T_{\text{train}},\, t - T_{\text{val}} - 1]$; val $= [t - T_{\text{val}},\, t - 1]$; test = month t.
2. **Horizon selection:** train both tasks for each candidate horizon $H \in \{5, 15, 30, 60, 120, 180, 390\}$ min on pure-train; select $H^* = \arg\max_H \text{IC}_{\text{val}}$ (pooled Spearman rank IC on val).
3. **Retrain:** refit selected-horizon models on train ∪ val combined.
4. **Predict:** append out-of-sample predictions for month t.

Final evaluation follows the same IC, Q5−Q1, and Sharpe metrics as all other strategies (§6.1).

C.2 Trading Pipeline: Feature → Signal → Weight → Position

Phase 1–2: Raw Features to Standardised Signals. EPS surprise:

$$\text{eps_surp}_i = \frac{\text{act_eps}_i - \text{est_eps}_i}{\text{est_eps}_i}$$

Standardised unexpected earnings (SUE):

$$\text{SUE}_i = \frac{\text{eps_surp}_i}{\sigma(\varepsilon_i)}$$

where $\sigma(\varepsilon_i)$ is the rolling standard deviation of historical forecast errors (window = 8 quarters, min 4), per ticker, shifted one quarter to prevent look-ahead. Revenue surprise (SUR) is computed analogously.

Note: *FirstActual* strategies use ΔEPS% (percentage surprise relative to estimate) as the primary signal, capturing the immediate market-relevant deviation; *NextOpen* strategies use SUE (standardised by historical forecast error), as overnight processing allows incorporating the stock's historical surprise distribution for better cross-sectional comparability.

Joint signal variants. The blended signal $\text{joint}_i = \alpha \cdot \text{SUE}_i + (1 - \alpha) \cdot \text{SUR}_i$ admits two gating variants, differing only in the position entry condition:

- **Joint**: position weight w_i is assigned whenever $|\text{joint_pctl}_i - 0.5| \geq \tau - 0.5$, using a per-sector expanding-window CDF fitted on the training set.

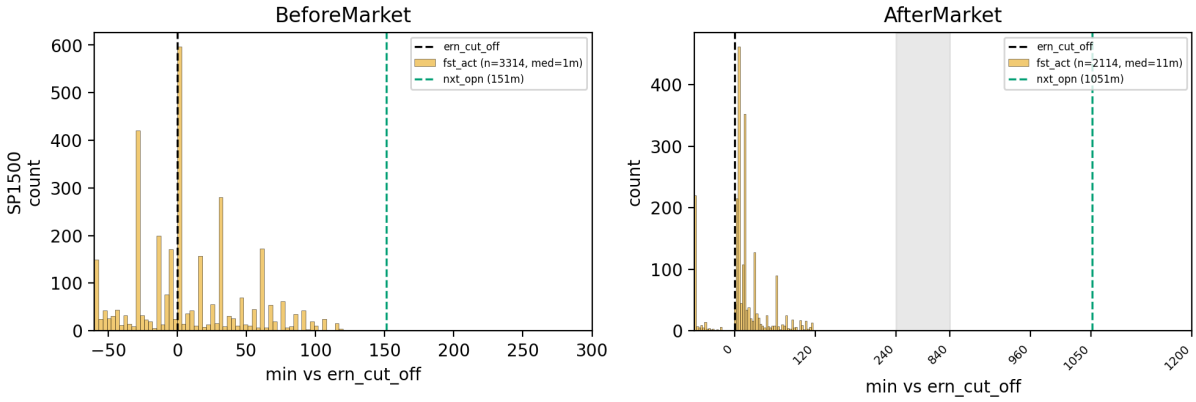

Figure 3: Anchor timestamp offset from earnings cutoff t_e (minutes) for BMO (before-market open, $t_e \approx$ 07:00 ET, left) and AMC (after-market close, $t_e \approx$ 16:00 ET, right) sessions. *fst_act* = *FirstActual* anchor (median 1 min for BMO, 11 min for AMC); *nxt_opn* = *NextOpen* anchor (09:30 ET open, 151 min after BMO cutoff, 1051 min after AMC cutoff). Gray shading (AMC) marks the overnight window between market close and next open. BMO news arrives near-simultaneously with the earnings release; AMC news arrives within minutes of the release but trades at the next morning's open.

- **Confirmed**: same percentile gate as Joint, with an additional directional agreement filter: a position is opened only when $\text{sign}(\text{SUE}_i) = \text{sign}(\text{SUR}_i)$, requiring EPS and revenue surprises to point in the same direction. This reduces coverage but filters mixed-signal events where EPS and revenue surprises disagree.

Both variants use the same α and τ selected on the validation set (2023) and share the same market-wide joint percentile CDF.

Phase 3: Signals to Percentile Ranks. An expanding-window empirical CDF fitted market-wide transforms each signal into a percentile $p_i \in [0, 1]$:

$$p_i = \hat{F}_{\text{market}}(\text{SUE}_i)$$

The CDF refreshes monthly after the training cutoff.

Phase 4: Percentiles to Weights (Historical Percentile). Position weight:

$$w_i = \frac{G}{N} \times 2 \times (p_i - 0.5) \quad \text{if } |p_i - 0.5| \geq \tau - 0.5$$

$$w_i = 0 \quad \text{otherwise}$$

where $G = 2.0$ (gross target: \$1 long + \$1 short per session), N = number of events in the session, $\tau = 0.9$ (fixed, top/bottom 10% trade). Dollar-neutral by construction.

Phase 4 (PEAD): Cross-Sectional Sort. Long leg (top decile):

$$w_i = +\frac{G/2}{N_{\text{long}}}$$

Short leg (bottom decile):

$$w_i = -\frac{G/2}{N_{\text{short}}}$$

Middle 80%:

$$w_i = 0$$

Sessions with fewer than 10 valid events are skipped.

Hyperparameter Selection. Each signal is evaluated at horizons $H \in \{5, 15, 30, 60\}$ min for *FirstActual* and $H \in \{5, 15, 30, 60, 120, 180, 390\}$ min for *NextOpen*. The blending weight $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ is selected on the validation set (2023) by maximising portfolio Sharpe ratio, subject to per-session IC > 0 to ensure positive signal direction, and held fixed for test (2024–2025); τ=0.9 is fixed throughout.

Phase 5–6: Execution. Portfolio PnL per session:

$$\text{PnL}_d = \sum_i w_i \times r_{i,H}$$

where $r_{i,H}$ is the forward return at horizon $H \in \{5, 15, 30, 60, 120, 180, 390\}$ min. Risk controls (sector caps, exposure limits) are disabled by default.

D Evaluation Metric Definitions

IC (Information Coefficient). Spearman rank correlation between the signal and realized forward returns, measuring signal quality independently of position sizing.

Per-session IC (primary metric): computed cross-sectionally within each event date d:

$$\mathrm{IC}_d = \mathrm{Spearman}\big(\mathbf{s}_d,\, \mathbf{r}_d\big)$$

where $\mathbf{s}_d$ = signal scores and $\mathbf{r}_d$ = forward returns for all stocks announcing on date d. The reported IC is the time-average $\mathrm{E}[\mathrm{IC}_d]$ across all sessions.

Pooled IC: computed across all events simultaneously (used in Figure 2 for characterising signal decay); typically higher than per-session IC due to cross-sectional variation in base returns. t_{NW} **(Newey-West t-statistic)**: tests whether $\mathrm{E}[\mathrm{IC}_d] \neq 0$, correcting for serial correlation across sessions:

$$t_{\mathrm{NW}} = \frac{\bar{\mathrm{IC}}}{\hat{\sigma}_{\mathrm{NW}}/\sqrt{T}} \tag{1}$$

where $\bar{\mathrm{IC}}$ is the time-averaged IC, T is the number of sessions, and $\hat{\sigma}_{\mathrm{NW}}$ is the Newey-West standard deviation correcting for serial correlation up to lag L (we use $L = 5$). $|t_{\mathrm{NW}}| > 2$ corresponds to $p < 0.05$.

Q5−Q1 bp (t). Fama-MacBeth session-averaged quintile spread:

$$\mathrm{Q5{-}Q1}_d = \bar{r}_{d,Q5} - \bar{r}_{d,Q1}$$

where $\bar{r}_{d,Q5}$ and $\bar{r}_{d,Q1}$ are mean returns of the top and bottom quintile on date d, expressed in basis points (bp). The t-statistic uses Fama-MacBeth standard errors (Bernard and Thomas, 1989). Measures tail separability independent of the full cross-section.

Sharpe Ratio.

$$\mathrm{Sharpe} = \frac{\bar{R}_p - R_f}{\sigma_p}$$

where $\bar{R}_p$ is mean annualised portfolio return, R_f the risk-free rate (set to zero for simplicity), and σ_p the annualised return standard deviation. Measures overall strategy quality by incorporating signal quality, selectivity, and sizing.

Tot. Ret. Cumulative total return over the test period (2024–2025), expressed as a percentage. Under dollar-neutral long/short sizing, equivalent to market-adjusted cumulative abnormal return (CAR).

N**.** Number of trades (long + short positions) over the test period.

E Extraction Validation: News vs. EOD Vendor

Figures 4 and 5 compare LLM-extracted EPS and revenue values from EARNINGSINONE against ground-truth values from an EOD financial data vendor, across all four equity universes.

F Dataset Construction Details

News Collection News articles are sourced from a licensed financial wire aggregator. We filter to earnings-related articles by requiring at least one of: (a) ticker mention in headline or lede matching the announcing company, (b) keyword match on earnings-related terms ("earnings", "EPS", "quarterly results") Duplicate articles (same headline within a 10-minute window) are deduplicated by URL hash; the collection window spans [−60,min, +120,min] around each 8-K filing timestamp.

Privacy and Content. Financial news articles and earnings conference call transcripts may contain names of corporate executives and financial analysts acting in their professional capacity; these are already part of the public record and do not constitute private personally identifiable information (PII). No private individual data is collected or released. To further limit exposure, we release article IDs and extracted metadata only, not full article text. No offensive content was identified in the corpus; all sources are regulated financial media outlets subject to editorial standards.

Dataset Documentation. EARNINGSINONE covers English-language U.S. financial news and earnings conference call transcripts for SP 1500 companies (2022–2025). The corpus is domain-specific (financial reporting) and monolingual (English). Company coverage skews toward large- and mid-cap firms (SP 500 and SP 400 represent 98% of events); SP 600 small-cap firms are underrepresented due to lower news coverage. No individual-level demographic data is present; all entities are publicly listed U.S. corporations

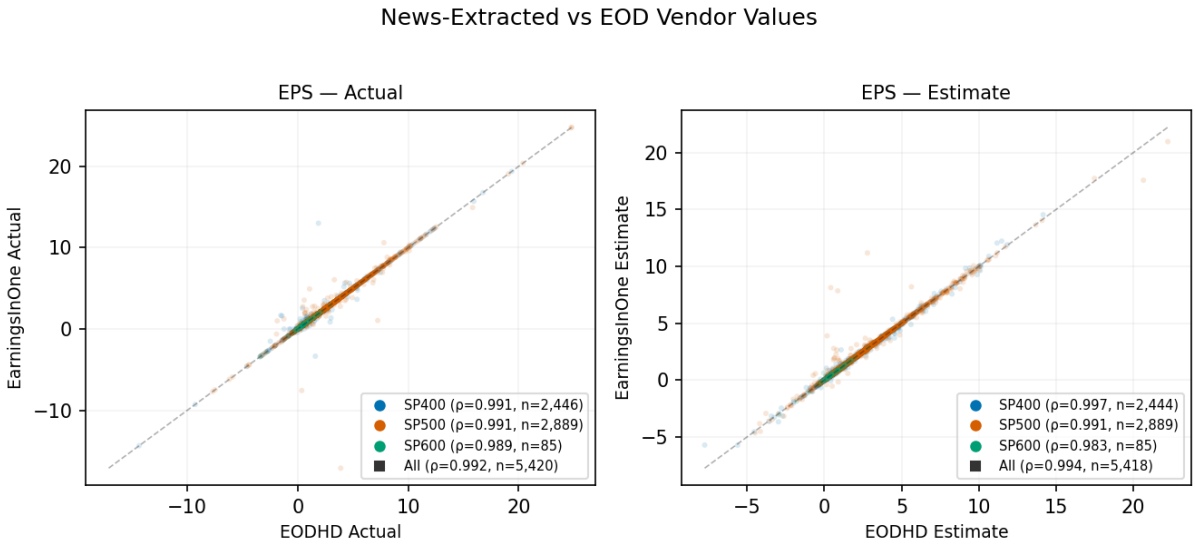

Figure 4: Scatter plots of EARNINGSINONE LLM- extracted values vs. EODHD vendor values for EPS actual (left) and EPS estimate (right), by index. Spearman $\rho \geq 0.989$ for actuals and $\rho \geq 0.983$ for estimates across all indices, confirming near-perfect agreement. Dashed line = identity.

subject to SEC reporting requirements. Linguistic phenomena include financial jargon, numerical expressions, forward-looking statements, and structured analyst Q&A dialogue.

Data Consent. All data in EARNINGSINONE is sourced from publicly available corporate disclosures and licensed commercial APIs. Earnings conference call transcripts are publicly disclosed under SEC Regulation FD, which mandates that material information be made available to all investors simultaneously. Financial news articles and price data are accessed via the EODHD Historical Data API under a commercial license agreement. No individual consent is required as all sources constitute public corporate records.

Ethics Review. This work does not involve human subjects research, collection of personal data, or any procedures requiring ethics board review. All data sources are publicly available corporate disclosures or licensed commercial data products.

Dataset Statistics Table 5 reports event and ticker counts by index and temporal split. The corpus totals 5,428 earnings events across 701 unique tickers over 2022–2025. SP 500 and SP 400 dominate (2,897 and 2,446 events respectively), while SP 600 contributes only 85 events ($\sim$1.6% of the corpus), reflecting the lower financial news coverage of small-cap firms. The test set (2024–2025) is the largest split by design, covering 2,468 events across 640 unique tickers. Ticker counts across splits sum to more than the "All" column because some tickers report in multiple years.

G Prompt Templates

Full prompt templates for IE extraction and sentiment scoring are available in the released code repository. IE extraction uses a structured JSON output schema with chain-of-thought reasoning; sentiment scoring uses a single-pass instruction with a $[-1, +1]$ scale and verbal anchors at ± 0.5 and ± 1.0.

ECT Sentiment Scoring Prompt.

> You are a financial analyst predicting how the market will react to this earnings call. You will be given the full transcript of a company's quarterly earnings call.
>
> **Your task:** assess whether the call conveys *incrementally* positive or negative information — i.e., things that would surprise the market beyond what was already expected from the headline numbers (EPS/revenue beat or miss).
> Focus on (in order of importance):
>
> 1. **Q&A section** — analyst tone (hostile vs. friendly), management evasiveness or defensiveness, unexpected admissions, stumbling on questions
> 2. **Forward guidance** — specific vs. vague, raised/maintained/lowered, hedging language ("we hope", "we'll try" vs. "we will", "we expect")
> 3. **Tone shifts** — prepared remarks upbeat but Q&A defensive = bearish signal
> 4. **Incremental disclosure** — new risks mentioned, customer losses, pipeline changes, margin pressure, one-time items framed as recurring

	Earnings				Tickers			
	Train (2022)	Val (2023)	Test (2024-25)	All	Train (2022)	Val (2023)	Test (2024-25)	All
SP400	768	589	1089	2446	277	252	304	354
SP500	810	766	1321	2897	285	280	319	346
SP600	11	16	58	85	5	12	29	31
SP1500	1589	1371	2468	5428	565	541	640	701

Table 5: EarningsInOne dataset statistics by index and split. Earnings = number of earnings events; Tickers = number of unique tickers. SP 600 (small-cap) is underrepresented; small-cap firms attract less financial news coverage, consistent with their lower analyst coverage and trading volume. Strict temporal split: train 2022, validation 2023, test 2024–2025.

Scoring guidance (use the full range):

- sentiment near $+1.0$: management clearly signalling acceleration, raising guidance with specifics, comfortable Q&A, analysts impressed
- sentiment near $+0.3$: slightly better than expected tone, modest positive guidance, no red flags
- sentiment near 0.0: neutral, in-line, no surprises either direction
- sentiment near -0.3: slight hedging, vague guidance, some analyst pushback
- sentiment near -1.0: defensive Q&A, guidance cut, evasive on key metrics, analyst frustration visible

Respond with a JSON object:

```
{ "sentiment": <float between -1.0 and 1.0> }
```

ECT Decomposition Scoring Prompt.

You are a financial analyst predicting how the market will react to this earnings call. You will be given the full transcript of a company's quarterly earnings conference call.

Your task: score each of the following dimensions independently, based *only* on what is discussed in the call. Each dimension captures a distinct aspect of the call's information content — score them as independently as possible.
Focus on (in order of importance):

1. **Q&A section** — analyst tone (hostile vs. friendly), management evasiveness or defensiveness, unexpected admissions, stumbling on questions
2. **Forward guidance** — specific vs. vague, raised/-maintained/lowered, hedging language ("we hope", "we'll try" vs. "we will", "we expect")
3. **Tone shifts** — prepared remarks upbeat but Q&A defensive = bearish signal
4. **Incremental disclosure** — new risks mentioned, customer losses, pipeline changes, margin pressure, one-time items framed as recurring

Scoring guidance (use the full range):

- **Directional dimensions** (`eps_beat_framing`, `rev_quality`, `margin_trajectory`, `cash_flow_quality`, `guidance_direction`, `long_term_outlook`, `demand_momentum`, `pricing_power`, `market_share`, `investment_cycle`, `competitive_threat`, `regulatory_risk`, `supply_chain_health`, `prepared_vs_qa_gap`, `analyst_tone`, `analyst_surprise`): near $+1.0$ = strongly positive for the stock; near 0.0 = neutral/in-line; near -1.0 = strongly negative for the stock
- **Intensity dimensions** (`guidance_specificity`, `guidance_conviction`, `macro_sensitivity`, `mgmt_confidence`, `transparency`, `surprise_novelty`, `qa_depth`): near 1.0 = strong presence; near 0.0 = absent/minimal
- Use `null` for any dimension not assessable from the transcript

Respond with a JSON object:

```
{ "eps_beat_framing":
  <float -1.0 to 1.0 or null>,
  "rev_quality":
  <float -1.0 to 1.0 or null>,
  "margin_trajectory":
  <float -1.0 to 1.0 or null>,
  "cash_flow_quality":
  <float -1.0 to 1.0 or null>,
  "guidance_direction":
  <float -1.0 to 1.0 or null>,
  "guidance_specificity":
  <float  0.0 to 1.0 or null>,
  "guidance_conviction":
  <float  0.0 to 1.0 or null>,
  "long_term_outlook":
  <float -1.0 to 1.0 or null>,
  "demand_momentum":
  <float -1.0 to 1.0 or null>,
  "pricing_power":
  <float -1.0 to 1.0 or null>,
  "market_share":
  <float -1.0 to 1.0 or null>,
  "investment_cycle":
  <float -1.0 to 1.0 or null>,
  "competitive_threat":
  <float -1.0 to 1.0 or null>,
  "regulatory_risk":
  <float -1.0 to 1.0 or null>,
  "supply_chain_health":
  <float -1.0 to 1.0 or null>,
  "macro_sensitivity":
```

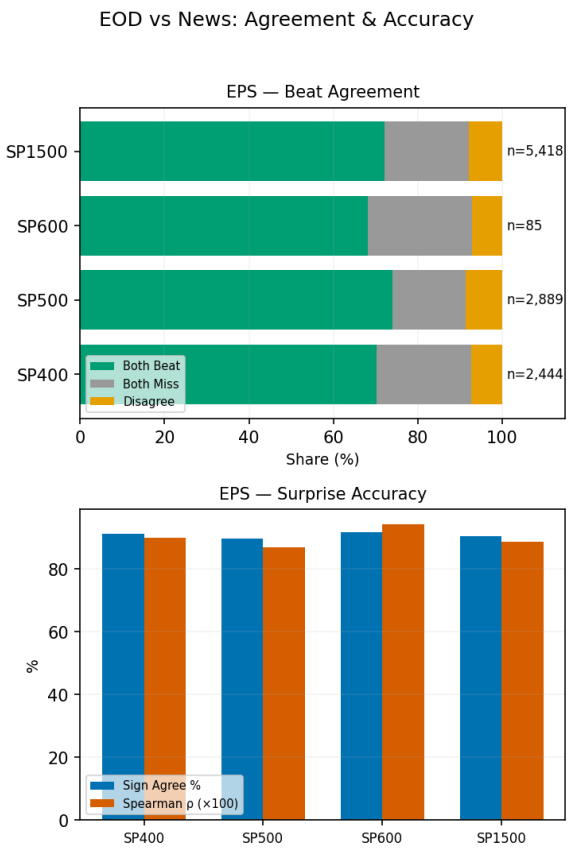

Figure 5: Beat/miss agreement and surprise accuracy between EARNINGSINONE and EODHD. **Top:** stacked bars show the share of both-beat, both-miss, and disagreement events; overall agreement exceeds 90% across all indices. **Bottom:** sign agreement rate (%) and Spearman ρ ($\times 100$) by index, both consistently above 85%, confirming reliable directional extraction.

```
    <float  0.0 to 1.0 or null>,
    "mgmt_confidence":
    <float  0.0 to 1.0 or null>,
    "transparency":
    <float  0.0 to 1.0 or null>,
    "prepared_vs_qa_gap":
    <float -1.0 to 1.0 or null>,
    "surprise_novelty":
    <float  0.0 to 1.0 or null>,
    "analyst_tone":
    <float -1.0 to 1.0 or null>,
    "analyst_surprise":
    <float -1.0 to 1.0 or null>,
    "qa_depth":
    <float  0.0 to 1.0 or null> }
```

H Full Results Table

Table 6 reports all quantitative and qualitative variants, including those omitted from the main table for concision. The table is organised into four groups. **Naïve baselines** provide no-signal floors: passive index benchmarks (SPY, QQQ), random long/short, long-only, and two prior-NLP baselines (Ticker-Mean and ECT Embedding). **Quantitative (Fast)** variants isolate four factors: ranking mechanism (PEAD cross-sectional sort vs. historical percentile), entry anchor (*FirstActual* vs. *NextOpen*), source breadth (EPS-only vs. Joint EPS+revenue), and data quality (LLM-extracted EN vs. vendor EOD). **Qualitative (Slow)** variants cover ECT sentiment under multiple ablations (number masking, numeric grounding, alternative LLM backends, prepared-remarks-only, Q&A-only) and news sentiment aggregated to *NextOpen*. **Combined (Combo)** variants directly merge ECT sentiment with quantitative signals at *NextOpen*, isolating the effect of signal combination at a shared entry anchor.

Category	Strategy	Entry anchor	IC (t_{NW})	Q5−Q1 bp (t)	Sharpe	Tot. Ret.	N
Naïve	Random L/S	NextOpen	—	—	−0.88	−10.5%	2426
	Long-only	NextOpen	—	—	0.11	5.8%	2077
	Ticker-Mean	NextOpen	—	—	−3.13	−2.2%	2360
	Embedding (sign)	NextOpen	—	—	0.80	23.3%	2468
	Embedding (mag)	NextOpen	—	—	0.56	13.9%	2468
Quantitative (Fast)	*PEAD-EPS(EN)*	NextOpen	0.03 (0.97)	33.9 (1.36)	3.31	29.1%	380
	PEAD-Joint(EN)	NextOpen	0.03 (0.84)	42.1 (1.60)	3.18	25.2%	326
	EPS(EN)	NextOpen	−0.02 (−0.56)	−5.4 (−0.12)	−0.31	−4.2%	400
	Joint(EN)	NextOpen	−0.02 (−1.03)	19.87 (0.47)	0.23	6.3%	242
	EPS(EN)	FirstActual	0.03 (0.94)	**137.7 (3.54)**	1.81	50.4%	400
	Joint(EN)	FirstActual	0.05 (1.58)	61.3 (1.94)	1.79	39.9%	296
	Joint(EOD)[†]	FirstActual	**0.08 (2.40)**	118.2 (4.13)	**2.43**	**64.5%**	343
Qualitative (Slow)	*Sentiment(ECT)*	NextOpen	**0.11 (4.04)**	153.0 (4.30)	2.28	63.6%	591
	Sentiment(ECT-mask)	NextOpen	0.07 (2.62)	118.7 (3.96)	0.40	7.0%	412
	Sentiment(ECT-informed)	NextOpen	0.06 (2.10)	94.1 (2.62)	**2.65**	**72.0%**	599
	Sentiment(ECT-Haiku)	NextOpen	0.08 (2.43)	**213.2 (4.01)**	1.55	31.4%	455
	Sentiment(ECT-Gemini)	NextOpen	0.08 (2.77)	71.1 (2.01)	1.16	12.6%	110
	Sentiment(EN-agg)	NextOpen	0.05 (1.89)	−4.6 (−0.24)	0.42	5.9%	496
	Sentiment(ECT-Prepared)	NextOpen	0.05 (1.69)	31.5 (0.98)	2.15	40.8%	430
	Sentiment(ECT-QA)	NextOpen	0.07 (2.42)	76.6 (2.83)	1.47	27.9%	492
	PEAD-Sentiment(ECT)	NextOpen	0.08 (2.70)	136.4 (3.94)	−0.60	−8.1%	578
	PEAD-Sentiment(ECT-mask)	NextOpen	0.04 (1.25)	18.9 (1.21)	−0.92	−2.2%	568
	PEAD-Sentiment(ECT-informed)	NextOpen	0.09 (2.76)	41.2 (1.77)	1.63	11.5%	590
	PEAD-Sentiment(ECT-Haiku)	NextOpen	0.0713 (2.59)	138.3 (4.23)	-0.15	-2.0%	621
	PEAD-Sentiment(ECT-Gemini)	NextOpen	0.0603 (2.22)	20.7 (1.06)	0.77	9.5%	561
Combined (Combo)	*Sentiment(ECT)+Joint(EN)*	NextOpen	0.05 (1.99)	20.3 (0.79)	0.59	6.3%	271
	Sentiment(ECT)+EPS(EN)	NextOpen	0.06 (2.37)	86.3 (2.00)	0.12	2.1%	437

Table 6: Out-of-sample trading performance, test set (2024–2025 with 2,468 events), SP 1500. **Entry anchors:** *FirstActual* = first news article with confirmed EPS/revenue actuals; *NextOpen* = 09:30 am on the following trading day. **Metrics:** IC (t_{NW}) = Spearman rank IC (Grinold and Kahn, 2000) computed per event date (t_{NW}: Newey-West t-statistic correcting for serial correlation); Q5−Q1 bp (t) = Fama-MacBeth date-averaged quintile spread (Bernard and Thomas, 1989) (t: Fama-MacBeth standard errors); $|t| > 2$ corresponds to $p < 0.05$ for both; N = number of trades. **Signals:** *Joint* = EPS + revenue surprise combined; [†]EOD: EODHD Historical Data pre-release estimates; professional-grade vendor actuals with superior data quality, but not available in real time for live deployment; *Embedding* (sign/mag variants, text-embedding-3-large, OpenAI) = ECT embeddings passed through a 2-layer MLP; *PEAD-Sentiment(ECT)* variants = ECT sentiment strategies using canonical PEAD contemporaneous cross-sectional sort (vs. historical percentile for vanilla ECT); *Sentiment(ECT-Prepared)* = sentiment scored over prepared remarks only; *Sentiment(ECT-QA)* = sentiment scored over Q&A section only; *Sentiment(ECT-Haiku/Gemini)* = ECT sentiment using alternative LLM backends; *Combo* = direct merger of *Sentiment(ECT)* with quantitative signal at *NextOpen*. All other variants use historical percentile calibration unless labelled PEAD.s **Metric Interpretation:** IC measures overall signal quality (does the signal rank stocks correctly?); Q5−Q1 measures tail separability (do the extreme quintiles trade well?); Sharpe measures full strategy performance (does it make money risk-adjusted?). High IC with low Sharpe indicates the signal ranks stocks well but individual trade returns are noisy and volatile; high Sharpe with low IC signals selectivity-driven returns (strong performance concentrated in a small number of high-conviction trades). **Bold** = best per column within group. **Passive Market benchmarks:** SPY and QQQ are the most widely traded ETFs tracking the S&P 500 and Nasdaq-100, serving as buy-and-hold baselines: SPY total return 48% (Sharpe 1.02); QQQ total return 54% (Sharpe 0.94).

I ECT Decomposition: Per-Dimension IC

Table 7 reports Spearman IC at 390 min (*NextOpen*) for each of the 23 orthogonal ECT dimensions, computed on the full dataset (2022–2025). Dimensions are scored independently by a structured LLM prompt (Appendix G); see §6.3 for interpretation.

The 23 dimensions span six categories: **A. Financial Performance** (backward-looking): EPS beat framing, revenue quality, margin trajectory, cash flow quality — capturing how management characterises reported results and whether earnings quality is sustainable. **B. Forward Guidance**: guidance direction (raised/maintained/lowered), specificity, conviction, long-term outlook — measuring the strength and credibility of management's forward outlook. **C. Growth & Demand**: demand momentum, pricing power, market share, investment cycle — reflecting whether the business is accelerating or decelerating in its competitive environment. **D. Risk & External**: competitive threat, regulatory risk, supply chain health, macro sensitivity — assessing exposure to external forces beyond management control. **E. Management Communication Quality**: overall confidence, transparency, prepared-vs-Q&A tone gap, surprise novelty — evaluating how clearly and honestly management communicates, and whether new information is disclosed. **F. Analyst Reception**: analyst tone, analyst surprise, Q&A depth — capturing real-time expert reaction as a proxy for institutional investor sentiment. Each dimension is scored on $[-1, +1]$ independently, capturing incremental information beyond the headline EPS/revenue outcome.

Table 7: Per-dimension IC at 390 min (*NextOpen*) for the 23 ECT decomposition dimensions, SP 1500 (full dataset, 2022–2025). Ordered by IC descending within each category. Positive = directional signal toward next-open return; negative = contrarian or noise.

Category	Dimension	IC
A. Financial Performance	Cash flow quality	0.055
	Margin trajectory	0.022
	EPS beat framing	0.025
	Revenue quality	0.018
B. Forward Guidance	Guidance direction	0.020
	Guidance specificity	0.012
	Guidance conviction	0.008
	Long-term outlook	0.007
C. Growth & Demand	Market share	0.028
	Pricing power	0.023
	Demand momentum	0.001
	Investment cycle	−0.003
D. Risk & External	Competitive threat	0.041
	Regulatory risk	0.021
	Supply chain health	0.020
	Macro sensitivity	−0.007
E. Management Communication	Mgmt confidence	0.021
	Transparency	0.014
	Prepared vs. Q&A gap	0.004
	Surprise novelty	−0.029
F. Analyst Reception	Analyst surprise	0.038
	Analyst tone	0.037
	Q&A depth	−0.009