Trading the Tonality Dispersion in Earnings Call Q&A Sections

Mert Ülgüner
Linze Li

April 14, 2026

Abstract

We construct a novel text-based measure of disagreement using earnings call Q&A sections. For each transcript, we compute the dispersion of sentiment across analyst questions, management answers, and their differences, where sentiment scores are obtained using a financial-domain language model (FinEAS). We show that these dispersion measures capture information uncertainty and disagreement between analysts and managers. Using a weekly rebalanced portfolio strategy, we find that firms with higher sentiment dispersion significantly outperform the market. A strategy based on the intersection of multiple dispersion signals delivers the strongest performance, with a Sharpe ratio substantially higher than the S&P 500 benchmark. Our results highlight the importance of information heterogeneity in earnings calls and show that textual disagreement can be systematically exploited to generate significant excess returns.

Keywords: earnings calls, textual disagreement, sentiment dispersion, Q&A, FinEAS

1 Introduction

Earnings calls are a major channel through which firms communicate with market participants. While prior studies mostly emphasize the average tone or informational content of management disclosures, the interactive Q&A section contains an additional and potentially richer dimension: disagreement. Analysts often ask questions that differ in emphasis, concern, and sentiment, while managers respond with varying degrees of clarity, confidence, and optimism. Such heterogeneity may reflect information uncertainty, disagreement in interpretation, or communication frictions between analysts and managers.

This paper studies whether disagreement embedded in earnings call Q&A sections predicts future returns. We construct transcript-level disagreement measures by first extracting sentiment at the Q&A-pair level using FinEAS and then computing the dispersion of sentiment across the set of questions and answers within each transcript. Specifically, for each transcript we calculate three measures: the standard deviation of analyst-question sentiment, the standard deviation of management-answer sentiment, and the standard deviation of the difference between answer and question sentiment. These measures capture, respectively, disagreement among analysts, inconsistency in management communication, and mismatch between analyst framing and management responses.

We then examine whether these text-based disagreement signals have cross-sectional return predictability. Using weekly portfolio rebalancing, we rank firms by each disagreement measure and form long portfolios based on the top two bins. We also construct intersection and union strategies across multiple disagreement signals. The intersection strategy, which focuses on firms that simultaneously exhibit high disagreement across several dimensions, delivers the strongest performance and substantially outperforms the market benchmark.

Our paper makes three contributions. First, we introduce a new Q&A-based disagreement framework that moves beyond average tone and captures information heterogeneity within a transcript. Second, we show that the cross-sectional dispersion of textual sentiment contains economically meaningful and profitable information. Third, we demonstrate that combining multiple disagreement dimensions through an intersection strategy enhances signal quality and delivers superior portfolio performance.

2 Data and Preprocessing

2.1 Earnings Call Transcripts

Our sample consists of 16,214 earnings call transcripts. Each transcript is segmented into Q&A pairs, where each pair contains one analyst question and the corresponding management answer. Let j index transcripts and $\ell = 1, 2, \ldots, m_j$ index Q&A pairs within transcript j.

Field	Description
Transcript Date	Date of the earnings call
Ticker	Firm identifier
Q&A Pairs	Analyst question and corresponding management answer

Table 1: Key identifiers retained for each transcript.

These transcripts are extracted from Refinitiv workspace by filtering on S&P CBOE index where we believe the constituents are updated dynamically, preventing the analysis from suffering from **survivorship bias**.

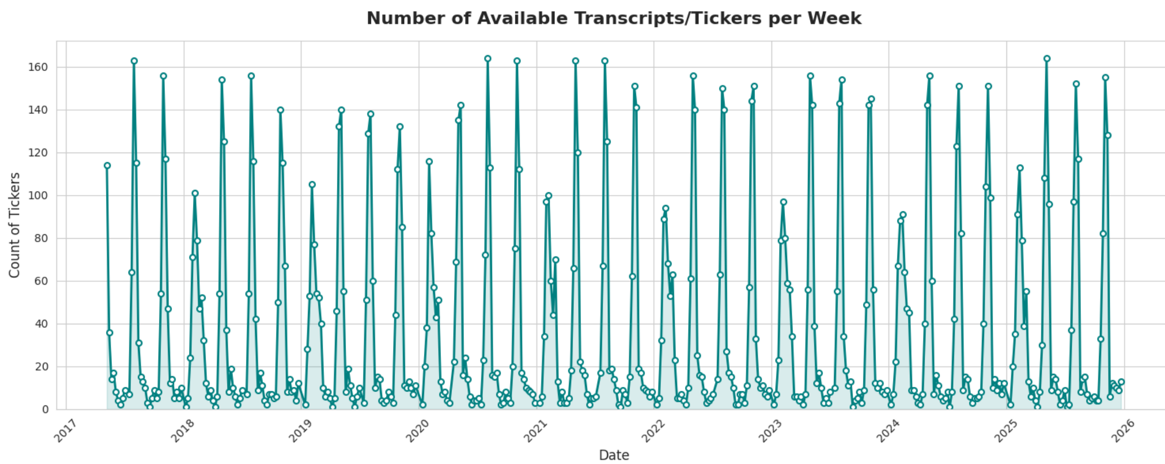


Figure 1: Weekly number of transcripts across the historical data

Earnings calls are released in a quarterly fashion where the periods are refered to as earnings season as can be seen in Figure 1.

2.2 Sentiment Extraction

For each question and answer, we compute text-based sentiment scores using pretrained financial language models. In our implementation, we choose the **Financial Embedding Analysis of Sentiment (FinEAS) model** [1] because it is fine-tuned on financial text data and employs a regression-based output layer, which enables continuous sentiment scoring rather than discrete classification. The use of this regression-based sentiment model is particularly important in our setting, as it allows us to capture fine-grained variations in tone across Q&A interactions, which are essential for measuring within-transcript dispersion.

[1] The model is available via the HuggingFace repository: https://huggingface.co/LHF/FinEAS

Let

$$S_{j,\ell}^{Q} := f_{\text{sent}}(Q_{j,\ell}),$$
$$S_{j,\ell}^{A} := f_{\text{sent}}(A_{j,\ell}),$$

where $f_{\text{sent}}(\cdot)$ denotes the sentiment scoring function derived from the underlying language model.

Our approach leverages model-based sentiment representations that incorporate contextual embeddings and attention mechanisms.

We further define the **Tone Difference** between analysts and managers as

$$S_{j,\ell}^{D} = S_{j,\ell}^{A} - S_{j,\ell}^{Q}. \tag{1}$$

Intuitively, $S_{j,\ell}^{D}$ captures the extent to which management responds with a more optimistic or pessimistic tone relative to the analyst's framing of the issue.

3 Methodology

3.1 Project Pipeline

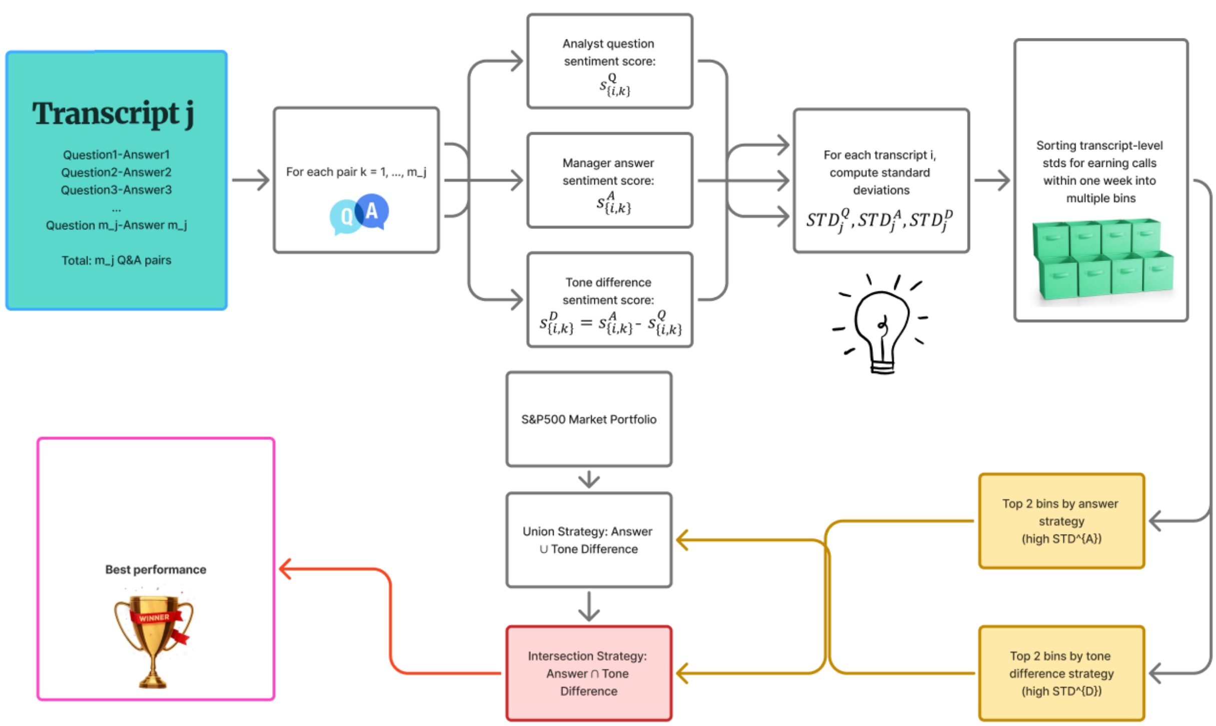


Figure 2: Overview of the earnings call disagreement pipeline.

Figure 2 illustrates how Q&A pairs are transformed into sentiment scores, aggregated into dispersion measures, and used to construct portfolio strategies.

3.2 Transcript-Level Disagreement Measures

For each transcript j, containing m_j Q&A pairs, we aggregate the pair-level sentiment scores into transcript-level disagreement measures using cross-pair dispersion.

Question sentiment dispersion. We define analyst-question sentiment dispersion as

$$\mathrm{STD}_j^Q = \mathrm{std}\left(S_{j,1}^Q, S_{j,2}^Q, \ldots, S_{j,m_j}^Q\right). \tag{2}$$

This measure captures disagreement among analysts in how they frame the firm's outlook, performance, or risks.

Answer sentiment dispersion. We define management-answer sentiment dispersion as

$$\mathrm{STD}_j^A = \mathrm{std}\left(S_{j,1}^A, S_{j,2}^A, \ldots, S_{j,m_j}^A\right). \tag{3}$$

This measure captures inconsistency or heterogeneity in management responses across different topics raised during the call.

Q&A sentiment-difference dispersion. We define disagreement in the question–answer gap as

$$\mathrm{STD}_j^D = \mathrm{std}\left(S_{j,1}^D, S_{j,2}^D, \ldots, S_{j,m_j}^D\right). \tag{4}$$

This measure captures heterogeneity in how management responds relative to analysts' sentiment. High values of STD_j^D indicate that the answer–question mismatch varies substantially across the Q&A section.

3.3 Economic Interpretation

The three transcript-level measures capture related but distinct dimensions of disagreement.

First, STD_j^Q proxies for disagreement among analysts. If analysts ask questions with very different sentiments, they may hold heterogeneous views about the firm's fundamentals, risks, or future performance.

Second, STD_j^A proxies for inconsistency in management communication. Managers may respond confidently to some issues but more defensively or ambiguously to others, generating high dispersion in answer sentiment.

Third, STD_j^D captures the extent to which the answer–question sentiment gap varies across topics. This is especially relevant when management does not respond uniformly to analyst concerns, which may reflect information asymmetry, selective disclosure, or communication frictions.

We hypothesize that these dispersion measures proxy for information uncertainty and disagreement within earnings discussion of companies, and is a strong indicator implying that the management is speaking their minds and being honest about their operations, increasing trust and future capital flow for the companies. We further argue that this insight can be converted into a robust trading strategy.

3.4 Trading Signals

For each weekly rebalancing date t, we compute one or more of the following signals:

$$\mathrm{STD}^{Q}_{j(t)}, \quad \mathrm{STD}^{A}_{j(t)}, \quad \mathrm{STD}^{D}_{j(t)}. \tag{5}$$

Single-signal strategies. For each dispersion signal separately, we form a long portfolio containing firms in the top bins of the cross-sectional distribution. Let $\mathcal{P}^A_t$, $\mathcal{P}^Q_t$, and $\mathcal{P}^D_t$ denote the long portfolios formed from STD^A, STD^Q, and STD^D, respectively.

Intersection strategy. We define the intersection portfolio as the set of firms that simultaneously rank in the top bins across multiple disagreement measures. For example, for answer and difference dispersion,

$$\mathcal{P}^{\cap}_t = \mathcal{P}^A_t \cap \mathcal{P}^D_t. \tag{6}$$

The intersection strategy sharpens the signal by selecting firms with consistently high disagreement across dimensions.

Union strategy. We also define the union portfolio as

$$\mathcal{P}^{\cup}_t = \mathcal{P}^A_t \cup \mathcal{P}^D_t. \tag{7}$$

This broader strategy includes firms flagged by at least one dimension of disagreement, but it is generally noisier.

3.5 Trading Strategy

3.5.1 Portfolio Formation

At each weekly rebalancing date t, we construct portfolios based on the most recently available transcript-level signals. To avoid look-ahead bias, we use only transcripts released prior to portfolio formation (prior week).

Each week, we sort the available transcripts into equal-width bins. We then select the tickers from the top two bins and assign them equal weights, as defined in Equation 8. Finally, we take a long position in the selected portfolio and hold it for one week.

Portfolio returns. Let $r_{i,t+1}$ denote the return of stock i over the subsequent holding period. The equal-weighted portfolio return is

$$R_{t+1}^{\mathcal{P}} = \frac{1}{|\mathcal{P}_t|} \sum_{i \in \mathcal{P}_t} r_{i,t+1}. \tag{8}$$

We accumulate these returns over time and compare performance across strategies and against a market benchmark.

3.5.2 Hyperparameters

In order to augment our strategy, we observe whether changing the `number_of_bins` and `top_bins` jointly increases the historical performance of our strategy.

`number_of_bins` determines how many groups we split stocks into before selecting the top portfolios. With fewer bins, the selected portfolio is broader and more diversified, but the signal may be weaker. With more bins, the selected portfolio is narrower and more concentrated, which may improve the signal but also makes the strategy more sensitive to noise. For this reason, performance can change a lot across different values of `number_of_bins`.

`top_bins` determines how many of the highest-ranked groups are included in the portfolio after the stocks have been split into bins. With fewer top bins, the strategy is more selective and focuses only on the strongest signals, which can improve performance but also makes the portfolio more concentrated. With more top bins, the portfolio includes a wider set of stocks, which increases diversification but may dilute the strength of the signal. For this reason, performance can change meaningfully across different values of `top_bins`.

4 Empirical Results

4.1 Main Findings

Our empirical analysis shows that disagreement measures derived from earnings call Q&A sections contain substantial predictive information for future returns.

Specifically, we explore the performance of **single-signal strategies** based on (STD_j^Q, STD_j^A, STD_j^D), **intersection strategy** $\mathrm{STD}_j^{A\cap D}$, and the **union strategy** $\mathrm{STD}_j^{A\cup D}$ which are based on STD_j^A, STD_j^D. When firms are ranked w.r.t. the dispersion of answer sentiment, question sentiment, or the answer–question sentiment gap, the long portfolios formed from the top two bins earn economically meaningful returns. Among these signals, the intersection strategy performs best 3. Firms that simultaneously exhibit high answer-sentiment dispersion and high sentiment-difference dispersion significantly outperform both the market benchmark and broader union portfolios.

The cumulative return plot 3 indicates that the intersection portfolio delivers the highest return path over the sample period. Its Sharpe ratio is substantially above that of the standard S&P 500 benchmark, suggesting that combining multiple disagreement dimensions improves the quality of the signal.

4.2 Backtest Results

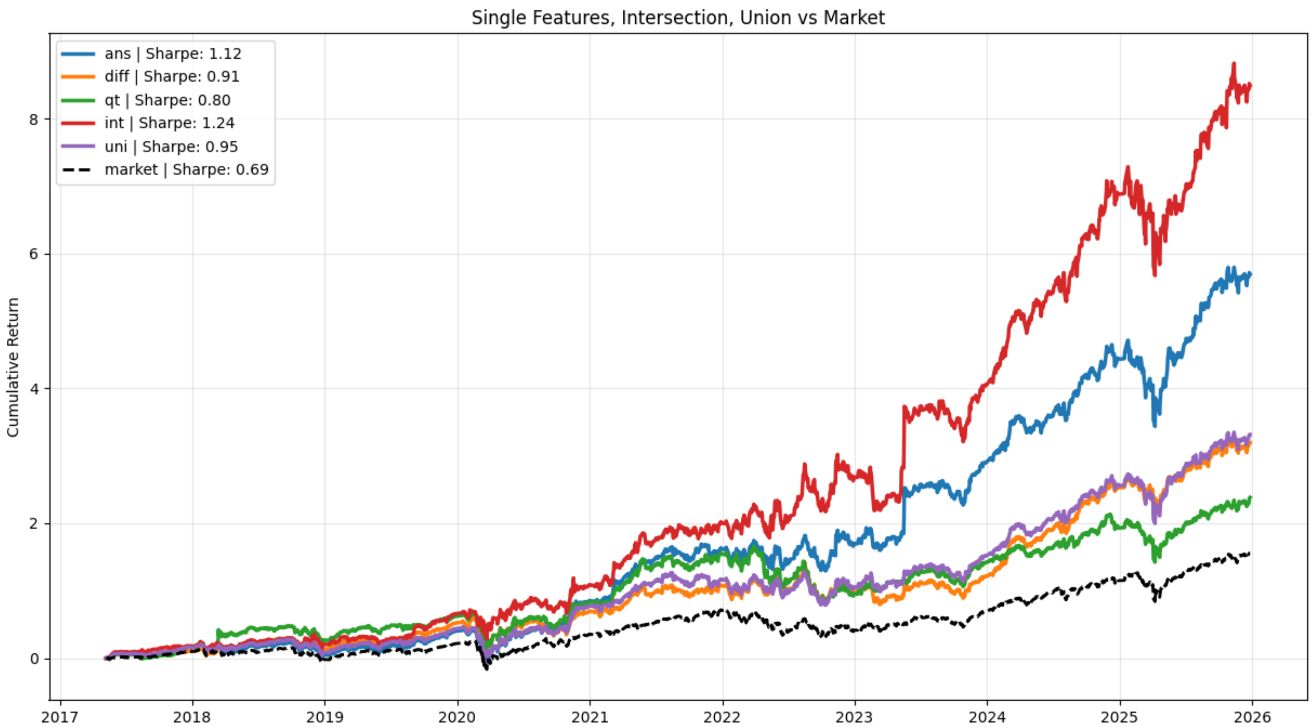


Figure 3: Cumulative returns of Intersection portfolio of dispersion in Answers and Q&A sentiment difference

Without tweaking the mentioned parameters, the results show that the intersection strategy has a historical Sharpe Ratio of 1.24, beating the union strategy with 0.95 and the S&P500 benchmark with 0.69.

4.3 Why Does the Intersection Strategy work better?

The outperformance of the intersection strategy is consistent with the idea that multiple disagreement dimensions reinforce one another. A firm with high STD_j^A alone may simply have heterogeneous management answers across topics. A firm with high STD_j^D alone may reflect isolated mismatches between analysts and managers. However, when both signals are simultaneously high, the transcript likely reflects broader communication frictions, substantial uncertainty, or heterogeneous beliefs. This stricter selection rule reduces noise and increases signal precision.

Table 2: Performance metrics by strategy, for different `bin_n` and `top_bins` values

bin_n	top_bins	Strategy	Sharpe	Sortino	MaxDD	Calmar
12	2	ans	**1.27**	1.79	-34.04%	0.93
		qt	1.02	1.41	-35.29%	0.66
		diff	1.10	1.47	-27.33%	0.93
		uni	1.08	1.41	-33.76%	0.70
		int	1.22	1.73	-24.98%	1.26
		market	0.69	0.82	-33.93%	0.35
12	3	ans	1.12	1.51	-34.94%	0.74
		qt	0.80	1.04	-35.38%	0.45
		diff	0.91	1.18	-32.60%	0.58
		uni	0.95	1.21	-35.14%	0.55
		int	**1.24**	1.71	-25.17%	1.24
		market	0.69	0.82	-33.93%	0.35

In the combination (`number_of_bins` = 12 and `top_bins` = 2) 4, we observe that the answer dispersion-based strategy overperforms the intersection strategy (Sharpe Ratio 1.27 vs. 1.24). However, we also observe that the calmar ratio for the answer-based strategy is lower (0.93 vs. 1.24) due to the increase in maximum drawdown percentage (-%34.04 vs. -%25.17). Therefore, in order to minimize the exposed risk of the strategy, we conclude that the intersection strategy given the, `number_of_bins` = 12 and `top_bins` = 3, combination is more robust and going to be implemented in live trading.

5 Conclusion

This paper develops a novel text-based disagreement framework using earnings call Q&A sections. By measuring the dispersion of sentiment across analyst questions, management answers, and their differences, we obtain transcript-level proxies for disagreement and information uncertainty. These measures have strong economic content: firms with higher disagreement subsequently earn higher returns, and the most powerful strategy is obtained by intersecting multiple disagreement signals.

Our findings suggest that the cross-sectional heterogeneity of communication within earnings calls matters for asset pricing. More broadly, the results show that the information contained in earnings calls extends beyond average tone and includes disagreement structure, which can be systematically measured and exploited in portfolio construction.

Future research may further examine the interaction between disagreement, forward guidance, analyst coverage, methods to control for firm-specific balance-sheet metrics, and market conditions (e.g., by analyzing macro variables). The next idea we aim to explore is the effect of the tonality dispersion on the implied volatility figures of the companies. This strategy can be implemented via an options strategy which strictly takes a long position in the implied volatility.

References

Al Nuaimi, K., Marti, G., Marchal, A., and Henschel, A. (2025). Detecting evasive answers in financial Q&A. *Proceedings of FinNLP (EMNLP Workshop).*

Anderson, P., Janardhanan, M. V., He, J., Cheng, W., and Flanagan, C. (2024). Greenback bears and fiscal hawks: Finance is a jungle and text embeddings must adapt. *arXiv preprint arXiv:2411.07142.*

Gutierrez-Fandino, A., Noguer i Alonso, M., Kolm, P., and Armengol-Estapé, J. (2021). FinEAS: Financial embedding analysis of sentiment. *arXiv preprint arXiv:2111.00526.*

Li, S. Z., and Luan, Z. (2025). News-based investor disagreement and stock returns. *Review of Accounting Studies.*

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance.*

Appendix

A Backtest Results

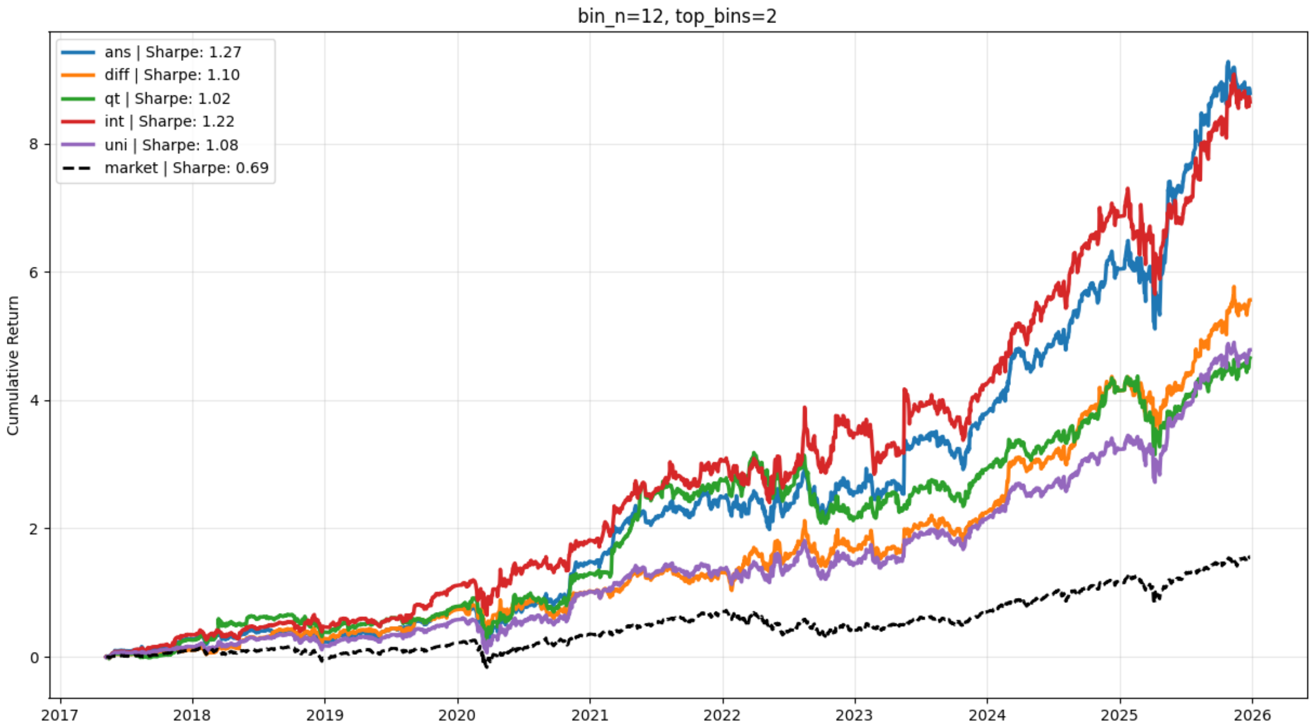


Figure 4: Backtest result for `number_of_bins` = 12 and `top_bins` = 2

Above, only the 5 strategies are shown given the distinct combination of `number_of_bins` = 12 and `top_bins` = 2. In this case, we still observe that the intersection strategy overperforms the others.

B Alternative Textual Features

In addition to our baseline sentiment dispersion measures, we explore several alternative text-based features to capture interactions between analysts and management during earnings call Q&A sessions.

B.1 Embedding-Based Similarity

We consider semantic similarity between analyst questions and management answers using pre-trained sentence embeddings. Specifically, we use the **Bidirectional Attention Model (BAM)** (Anderson et al., 2024), which produces domain-adapted embeddings suitable for financial text.

The model maps each text segment into a high-dimensional vector representation, and similarity is computed using cosine similarity:

$$\mathrm{Sim}_{j,\ell} = \frac{\langle e(Q_{j,\ell}), e(A_{j,\ell}) \rangle}{\|e(Q_{j,\ell})\| \cdot \|e(A_{j,\ell})\|} \tag{9}$$

where $e(\cdot)$ denotes the embedding function. This measure captures the semantic alignment between questions and answers, and has been widely used to detect response relevance and potential evasion in financial Q&A settings (Al Nuaimi et al., 2025).

B.2 Forward-Looking Statements

For each answer block (j, ℓ), we define forward-looking intensity as

$$\mathrm{FLS}_{j,\ell} = 0.5\, p_{j,\ell}^{\text{non-spec}} + p_{j,\ell}^{\text{spec}}, \tag{10}$$

where $p_{j,\ell}^{\text{non-spec}}$ and $p_{j,\ell}^{\text{spec}}$ are the predicted probabilities of non-specific and specific forward-looking content.

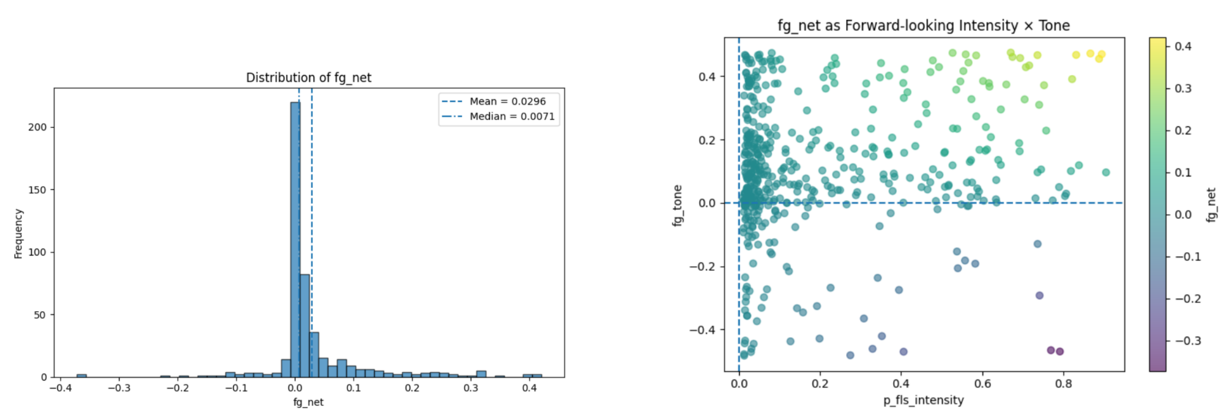
Sentiment tone is measured as

$$\mathrm{Tone}_{j,\ell} = \frac{p_{j,\ell}^{+} - p_{j,\ell}^{-}}{p_{j,\ell}^{+} + p_{j,\ell}^{-} + 1}, \tag{11}$$

where $p_{j,\ell}^{+}$ and $p_{j,\ell}^{-}$ denote the positive and negative sentiment probabilities from FinBERT.

We then construct the forward-guidance measure as

$$\mathrm{FG}_{j,\ell} = \mathrm{FLS}_{j,\ell} \times \mathrm{Tone}_{j,\ell}. \tag{12}$$



(a) Distribution of $\mathrm{FG}_{j,\ell}$

(b) $\mathrm{FG}_{j,\ell}$ against forward-looking intensity and tone

Figure 5: Distribution and structure of the forward-guidance measure.

Figure 5 shows that $\mathrm{FG}_{j,\ell}$ is concentrated around zero, while still exhibiting non-trivial positive and negative tails. The scatter plot further suggests that dispersion increases with forward-looking intensity, whereas observations with weak forward-looking content remain close to zero.

As a possible extension, the measure can be aggregated to the firm-event level and used to form portfolio-sorting signals. One may also restrict attention to blocks with sufficiently high forward-looking intensity to reduce noise.

B.3 Distributional Distance Measures

To capture changes in the distribution of textual features across earnings calls, we explore several distributional distance measures. We aim to quantify shifts in the entire distribution of Q&A-level signals.

For each transcript j at time t, let $\mathcal{F}_{j,t}^{(k)}$ denote the empirical distribution of feature k (e.g., sentiment, cosine similarity, or forward guidance scores) across all Q&A pairs within the transcript. We consider measuring the change between consecutive transcripts using distance metrics of the form:

$$D_{j,t}^{(k)} = \mathrm{Dist}(\mathcal{F}_{j,t}^{(k)}, \mathcal{F}_{j,t-1}^{(k)}), \tag{13}$$

where $\mathrm{Dist}(\cdot, \cdot)$ denotes a distributional distance.

We experiment with two commonly used metrics. First, the Wasserstein distance measures the minimal cost of transporting one distribution into another and captures differences in both location and shape. Second, Maximum Mean Discrepancy (MMD) provides a kernel-based measure of distance between distributions in a reproducing kernel Hilbert space.

However, in practice, we find that these distributional distance measures exhibit relatively low signal-to-noise ratios. This is likely due to the limited number of Q&A observations within each transcript and the resulting estimation noise in empirical distributions. Consequently, these measures do not provide strong predictive performance in our portfolio tests and are not included in the main specification.