EPFL — Project Summary

Trading the Tonality Dispersion in Earnings Call Q&A Sections

Mert Ülgüner, Linze Li

April 2026

Motivation

Earnings calls are one of the most important communication channels between firms and the market, but the most informative part is often not the prepared remarks. The Q&A section captures direct interaction between analysts and management, and that interaction reveals heterogeneity in concern, emphasis, confidence, and tone. Rather than focusing only on average sentiment, this project studies whether the *dispersion* of tone within a call contains tradable information.

The core intuition is that disagreement in Q&A language can proxy for uncertainty, communication frictions, or heterogeneous beliefs about the firm. When analysts frame issues very differently, or when management responds with uneven tone across topics, the transcript may reveal more than a single average sentiment score can capture. This is exactly where NLP becomes useful: it allows us to turn thousands of question–answer exchanges into consistent, scalable measurements of tone and disagreement.

Methodology

Data. The sample consists of 16,214 earnings call transcripts. Each transcript is segmented into Q&A pairs, where pair ℓ in transcript j contains one analyst question and the corresponding management answer. The transcripts are extracted from Refinitiv Workspace after filtering on the S&P CBOE index, with dynamically updated constituents to reduce survivorship bias. Because earnings calls arrive on a quarterly schedule, the signal naturally refreshes around earnings seasons.

Sentiment score generation. For each question and answer, sentiment is generated with FinEAS, a financial-domain language model fine-tuned on financial text and equipped with a regression-style output layer. This is important because it supports continuous sentiment measurement rather than coarse discrete labels. In our setup, the model output is represented on a $[-1, 1]$ scale so that we can compare fine-grained tone variation across Q&A interactions.

For transcript j and pair ℓ, we define

$$S^Q_{j,\ell} = f_{\text{sent}}(Q_{j,\ell}), \qquad S^A_{j,\ell} = f_{\text{sent}}(A_{j,\ell}),$$

and the tone difference between the answer and the question as

$$S^D_{j,\ell} = S^A_{j,\ell} - S^Q_{j,\ell}.$$

Transcript-level signals. We then aggregate pair-level scores into transcript-level dispersion measures:

$$\begin{aligned}
\text{STD}^Q_j &= \text{std}\Big(S^Q_{j,1}, S^Q_{j,2}, \ldots, S^Q_{j,m_j}\Big), \\
\text{STD}^A_j &= \text{std}\Big(S^A_{j,1}, S^A_{j,2}, \ldots, S^A_{j,m_j}\Big), \\
\text{STD}^D_j &= \text{std}\Big(S^D_{j,1}, S^D_{j,2}, \ldots, S^D_{j,m_j}\Big).
\end{aligned}$$

These three measures capture different aspects of disagreement:

- STD_j^Q measures dispersion across analyst questions.
- STD_j^A measures dispersion across management answers.
- STD_j^D measures variation in the answer–question tone gap.

Trading signals and portfolio construction. At each weekly rebalancing date t, firms are ranked cross-sectionally using the most recently available transcript signals, using only information released before portfolio formation to avoid look-ahead bias. For a given signal, stocks are sorted into equal-width bins and the equal-weighted portfolio is formed from the selected top bins. Let P_t^Q, P_t^A, and P_t^D denote the long portfolios formed from the three disagreement measures.

We also define combined strategies:

$$P_t^{\cap} = P_t^A \cap P_t^D, \qquad P_t^{\cup} = P_t^A \cup P_t^D.$$

Portfolio returns are equal-weighted:

$$R_{t+1}^P = \frac{1}{|P_t|} \sum_{i \in P_t} r_{i,t+1}.$$

Hyperparameters. Two design choices matter:

- **Number of bins.** More bins make the signal more selective and concentrated, but also more sensitive to noise. Fewer bins broaden diversification, but can dilute signal strength.
- **Selected number of top bins.** Choosing fewer top bins focuses on the strongest disagreement signals, while choosing more top bins increases breadth at the cost of lower purity.

These hyperparameters are introduced to control the trade-off between concentration and diversification, and between signal sharpness and robustness.

Results

The main result is that tonality dispersion in the Q&A section carries economically meaningful predictive information for future returns. Long-only portfolios formed from high-dispersion firms outperform the broad market, and combining multiple disagreement dimensions improves the quality of the signal.

The strongest baseline result comes from the intersection strategy based on answer dispersion and answer–question tone-difference dispersion, which delivers a historical Sharpe ratio of **1.24**. This is meaningfully stronger than both the union strategy and the market benchmark. An alternative specification based only on answer dispersion reaches a slightly higher Sharpe ratio of **1.27**, but that improvement comes with a larger maximum drawdown. For that reason, the intersection-based implementation is treated as the more robust live-trading candidate, as it gives up a small amount of upside in exchange for a clearly better risk profile.

Extension

First, we plan to explore additional textual scores, especially **forward-guidance measures** and **cosine similarity** between analyst questions and management answers. These features may capture whether management is forward-looking, evasive, or semantically aligned with analyst concerns.

Second, we aim to develop **macro-adjustments** to the raw sentiment scores. The goal is to distinguish firm-specific disagreement from market-wide tone regimes, so that sentiment is interpreted relative to the broader macro backdrop rather than in isolation.

Third, we want to test the **immediate effect of tonality dispersion on implied volatility**. If high disagreement in the Q&A section is quickly reflected in option prices, this would open the door to volatility-focused implementations in addition to the current equity strategy.