Transforming Financial Text into Return Signals: A Comparative Study of News Sentiment and SEC Filing Features

Ross E Cunningham, Ethan Davis

December 20, 2025

Abstract

This study investigates whether large language models (LLMs) can extract economically meaningful signals from unstructured financial text and how such signals can be integrated with traditional asset pricing frameworks to predict equity returns across multiple horizons. We examine two complementary information channels: corporate regulatory disclosures and firm-specific financial news. For regulatory filings, we collect SEC 10-K, 10-Q, and 8-K documents for S&P500 firms and process their narrative content using an LLM to generate structured linguistic features capturing sentiment, tone, and perceived risk. These features are used within a supervised machine-learning framework to predict five-day forward returns and to construct an event-driven long-short trading strategy. For financial news, we score article headlines using FinBERT, aggregate sentiment to the daily ticker level, and construct standardized sentiment and surprise measures that inform short-horizon trading rules and portfolio filters.

Both pipelines are evaluated under consistent horizon definitions, time-based train test splits to prevent information leakage, and performance metrics that account for turnover and implementation realism. We further benchmark the text-based strategies against the Fama-French Five Factor (FF5) model to assess their relationship to established sources of systematic risk. The results indicate the at while LLM-based strategies do not outperform FF5 on a standalone basis, filing-based signals generate statistically meaningful returns that are largely orthogonal to traditional factor exposures, suggesting that corporate disclosures contain incremental information beyond price-based models. In contrast, news sentiment signals exhibit strong but short-lived predictive power, with hihg-turnover long-short strategies proving fragile out of sample and lower-turnover implementations demonstrating greater robustness.

Taken together, the findings highlight that the economic value of LLM-dervied text signals depend critically on information source, horizon alignment, and execution design. By combining factor-based long-term return expectations with medium-horizon disclosure-driven signals and short-horizon new sentiment, the study outlines a unified framework in which unstructured text serves as a complementary input to a traditional quantitative asset pricing models rather than a replacement for them.

Contents

1 Introduction **3**
- 1.1 Unstructured Financial Text as an Information Channel . . . 3
- 1.2 Information Timing and Horizon Heterogeneity . . . 3
- 1.3 Text-Base Signals and Factor Models . . . 4
- 1.4 Contributions and Overview . . . 4

2 Related Literature **4**

3 Data **5**
- 3.1 Financial News (FINSPID & Finnhub) . . . 5
  - 3.1.1 FINSPID . . . 5
  - 3.1.2 Finnhub . . . 6
  - 3.1.3 Sample Window . . . 6
  - 3.1.4 Data Preprocessing . . . 6
- 3.2 News-Based Sentiment Pipeline(Short Horizon) . . . 6
- 3.3 SEC Filings . . . 8
- 3.4 Equity Returns . . . 8

4 Methodology **8**
  - 4.0.1 FINSPID Feature Construction and Exploratory Validation . . . 8
  - 4.0.2 Finnhub Integration . . . 12
- 4.1 LLM-Based Text Processing of SEC Filings . . . 13
- 4.2 Feature Engineering and Machine Learning Framework . . . 14
  - 4.2.1 Feature Construction . . . 14
  - 4.2.2 Predictive Model . . . 14
  - 4.2.3 Training and Evaluation . . . 14
- 4.3 Trading Strategy Construction . . . 14
- 4.4 Factor-Based Benchmark and Residualization (FF5) . . . 14

5 Empirical Results **16**
- 5.1 New Sentiment Strategy Results . . . 16
  - 5.1.1 Initial Strategy Versus Equal Weight Hold . . . 16
  - 5.1.2 Parameter Sweep Methodology and Selection Rule . . . 18
  - 5.1.3 Sweep Results and Top Parameter Settings . . . 18
  - 5.1.4 Interpreting the Cross Period Gap Using Coverage Diagnostics . . . 19
- 5.2 Filing-Based LLM Strategy . . . 19

6 Comparison with the Fama-French Five Factor Model **20**

7 Comparison of Signal Strategy Against S&P 500 **22**
- 7.1 Motivation . . . 22
- 7.2 Results . . . 22

8 An Extension: Text-Based Prediction of Factor-Adjusted Returns **24**
- 8.1 Methodological Extension: Integrating LLM-Based Signals with the Fama-French Framework . . . 24
- 8.2 Two-Stage Modeling Framework . . . 25
  - 8.2.1 Stage 1: Factor-Based Return Decomposition . . . 25
  - 8.2.2 Stage 2: Text-Based Prediction of Residual Returns . . . 25
- 8.3 Out-of-Sample Evaluation and Portfolio Construction . . . 25
- 8.4 Interpretation and Motivation for the Extended Approach . . . 26

9 Bridging the Two Pipelines: A Unified Text-Based Return Framework **27**

10 Limitations and Future Research **28**
10.1 News Sentiment Analysis Model 28
10.1.1 Success and Financial Intuition 28
10.1.2 Practical Constraints and Limitations 28
10.1.3 Areas for Improvement 28
10.2 SEC Filings LLM Model 29

11 An Integrated Multi-Horizon Strategy Framework **29**

12 Conclusion **30**

1 Introduction

1.1 Unstructured Financial Text as an Information Channel

Financial markets are shaped by a continuous flow of information, much of which is conveyed through structured text. Corporate disclosures and financial news represent two of the most important channels through which firms communicate information to investors, yet much of this information remains difficult to incorporate into systematic asset pricing models. Traditional quantitative approaches to return prediction have therefore focused primarily on price-based factors or structured accounting variables, leaving a large portion of qualitative information underutilized.

Corporate disclosures play a central role in the functioning of financial markets, acting as a primary channel through which firms communicate information to investor. Regulatory filings such as 10-Ks, 10-Qs, and 8-Ks provide detailed discussions of firm performance, strategic priorities, risks, and forward-looking expectations. These disclosures are legally constrained and subject to regulatory oversight, rendering them a particularly credible source of firm-level information. Despite their importance, the unstructured and narrative nature of these documents has historically limited their application in systematic quantitative asset pricing models.

Traditional approaches to modeling stock returns have largely relied on price-based factors or accounting variables, with relatively little emphasis placed on qualitative corporate language. This limitation arises primarily from methodological challenges: textual disclosures are long, heterogeneous, and difficult to convert into numerical form. As a result, much of the informational content embedded in corporate communication remains underutilized within quantitative frameworks.

Recent advances in large language models (LLMs) offer a potential solution to this constraint. Unlike earlier natural language processing techniques, LLMs are capable of interpreting semantic context, nuance, and syntactic structure across long passages of text. This capability allows qualitative information, such as managerial tone, uncertainty, and risk disclosure, to be transformed into structured features suitable for machine-learning models. Consequently, LLMs provide a novel mechanism through which textual disclosures may be systematically incorporated into return prediction frameworks.

1.2 Information Timing and Horizon Heterogeneity

While regulatory filings represent a rich and credible source of firm-specific information, they are released infrequently and inherently event-driven. In contrast, financial news articles are produced continuously and disseminated rapidly, often reflecting real-time developments, market narratives, and shifts in investor sentiment. This distinction suggests that different types of financial text may influence asset prices over different horizons.

News based sentiment signals are typically fast moving, noisy, and short-lived, as markets react quickly to headline information and narrative framing. In contrast, filing-based disclosures convey more structured and information-dense content related to firm fundamentals, risk exposure, and strategic direction, which may take longer for markets to fully process. As a result, the predictive power of textual signals is inherently horizon-specific, and evaluating all forms of financial text using a single modeling or trading framework may obscure important differences in how information is incorporated into prices. Recognizing this heterogeneity in information timing is essential for both interpretation and implementation. Short-horizon strategies based on news sentiment may require frequent rebalancing and are sensitive to execution costs, while event-driven strategies based on regulatory filings operate

at lower frequency and may exhibit greater stability. These differences motivate a framework that explicitly aligns each text source with the horizon at which it is most informative.

1.3 Text-Base Signals and Factor Models

At the same time, a substantial body of empirical evidence demonstrates that a large portion of long-run equity returns is explained by systematic risk premia, as captured by factor models such as the Fama-French Five Factor (FF5) framework. These factors reflect persistent sources of expected returns related to market risk, size, value, profitability, and investment, rather than the firm-specific information shocks.

Consequently, any return predictability derived from textual information should be evaluated not only on a standalone basis, but also in relation to established factor exposures. Without such benchmarking, it is difficult to determine whether text-based strategies capture genuinely new information or simply proxy for known risk premia. Incorporating a factor-based perspective therefore provides a disciplined benchmark against which the incremental value of LLM-derived signals can be assessed.

1.4 Contributions and Overview

Motivated by these considerations, this study investigates whether LLM-derived features extracted from unstructured financial text can generate economically meaningful return signals, and how much signals operate across different horizons. We examine two complementary information channels: corporate regulatory filings and firm-specific financial news. For filings, we extract narrative content from SEC 10-K, 10-Q, and 8-K documents and use an LLM to generate structured linguistic features capturing sentiment, tone, and perceived risk. These features are used within a supervised machine-learning framework to predict short-horizon equity returns and to construct event-driven trading strategies. For financial news, we apply a sentiment model to article headlines, aggregate sentiment at the daily ticker level, and construct standardized sentiment and surprise measures that inform short-horizon trading rules and portfolio filters.

Both pipelines are evaluated under consistent horizon definitions, time based train-test splits designed to prevent information leakage, and performance metrics that account for turnover and execution realism. In addition, we benchmark whether their performance reflects exposure to known risk premia or incremental information content.

The contributions of this paper are threefold. First, we provide a direct comparison of two LLM-based text pipelines applied to distinct financial information sources. Second, we document that the predictive power of textual signals is horizon specific: news sentiment is most informative at very short horizons, while regulatory filings convey information that unfolds more gradually. Third we propose an integrated multi-layer framework in which long-run expected returns are anchored by factor models, medium-horizon alpha is derived from corporate disclosures, and short-horizon positioning is informed by news sentiment. Together, these results suggest that unstructured text is best viewed as a complementary information channel that enhances, rather than replaces, traditional quantitative asset pricing models.

2 Related Literature

A substantial body of research has documented the influence of textual information on financial markets. Early studies, such as Tetlock (2007), demonstrate that negative language in media content predicts subsequent market returns and volatility. Subsequent research extends these insights to a variety of financial text sources including corporate disclosures, earnings call transcripts, and analyst reports, highlighting the role of sentiment, tone, and uncertainty in shaping investor expectations.

However, much of this literature relies on dictionary-based or frequency-driven sentiment measures. While intuitive, these approaches struggle to capture context-dependent meaning and often fail to distinguish between economically relevant and boilerplate language. This limitation is particularly pronounced in regulatory filings, where language is often formal, repetitive, and subject to legal constraints that dilute the effectiveness of simple word-count-based sentiment measures. As a result, traditional text-based methods may understate the informational content of corporate disclosures.

Recent advances in large language models represent a structural shift in textual analysis. By leveraging deep neural architectures trained on large corpora of natural language, LLMs are able to infer semantic meaning, contextual nuance, and narrative structure in ways that surpass traditional sentiment dictionaries. A growing empirical literature has begun to explore the application of LLMs and transformer-based models in finance, particularly in the analysis of earnings call transcripts and financial news, where such models have been shown to improve sentiment measurement and predictive performance.

Empirical work applying LLMs to SEC filings remains more limited, especially within a systematic trading or asset pricing framework. Existing studies often focus on single text sources or evaluate predictive performance without explicitly benchmarking against established factor models. As a result, it remains an open question whether LLM-derived signals extracted from regulatory disclosures capture information that is distinct from traditional sources of expected returns, or whether they primarily reflect exposure to known risk premia.

Similarly, the literature on news-based sentiment has documented short-horizon return predictability, but has also highlighted concerns related to signal instability, high turnover, and sensitivity to market regimes. These findings underscore the importance of aligning textual signals with appropriate investment horizons and evaluating them under realistic implementation constraints.

By situating both filing-based and news-based LLM signals within a unified framework and explicitly comparing them to factor-based benchmarks, this study contributes to the emerging literature on text-based asset pricing by clarifying when, how, and over what horizons unstructured textual information adds incremental value.

3 Data

The empirical analysis in this study draws on multiple data sources that correspond to distinct channels of information flow in financial markets. To reflect the multi-horizon framework outlined in the Introduction, we combine corporate regulatory disclosures, firm-specific financial news, and standard market and factor data within a unified empirical setting. Each data source is selected to capture a different dimension of information relevant for return prediction, while maintaining careful alignment in timing, coverage, and frequency to ensure comparability across models and strategies.

Regulatory filings provide structured, legally constrained disclosures that convey firm-level information at discrete points in time. These documents serve as the foundation for the event-driven LLM-based filing analysis. In contrast, financial news articles represent a high-frequency and continuously updating information stream that captures short-term sentiment, narrative framing, and market attention. Market return data and Fama–French factor series are used both to construct trading strategy returns and to benchmark text-based signals against established sources of systematic risk.

Importantly, all datasets are processed and aligned to prevent look-ahead bias and to ensure that textual features are matched only with returns that occur after the corresponding information becomes publicly available. This design allows for a consistent evaluation of filing-based and news-based signals under common horizon definitions, while preserving the distinct informational role of each data source. The following subsections describe each dataset in detail and outline the preprocessing steps used to construct the final analysis panels.

3.1 Financial News (FINSPID & Finnhub)

3.1.1 FINSPID

Our primary historical news source is FINSPID, commonly referenced as FINSPID, the Financial News and Stock Price Integration Dataset. The key motivation is scale and alignment. The dataset was designed to integrate financial news with equity price records under a shared time series framework, which makes it well suited for building ticker day sentiment measures and evaluating their relationship to forward returns. This scale is valuable for two reasons. First, it reduces the likelihood that results are driven by idiosyncratic events in a narrow window, which is especially important when testing short horizon signals that can be noisy. Second, it supports cross sectional portfolio construction, where daily signals require consistent coverage across many tickers. A practical constraint is size. The raw FINSPID files used in our workflow are approximately 22.6 GB locally, which affects storage, compute, and what can reasonably be packaged for submission. For that reason, we do not include

the full raw FINSPID dump in the submission. Instead, we provide the scripts that reproduce the transformation into the final ticker day panel and include the processed outputs used in model and strategy evaluation.

3.1.2 Finnhub

Finnhub and Yahoo Finance via yfinance out of sample extension. To evaluate robustness under an independently constructed dataset, we build a second panel for 2025 using Finnhub for company news and Yahoo Finance data accessed through the yfinance package for daily prices. Finnhub provides article level records by ticker and date range, including timestamps and short text fields such as headlines and summaries. These fields are sufficient to apply the same FinBERT sentiment scoring approach and daily aggregation logic used in the historical panel. The daily price series is constructed from adjusted close data so that returns are computed consistently across tickers. We treat this panel as deliberately independent from FINSPID. It uses different collection mechanisms and shows materially different news coverage patterns, so differences in performance help identify whether our feature choices reflect a robust relationship or a result that depends on one specific dataset construction.

3.1.3 Sample Window

We restrict attention to a fixed universe of 40 large, liquid US equities that receive consistent media coverage examples being AAPL, TSLA, MSFT, JNJ. This choice supports the core objective of the project, which is to construct daily sentiment features that appear frequently enough to support cross sectional signals, and to evaluate strategies on assets with reliable daily pricing and low missingness. A stable ticker universe also improves comparability between the FINSPID based panel and the Finnhub based panel, since both panels are evaluated on the same set of firms and differences in outcomes are less likely to be driven by universe selection.

We focus the main historical analysis on a post 2020 window. This decision is partly methodological and partly practical. Methodologically, the post 2020 period reflects more recent market structure and information environments, which is relevant for a short horizon sentiment based signal. Practically, working with the full FINSPID date range is computationally expensive, and narrowing the window makes it feasible to iterate on feature definitions, validate preprocessing, and run repeated strategy tests within typical course project constraints. The 2025 period is then used as an out of sample extension built from an independent pipeline. The goal is not to maximize performance in that short window, but to test whether the same feature definitions and strategy logic behave similarly under a different regime and a different news source.

3.1.4 Data Preprocessing

This section describes how raw article level text and raw prices are transformed into a single daily ticker panel used for modeling and strategy testing. The preprocessing choices were guided by two requirements that connect directly to the evaluation rubric. First, the pipeline must avoid time leakage, meaning features at date t must use only information available up to date t, while labels and realized returns must refer to future price movement. Second, the feature set must be interpretable and motivated. For each sentiment feature, we aim to capture a distinct mechanism by which news may affect short horizon returns, so that results can be explained rather than reported as a black box outcome.

3.2 News-Based Sentiment Pipeline(Short Horizon)

Before feature construction, we standardize tickers and dates across sources. Tickers are uppercased to prevent merge inconsistencies, and all timestamps are converted into a daily date key that matches the trading calendar. Data are sorted by ticker then date before any rolling computations or shifts. This step is essential because both the forward return label and rolling statistics depend on correct time ordering. For Finnhub news, we also deduplicate across weekly pull windows to avoid counting the same article multiple times when API windows overlap. These cleaning steps reduce measurement error in daily sentiment features and prevent artificial spikes in volume or sentiment intensity.

Each article is scored using FinBERT probabilities for positive, neutral, and negative sentiment. From these probabilities, we construct two article level primitives that carry most of the information

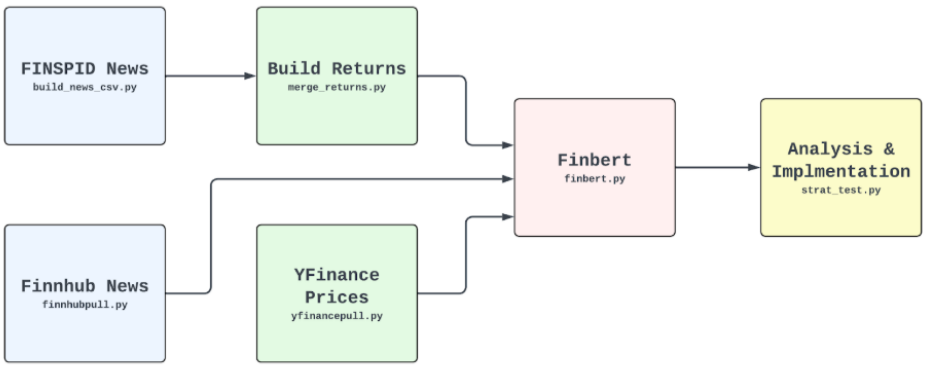

Figure 1: Financial News Python Pipeline

used downstream. Net sentiment measures directional tone and is defined as positive minus negative probability. Sentiment intensity measures strength and is defined as the absolute value of net sentiment. We prefer these continuous scores over discrete labels because they preserve confidence information, aggregate naturally across multiple articles, and map directly to trading intuition. A strongly positive net score represents favorable language, while a near zero net score often corresponds to neutral language or mixed sentiment.

We then aggregate article level measures to daily ticker features. This aggregation is a deliberate methodological choice because the number of articles per day varies substantially by firm and by event intensity. Trading and prediction are evaluated at the daily frequency, so a daily panel also simplifies time based splitting and strategy implementation. The first daily feature is mean net sentiment, which captures the overall direction of news on that day. The second is mean intensity, which captures how strongly worded the coverage is, regardless of direction. We include intensity because price reactions are often driven by extreme news rather than mild tone. The third feature is sentiment disagreement, computed as the within day standard deviation of net sentiment, which captures whether coverage is consistent or conflicting across articles. This feature is included because mixed narratives may weaken or delay market reaction and also because a mean near zero can represent either neutrality or offsetting extremes. The fourth feature is sentiment volume, the number of articles on that ticker day, which acts as a proxy for attention and information arrival. Volume is included because a single article day produces a noisier estimate than a high volume day and because large news days are often event driven.

To focus on tail events, we also construct strong positive and strong negative ratios based on a fixed threshold applied to article net sentiment. For each ticker day, we count the number of articles above the positive threshold and below the negative threshold, then divide by daily volume. The rationale is that extreme language is more likely to coincide with genuinely new information and larger price responses, and using ratios makes the feature comparable across low and high volume days. Finally, we create a short horizon surprise feature by comparing today's net sentiment to a rolling mean of recent net sentiment for the same ticker. The motivation is that markets adapt to a persistent tone, so changes relative to recent history may better capture novelty. We keep the window short to match the focus on short horizon predictability.

On the price side, we compute daily returns from adjusted close prices and construct rolling realized volatility as a simple finance baseline feature. For prediction, we define the label as the next day forward return, created by shifting returns within ticker so that the label stored on date t corresponds to the return from t to t plus one. This label definition is aligned with our evaluation design and is the primary control against time leakage. The final panel is formed by merging daily sentiment features into the daily price panel on ticker and date, keeping all price rows and treating sentiment as an additional information layer when present. This design also makes data coverage differences transparent, which is important for interpreting out of sample results. In particular, the Finnhub based panel naturally has lower sentiment coverage than FINSPID, which reduces the cross sectional breadth of daily signals and can materially affect strategy behavior even when the strategy code is unchanged.

Together, these steps produce a single daily panel where each row is a ticker day with aligned price based returns, forward return labels, and aggregated sentiment features. This standardized panel is

the input to all downstream experiments, which allows us to compare models and strategies without changing the underlying data structure.

3.3 SEC Filings

The primary dataset consists of corporate regulatory filings from the SEC's EDGAR system. The sample includes all available Form 10-K, 10-Q and 8-K filings for firms within the S&P500 index over a two-year period. These filings represent a broad cross section of mandatory corporate disclosures, ranging from comprehensive annual reports to unscheduled announcements of material events. Each filing is timestamped using its official submission date, which serves as the event date for subsequent return alignment.

Prior to analysis, filings are preprocessed to remove HTML markup, tabular data, and non-narrative sections. This preprocessing step is essential to ensure that the analysis focuses on qualitative language intended for human interpretation, including management discussion, risk disclosures, and forward-looking commentary, rather than mechanical reporting structures. By isolating narrative content, the methodology aims to capture linguistic signals that reflect managerial intent, uncertainty, and strategic emphasis.

3.4 Equity Returns

Daily equity return data are obtained from standard market data sources and aligned with filing dates at the firm level. Forward-looking returns are computed over a five-day horizon following each filing event. This horizon is chosen to balance two competing considerations. First, it is sufficiently short to capture market reactions to newly released information before it is fully incorporated into prices. Second, it mitigates the excessive noise and microstructure effects often associated with intraday or single-day return measures.

The use of an event-based forward return window is consistent with the notion that regulatory filings convey discrete information shocks, rather than continuous signals, and that markets require several trading sessions to process and respond to complex textual disclosures.

4 Methodology

We implement a reproducible pipeline that converts raw news and raw price data into a daily ticker panel used for feature engineering, prediction, and strategy backtesting. The workflow has two parallel data inputs. FINSPID provides the historical research grade news and price records used for development and exploratory validation. A second pipeline built from Finnhub news and yfinance prices provides a 2025 out of sample panel used to test whether the same feature definitions and trading rules transfer to a different data collection mechanism and market regime. Presenting the pipeline first is important for transparency because it makes clear where each transformation occurs and how intermediate outputs feed into the next step. Figure 1 summarizes the end to end workflow and highlights the intermediate outputs that make the downstream feature engineering, prediction, and backtesting steps auditable and reproducible.

In both pipelines, the same high level transformations are applied. News is mapped to ticker and date, scored with FinBERT at the article level, aggregated into daily sentiment features, and then merged onto daily prices to compute returns and forward return labels. The merged panel is the common input into our modeling and strategy functions, which allows us to keep the analysis code consistent across datasets and isolate performance differences to data coverage and regime effects rather than implementation differences

4.0.1 FINSPID Feature Construction and Exploratory Validation

We first develop and validate the sentiment feature design using the FINSPID panel, which provides a long sample and dense article coverage. This stage focuses on intermediate checks that demonstrate the pipeline is working as intended and that our modeling and strategy choices are motivated by observed patterns. We begin from FINSPID news at the article level and map each record to a ticker and a calendar date. Because the end goal is a daily tradable signal, we aggregate information to the daily frequency. A key intermediate check is whether FINSPID provides consistent coverage over time, since

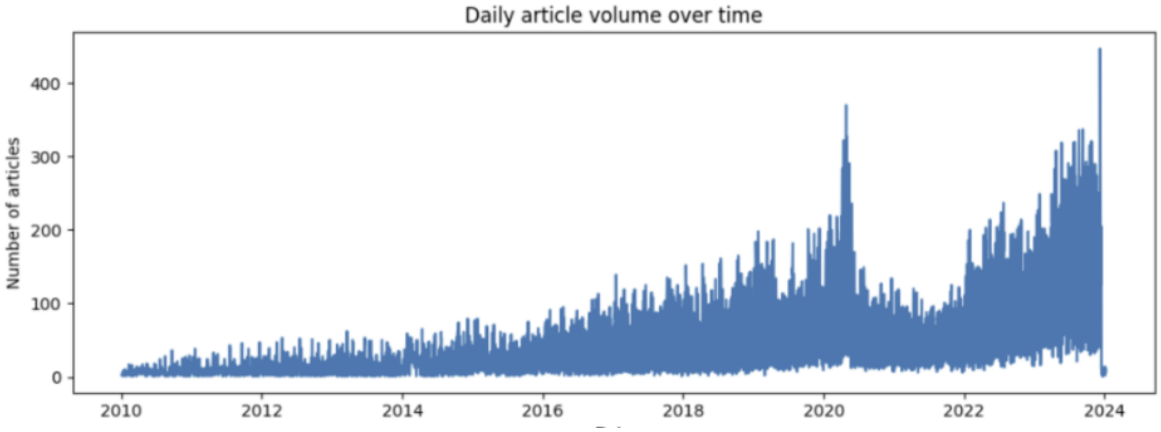

Figure 2: Daily Article Volumen Over Time (FINSPID)

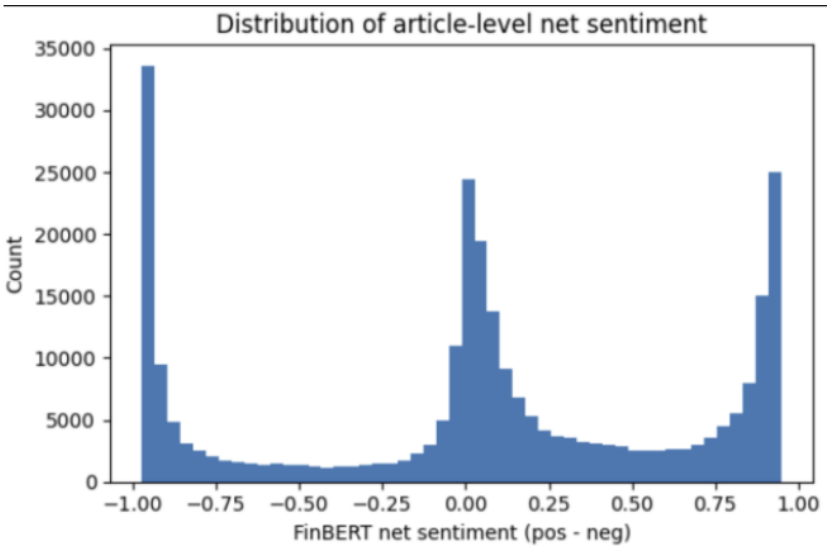

Figure 3: Article Level FinBERT Score Distributions

a daily sentiment signal requires frequent text observations. We therefore summarize the number of articles per day across the historical sample. The resulting series shows persistent coverage and clear spikes during periods of elevated information flow, supporting the decision to represent sentiment at the daily level.

Each article is then scored using FinBERT, producing probabilities for positive, neutral, and negative sentiment. We convert these probabilities into an article level net sentiment score defined as the positive probability minus the negative probability. We prefer a continuous net score over a discrete label because it preserves confidence information and aggregates naturally across multiple articles. Before constructing daily features, we validate that the FinBERT output has useful dispersion by examining the distribution of article level net sentiment. This step is an important intermediate result because a sentiment model that outputs mostly near zero values or saturates at extremes would not support meaningful ranking signals. The distribution exhibits substantial variation and clear mass in both positive and negative directions, supporting its use as a continuous input into daily aggregation. We then aggregate article level sentiment to the ticker day level. For each ticker and date, we compute daily mean net sentiment as the primary directional signal. We also compute complementary features that capture different intuitions about how news may map to short horizon returns. Daily sentiment intensity is defined as the mean absolute net sentiment and captures how strongly worded coverage is. Sentiment disagreement is the within day standard deviation of net sentiment and captures whether coverage is consistent or mixed, since a mean near zero can reflect either neutrality or offsetting ex-

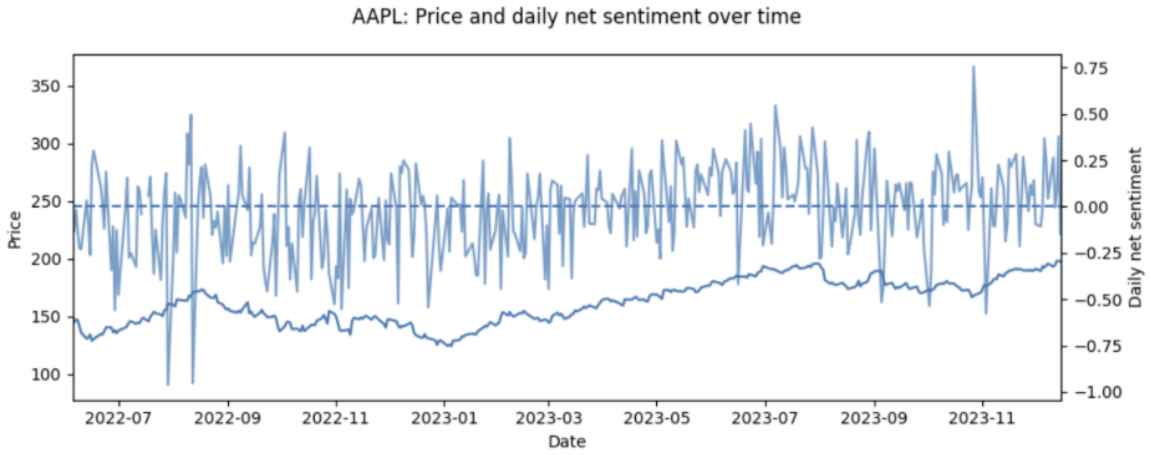

Figure 4: AAPL Price and Daily Net Sentiment Over Time(FINSPID)

tremes. Sentiment volume is the number of articles in the day and is included as a proxy for attention and as a reliability indicator for the daily mean. Finally, we compute a short horizon surprise feature defined as the deviation of today's net sentiment from a recent rolling mean for that ticker, intended to capture novelty relative to the recent narrative rather than persistent tone.

To illustrate that this aggregation step produces a coherent time series that aligns with market data, we plot an example ticker showing daily net sentiment together with the corresponding price path. The purpose is not to infer causality from one firm, but to demonstrate that the daily sentiment signal exhibits meaningful variation, merges cleanly onto prices, and can be interpreted at the same frequency as trading decisions. Next, we use a horizon analysis to motivate the holding period and label design used in prediction and strategy testing. For each horizon h, we compute the correlation between a sentiment feature observed on day t and the forward log return from t to t plus h, within each ticker, and then summarize across tickers. This intermediate result directly informs strategy construction. If the correlation decays quickly with horizon, it suggests that any information content in sentiment is short lived and that signals should be evaluated and traded at short horizons. In our results, the mean correlation is strongest at very short horizons and declines as the horizon increases, and for some features the sign changes at longer horizons. This motivates next day direction labels and short rebalance intervals rather than multi week holding periods. Because the horizon evidence indicates that any sentiment effect decays quickly, we use a simple, short horizon portfolio rule that only depends on information available at date t and is realized over the next few trading days. This motivates next day direction labels and short rebalance intervals rather than multi week holding periods. Because the horizon evidence indicates that any sentiment effect decays quickly, we use a simple, short horizon portfolio rule that only depends on information available at date t and is realized over the next few trading days. We implement a cross sectional ranking approach because it tests whether sentiment contains relative information across firms on the same day, rather than relying on a single stock's time series relationship, which can be unstable and highly idiosyncratic. We chose a quantile long short rule because it is the simplest way to test whether sentiment contains relative information across stocks on the same day, without relying on any single name's unstable time series. This structure also partially cancels broad market moves, so performance is more attributable to the sentiment ranking rather than overall index direction. Quantile based long and short selection focuses trading on the most extreme signals, where measurement error is smaller and any market reaction is more likely to be economically meaningful, while also keeping the rule transparent and easy to reproduce.

Finally, we document that the sentiment return relationship is heterogeneous across tickers. Even when the average correlation is positive at short horizons, individual firms can exhibit weaker or even negative relationships due to differences in news dynamics, investor attention, and coverage quality. We therefore compute per ticker correlations between next day returns and daily net sentiment and report the most positive and most negative cases. This diagnostic supports the decision to rely on cross sectional ranking across a universe rather than a single asset rule. At this point, the FINSPID development stage establishes the core feature definitions and the short horizon focus that we use in

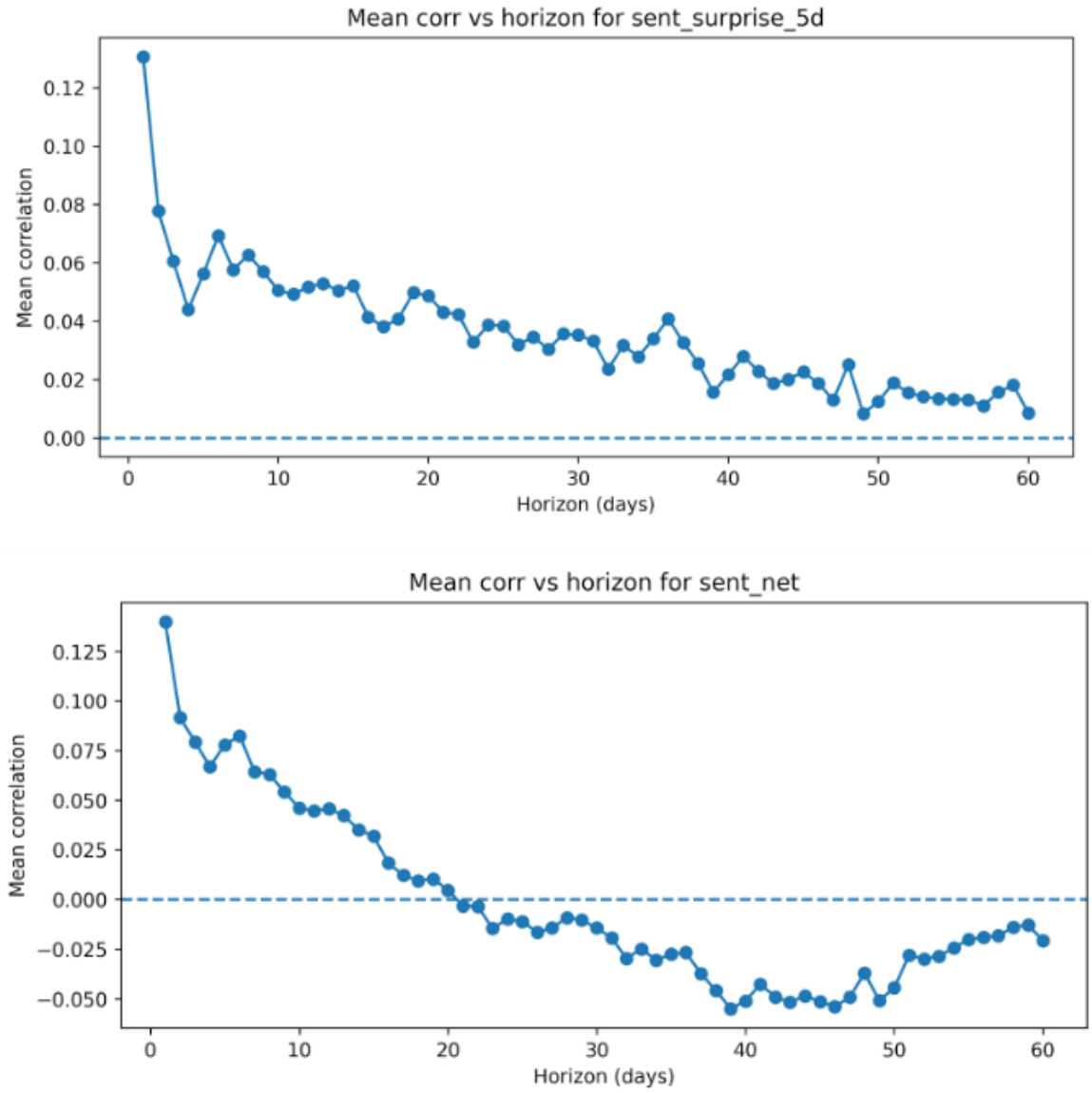

Figure 5: Mean Correlations vs Time Horizons for Sentiment Net Scores and Sentiment Surprise 5D

Ticker	Top 10 Negative Correlations	Ticker	Top 10 Positive Correlations
GE	0.007130	AVGO	0.440636
GOOG	0.007939	BAC	0.418164
KO	0.029100	PFE	0.386630
WMT	0.050198	TSLA	0.360577
CAT	0.55034	NVDA	0.357412
PEP	0.059567	JNJ	0.348873
ABBV	0.063592	SBUX	0.335488
JPM	0.077502	MSFT	0.290046
NFLX	0.083777	SCHW	0.259951
CVX	0.089893	AAPL	0.252549

Table 1: Individual Stock Based Correlation Table.

the remainder of the project. The next subsection applies the same pipeline to a 2025 panel built from Finnhub and yfinance to evaluate whether the implementation carries over and whether the same short horizon relationships appear under a different data source and regime. Having fixed the feature definitions and the short horizon strategy design using FINSPID, we then apply the same pipeline to the 2025 panel without redefining the signal, so that differences in performance can be interpreted as a robustness check rather than an artifact of changing methodology. At this stage, feature definitions and the trading rule family are treated as fixed. The 2025 panel is then constructed independently and evaluated using the same features, label definition, and strategy mechanics, so differences in outcomes are interpreted as robustness or regime effects rather than re-tuning.

4.0.2 Finnhub Integration

The FINSPID development stage establishes the feature definitions and the short horizon focus that guide the remainder of the analysis. To test whether these choices generalize beyond a single dataset construction, we then rebuild the same daily panel structure using a separate 2025 pipeline based on Finnhub company news and yfinance price data. The goal of this stage is not to re-optimize features or strategy rules for a new sample, but to hold the methodology fixed and evaluate whether the same transformations and trading logic remain coherent when the underlying news source, coverage patterns, and market regime differ.

We begin by collecting article level news for the same ticker universe using the Finnhub company news endpoint over the 2025 window. Requests are executed in weekly windows to manage rate limits and reduce the risk of missing data from overly large queries. Because weekly windows can overlap and Finnhub can return the same article in adjacent pulls, we deduplicate records using a stable key constructed from ticker, headline, and timestamp. This step is important for downstream feature integrity because duplicate articles would artificially inflate sentiment volume and could distort daily averages or tail ratios. In parallel, we collect daily prices from yfinance and standardize the output to match the FINSPID panel format, including a consistent ticker field, a daily date column, and a single price column used to compute returns.

After data collection, we apply the same FinBERT scoring procedure to the 2025 article text and construct the same article level primitives used in FINSPID, namely net sentiment and intensity. We then aggregate these scores to daily ticker features using identical definitions: daily mean net sentiment, daily mean intensity, daily disagreement as within day standard deviation, daily volume as the number of articles, strong positive and strong negative ratios based on the same threshold logic, and the short horizon surprise feature defined relative to a rolling mean. Maintaining identical feature definitions is intentional. It ensures that differences observed between FINSPID and 2025 can be interpreted as a robustness check rather than a consequence of redefining the signal.

To verify that the 2025 construction produces a clean daily time series aligned with prices, we repeat the same type of intermediate visualization used in the FINSPID section by plotting a representative ticker's daily sentiment measure alongside its price path. This figure serves as a structural validation that the article to day aggregation, time alignment, and merge logic behave as expected in an independently assembled dataset. It also provides a qualitative check that the 2025 sentiment series is not trivially constant and that there are visible periods of elevated positive or negative tone that correspond to realistic news regimes. Once the daily panel is constructed, we compute the same finance features and prediction labels used in the FINSPID analysis. Daily returns are computed from prices, and the forward one day return label is created by shifting returns within ticker so that the label stored at date t refers to the return from t to t plus one. This design matches the FINSPID setup and prevents time leakage, because sentiment features for date t are based only on news observed at date t, while the label refers to future price movement. The resulting 2025 panel therefore has the same column structure as the FINSPID panel and can be passed into the same modeling and strategy functions without changing the analysis logic.

In this 2025 stage, we do not repeat the full horizon correlation exploration. The Finnhub based period is shorter by construction and news coverage patterns differ from FINSPID, which makes long horizon correlation estimates less stable and less informative for the specific objective of demonstrating out of sample robustness. Instead, we carry forward the horizon choice motivated by FINSPID and apply the same next day label definition and short rebalance strategy framework. This ensures that the 2025 analysis functions as a true robustness check, where the central question is whether the pipeline and short horizon trading logic remain coherent when applied to a different dataset, rather

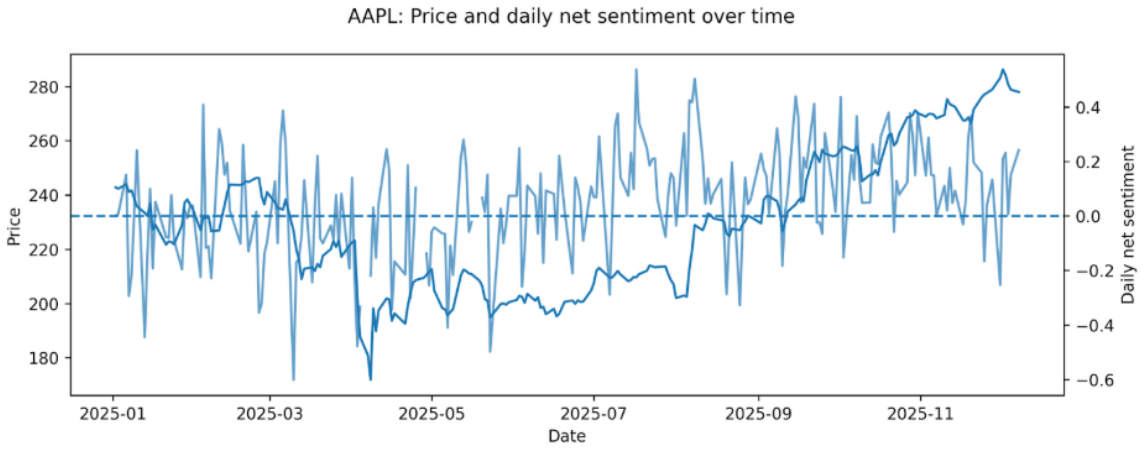

Figure 6: AAPL Price and Daily Net Sentiment Over Time (FINNHUB)

than whether we can discover a new horizon preference in a short window.

Finally, we evaluate the strategy in the same manner as in the FINSPID section by comparing the cumulative performance of the sentiment based long short portfolio to a simple equal weight hold benchmark. The trading rule is unchanged. On each rebalance date, tickers are ranked by the chosen sentiment signal, the highest quantile is assigned to the long leg, and the lowest quantile is assigned to the short leg, with positions held until the next rebalance. This construction is intentionally cross sectional because broad market moves can dominate single stock daily returns, and the long short structure reduces exposure to market direction so the test focuses on whether sentiment helps discriminate winners from losers within the universe. The quantile cutoffs are a practical way to trade only when the signal is strongest, which reduces sensitivity to small sentiment fluctuations and avoids continuously reweighting positions in a way that would inflate turnover. We report performance against an equal weight hold benchmark because it represents the simplest investable alternative on the same stock universe, so any outperformance is attributable to the sentiment overlay rather than differences in stock coverage or universe selection.The benchmark is computed as the equal weight average return across the same ticker universe. This consistent implementation allows the results section to focus on how performance and stability change between the historical FINSPID development sample and the 2025 out of sample panel, and to interpret differences in terms of data coverage and current regime behavior rather than methodological inconsistency.

4.1 LLM-Based Text Processing of SEC Filings

Each cleaned filing is processed using a large language model through a structured prompting framework designed to ensure consistency across observations. To control computational cost and maintain feasibility, the analysis focuses on the first 9,000 characters of each filing. This portion typically contains executive summaries, high-level performance discussion, and early risk disclosures, which are often the most informative sections for investors.

The LLM is instructed to output a strict JSON format, yielding a set of numerical and categorical features. These features capture dimensions such as overall sentiment, perceived risk and uncertainty, tone of forward-looking statements, and the presence of qualitative red-flag indicators. Importantly, this approach avoids reliance on static sentiment dictionaries or manually engineered keyword lists. Instead, it leverages the model's contextual understanding of language, allowing sentiment and risk assessments to vary with phrasing, context and narrative structure.

The extracted features are aggregated into a structured, tabular dataset suitable for supervised machine learning. At this stage, no price-based, accounting, or factor-related variables are included, ensuring that the predictive model relies exclusively on textual information derived from corporate disclosures.

4.2 Feature Engineering and Machine Learning Framework

4.2.1 Feature Construction

The LLM-derived features undergo standard preprocessing to ensure compatibility with the machine learning model. Continuous variables, such as sentiment intensity and risk scores, are standardized to account for differences in scale. Categorical features, including tone classifications, are transformed using one-hot encoding. This preprocessing step facilitates efficient model training while preserving the informational content of the original features.

The resulting feature matrix is intentionally parsimonious. Rather than maximizing feature count, the design prioritzes interpretability and robustness, reflecting the relatively low signal-to-noise ratio characteristic of short-horizon financial prediction tasks.

4.2.2 Predictive Model

A Light Gradient Boosting Machine (LightGBM) classifier is employed to predict whether the five-day forward return following a filing is positive. LightGBM is a gradient-boosted decision tree algorithm that is particularly well suited to this application for several reasons. It naturally captures nonlinear relationships between features, models interactions without explicit specification, and remains robust in the presence of noisy and heterogeneous inputs; a common feature of text-derived financial signals.

The prediction task is framed as a binary classification problem, where the target variable equals one if the forward return is positive and zero otherwise. This formulation reflects the empirical characteristics of short-horizon equity returns, which are often heavy-tailed and difficult to model in magnitude. Directional prediction is therefore more stable and aligns closely with the construction of long-short trading strategies.

4.2.3 Training and Evaluation

To avoid look-ahead bias, the dataset is sorted chronologically by filing date and split using a time-based train-test framework. The model is trained on the earliest 70% of observations and evaluated on the remaining 30%, ensuring that all test-period predictions are genuinely out of sample.

Model performance is assessed using both statistical and economic criteria. In addition to predictive accuracy metrics, the analysis evaluates the realized returns of trading strategies derived from the model's predictions. This dual evaluation framework ensures that statistical improvements translate into economically meaningful outcomes.

4.3 Trading Strategy Construction

Model predictions are translated into an event-driven long-short trading strategy. On each filing date, stocks are ranked according to their predicted probability of a positive forward return. Securities in the top signal quantile are assigned to the long portfolio, while those in the bottom quantile are assigned to the short portfolio. Positions are equal-weighted and held over a one-day horizon.

This construction ensures that trades are initiated exclusively in response to new information arrivals, closely aligning the strategy with the underlying economic mechanism of disclosure-driven price adjustment. By restricting trading activity to filling dates, the approach avoids unnecessary turnover and isolates the contribution of textual information to return generation.

4.4 Factor-Based Benchmark and Residualization (FF5)

A central objective of this study is to evaluate whether textual signals extracted using large language models provide incremental information beyond well-established sources of systematic risk. To this end, we adopt the Fama–French Five-Factor (FF5) model as a benchmark representation of long-horizon expected returns. The FF5 framework captures widely documented risk premia related to market exposure, firm size, value, profitability, and investment, and serves as a standard baseline in empirical asset pricing.

Using FF5 as a long-horizon anchor serves two complementary purposes. First, it provides a disciplined benchmark against which the economic significance of text-based signals can be assessed. If LLM-derived predictions simply proxy for factor exposure, then any apparent return predictability

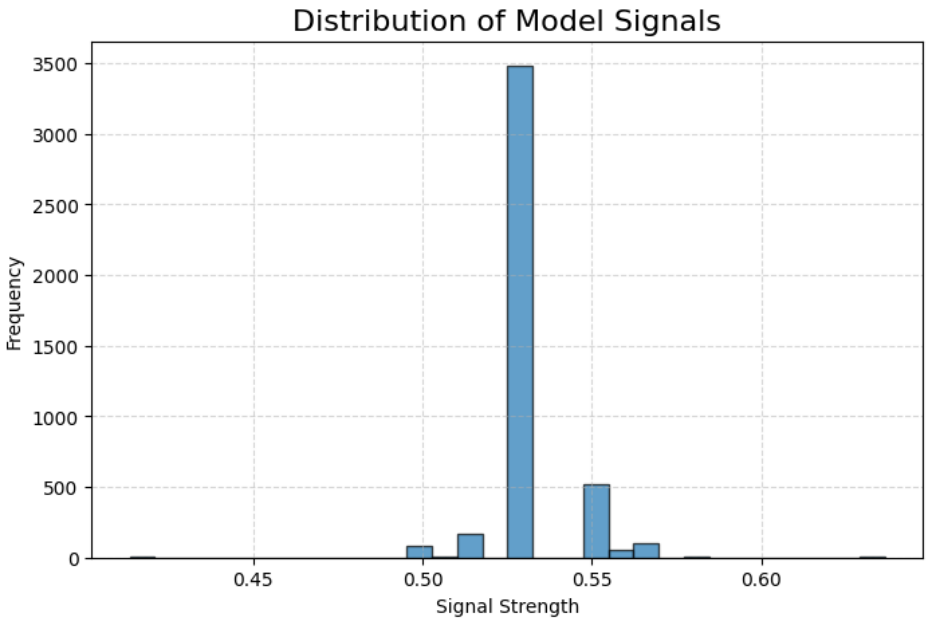

Figure 7: This figure shows the cross-sectional distribution of predicted probabilities produced by the LightGBM classifier. The distribution exhibits substantial dispersion, indicating that the model differentiates meaningfully across filings rather than producing near constant predictions.

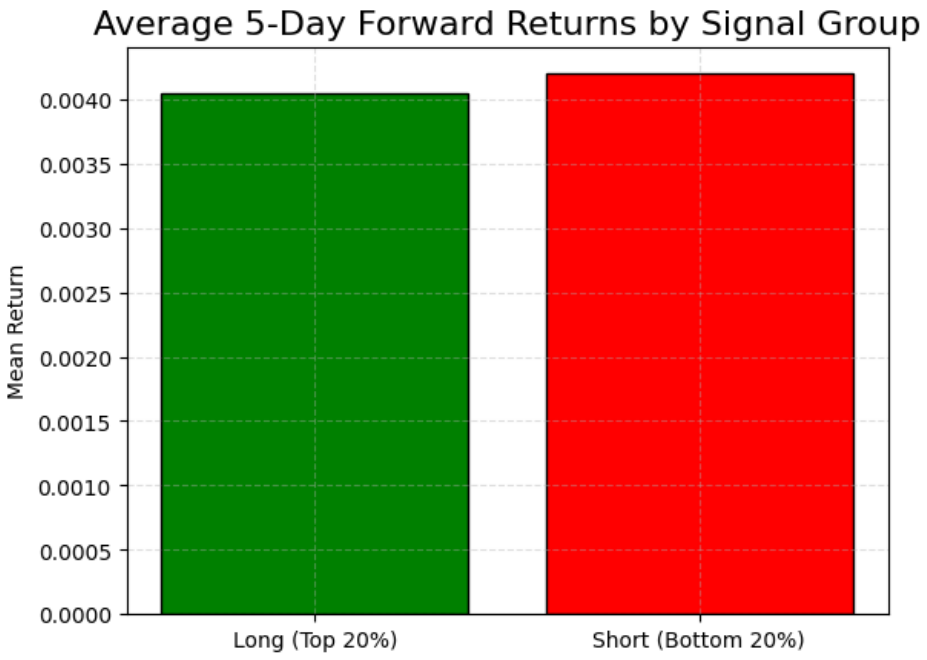

Figure 8: This figure illustrates average forward returns across signal-sorted portfolios. Returns increase monotonically with predicted signal strength, providing direct evidence that the model's outputs are economically informative.

may reflect compensation for systematic risk rather than information extracted from unstructured text. Benchmarking against FF5 therefore helps distinguish genuine informational content from implicit factor loading.

Second, the FF5 model facilitates a decomposition of realized returns into systematic and idiosyncratic components. For stock i on day t, excess returns are modeled as:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_{i,M}\,(r_{M,t} - r_{f,t}) + \beta_{i,S}\,\mathrm{SMB}_t + \beta_{i,H}\,\mathrm{HML}_t + \beta_{i,R}\,\mathrm{RMW}_t + \beta_{i,C}\,\mathrm{CMA}_t + \varepsilon_{i,t}, \quad (1)$$

where $r_{i,t}$ denotes the return on stock i, $r_{f,t}$ is the risk-free rate, $r_{M,t}$ is the market return, and SMB_t, HML_t, RMW_t, and CMA_t are the FF5 factor returns. The residual $\varepsilon_{i,t}$ represents the component of returns unexplained by systematic factor exposures and is interpreted as an idiosyncratic or abnormal return.

This residualization step plays a key role in the signaling strategy. Rather than predicting raw returns directly, the filing-based LLM model is tasked with predicting the direction of factor-adjusted residual returns following disclosure events. In practice, we construct a factor-adjusted forward return target over a horizon h (e.g., $h = 5$ trading days) by aggregating residual returns over the post-event window and defining a directional label as:

$$y_{i,t}^{(h)} = 1\left(\sum_{k=1}^{h} \varepsilon_{i,t+k} > 0\right), \quad (2)$$

where $1(\cdot)$ denotes the indicator function.

In the integrated framework, FF5 anchors long-run expected returns, while LLM-based signals operate as overlays that target deviations from factor-implied performance. This separation of horizons enhances interpretability and robustness by aligning each modeling component with the type of information it is intended to capture. By explicitly conditioning on factor exposures, the analysis ensures that any incremental predictive power attributed to textual features reflects information beyond that already embedded in traditional asset pricing models.

5 Empirical Results

5.1 New Sentiment Strategy Results

5.1.1 Initial Strategy Versus Equal Weight Hold

We begin with a single, fixed strategy specification to establish a clear baseline comparison before introducing any parameter search. Using the same cross sectional long short construction described in Methodology, we rank stocks each rebalance date by a chosen sentiment signal, go long the top quantile, and go short the bottom quantile, holding positions until the next rebalance. Performance is evaluated against an equal weight hold benchmark computed from the same universe of stocks, which isolates the incremental value of the sentiment overlay rather than changes in stock selection.

On the FINSPID out of sample window, the initial strategy based on daily net sentiment produces a positive cumulative return but generally trails the equal weight hold benchmark over the same period. The net sentiment strategy displays periods of outperformance early in the sample, followed by a flatter profile later, which is consistent with the idea that any sentiment edge is short lived and can be overwhelmed by sustained market trends captured by the hold portfolio. In contrast, the initial strategy using the surprise based feature performs weaker in the FINSPID window, remaining below zero for much of the period and only recovering near the end. This suggests that, in our implementation, the surprise definition may be noisy or may not align well with the short horizon label being traded.

The same fixed specification is then applied to the 2025 out of sample panel without altering the strategy mechanics. In this window, the equal weight hold benchmark trends upward, while both the net sentiment and surprise based strategies are flat to negative and finish below the benchmark. Because the 2025 period is shorter and has different news coverage density, this initial test is primarily used to check whether the pipeline produces coherent outputs and whether the sign and stability of the strategy are consistent with the earlier period. Taken together, these initial comparisons motivate a broader question. If a single parameter choice does not consistently improve on hold across both periods, is there a nearby configuration of the same transparent strategy family that performs more robustly. This motivates the parameter sweep reported next.

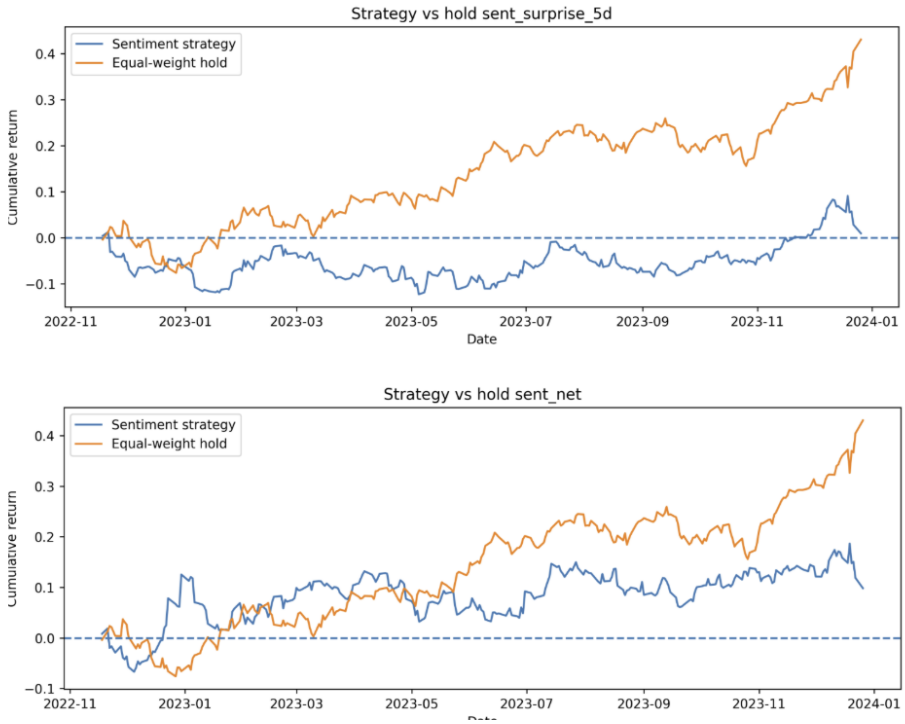

Figure 9: Strategy Tests vs Hold dor Sentiment Net & Surprise (FINSPID)

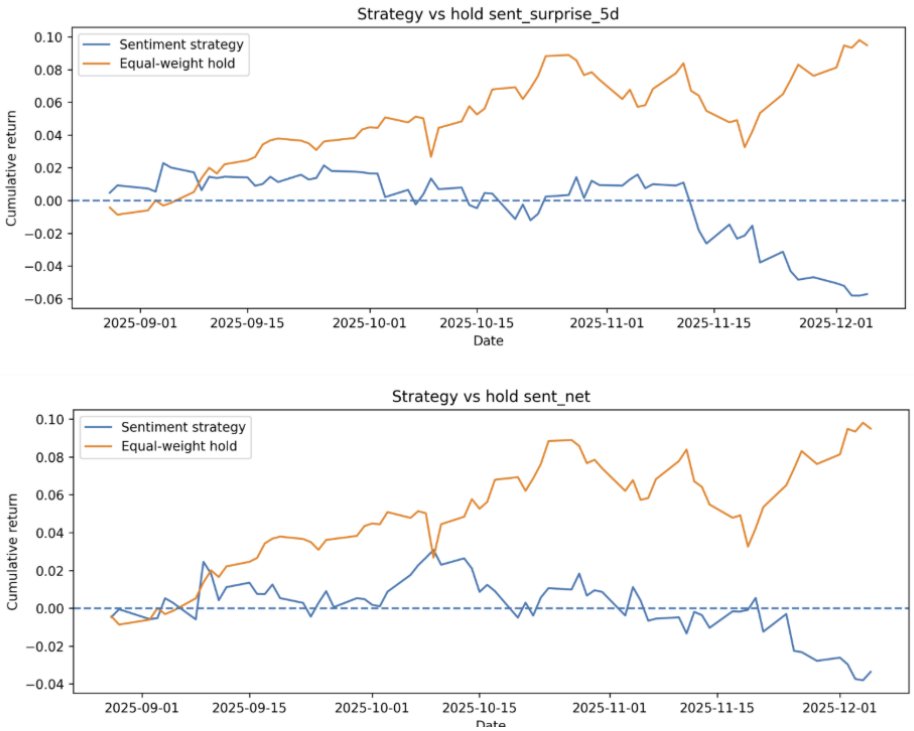

Figure 10: Strategy Tests vs Hold for Sentiment Net & Surprise (Finnhub)

Top strategies (ranked by avg Sharpe, then avg return across both periods)

	feature	rebalance_days	long_q	short_q	avg_sharpe	avg_total_return	sharpe_2020	ret_2020	sharpe_2025	ret_2025
0	sent_net	1	0.80	0.20	2.885138	1.195165	6.058159	2.403687	-0.287883	-0.013357
1	sent_net	1	0.75	0.20	2.765291	1.137384	6.228701	2.302479	-0.698119	-0.027710
2	sent_net	1	0.80	0.25	2.642797	1.019379	5.816894	2.059492	-0.531299	-0.020735
3	sent_net	1	0.70	0.20	2.641900	1.101644	6.210085	2.238480	-0.926284	-0.035191
4	sent_net	1	0.80	0.30	2.502567	0.923187	5.646611	1.870830	-0.641476	-0.024456
5	sent_net	1	0.75	0.25	2.491952	0.966445	5.962337	1.967890	-0.978433	-0.035000
6	sent_net	1	0.70	0.25	2.357597	0.933867	5.931990	1.910205	-1.216796	-0.042470
7	sent_net	1	0.75	0.30	2.332419	0.873001	5.778232	1.784599	-1.113394	-0.038596
8	sent_net	1	0.70	0.30	2.201558	0.842223	5.750534	1.730515	-1.347419	-0.046069
9	sent_surprise_5d	1	0.70	0.20	1.924406	0.663708	4.634650	1.353395	-0.785837	-0.025979

Figure 11: Top Strategies ranked by average Sharpe ratio and average return across both out of sample windows

5.1.2 Parameter Sweep Methodology and Selection Rule

To avoid over interpreting one arbitrary configuration, we run a small grid search over the same strategy family. Each candidate strategy is defined by the sentiment feature used to rank stocks, the rebalance interval in trading days, and the long and short quantile cutoffs that determine which names enter each side of the portfolio. For each parameter combination, we compute daily strategy returns and summarize performance with total return and an annualized Sharpe proxy based on daily returns. Importantly, we evaluate each candidate separately on the FINSPID out of sample window and on the 2025 out of sample window using the same benchmark definition.

To emphasize robustness rather than a single period fit, strategies are ranked by the average Sharpe across the two out of sample windows, with average total return used as a secondary tie breaker. This ranking rule penalizes configurations that perform well in only one period but fail in the other.

5.1.3 Sweep Results and Top Parameter Settings

To complement the single specification backtests, we ran a small sweep over strategy settings and ranked candidates by average Sharpe across the two out of sample periods, with average total return as a secondary tie breaker. This ranking rule is intentionally conservative because it rewards strategies that are strong in FINSPID and do not completely fail in the later 2025 robustness window. The above figure shows a clear pattern in the highest ranked configurations. The top of the ranking is dominated by the same feature, daily net sentiment, paired with very frequent rebalancing and relatively extreme long and short cutoffs. For example, the leading configurations all rebalance daily and use long quantiles between 0.70 and 0.80 with short quantiles between 0.20 and 0.30. This concentration is informative because it suggests that, within our simple strategy family, performance is not evenly spread across signals and parameters. Instead, the best outcomes occur when the strategy focuses on the most extreme cross sectional sentiment rankings and refreshes positions quickly, which is consistent with earlier evidence that any sentiment return relationship is strongest at short horizons and fades as the horizon length increases.

At the same time, Table 2 also highlights an important robustness limitation. The high average Sharpe values are driven primarily by very strong FINSPID performance, while the 2025 window contributes weak or negative Sharpe and slightly negative total return for the same parameter settings. This implies that the strategies that look best historically are not reliably transferable to the later period, even when we explicitly rank by an average across both windows. We therefore interpret the sweep as identifying which configurations best capture the FINSPID era dynamics, while the 2025 results serve as a check on whether those dynamics persist under a different market regime and under a dataset with substantially lower sentiment coverage.

In this afigure, the top five strategies generate large cumulative gains and remain clustered, indicating that performance is not isolated to a single parameter choice. In Figure 11, those same top ranked strategies do not replicate the same advantage and generally underperform the equal weight hold benchmark during 2025. This gap motivates the diagnostic discussion in the next subsection, where we connect performance differences to differences in sentiment coverage and the effective cross sectional sample size available on each rebalance date.

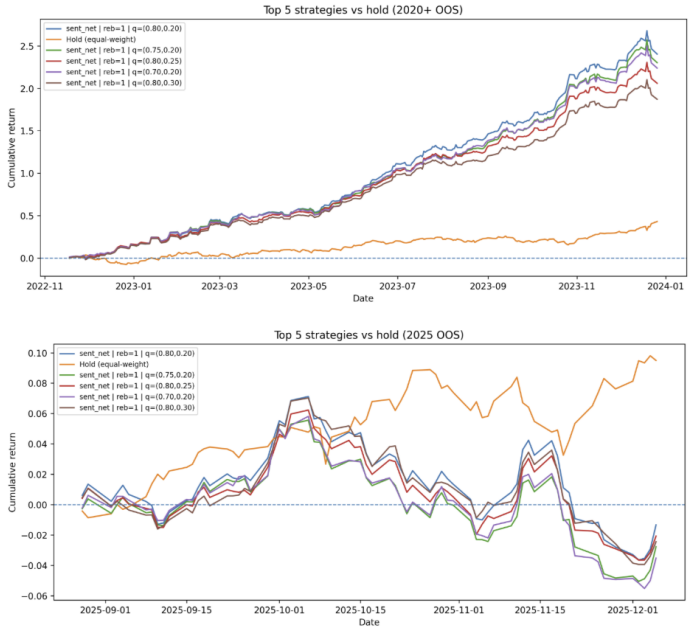

Figure 12: Top Five Strategies vs Hold for both periods

5.1.4 Interpreting the Cross Period Gap Using Coverage Diagnostics

A key diagnostic for interpreting the difference between Figure R5 and Figure R6 is the share of ticker days with usable sentiment information. In the FINSPID panel, sentiment features are available on the majority of ticker days, approximately 94 percent, which supports stable daily ranking across the universe. In the 2025 pipeline, sentiment coverage is substantially lower, approximately 13 percent, meaning that on many days most tickers have no news and therefore no reliable daily sentiment aggregates. Sparse coverage reduces the effective cross sectional sample size on each rebalance date, increases noise in quantile assignment, and can cause the strategy to behave more like a concentrated bet on a small subset of names.

This coverage gap provides a direct and measurable explanation for why the FINSPID backtest can produce strong looking performance while the 2025 robustness check is more challenging. It also motivates the discussion section, where we separate results driven by genuine predictive content from results that may depend on data availability, market regime, or implementation frictions.

5.2 Filing-Based LLM Strategy

The LLM-based strategy produces a cumulative out-of-sample return of approximately 6.3%, corresponding to an annualized return of 12.4% and a Sharpe ration of 1.55. Returns are episodic, with performance concentrated around periods of heightening filing activity, reflecting the event-driven structure of the strategy.

Sector-level analysis reveals meaningful heterogeneity in predictive performance. Technology and consumer discretionary firms exhibit stronger signals, while more regulated sectors show weaker effects. This variation is consistent with differences in disclosure practices and managerial discretion across industries. Firms with greater latitude in narrative framing appear to convey more informative linguistic signals, which the model is able to exploit.

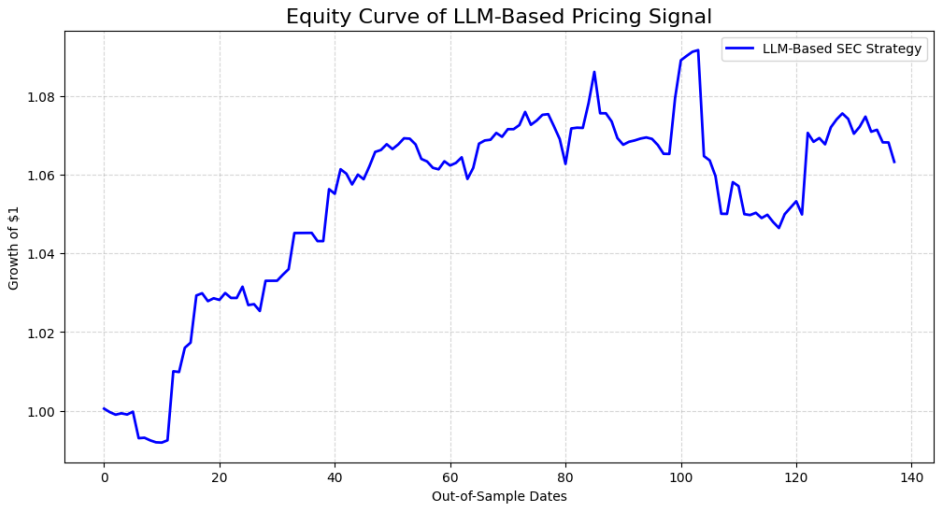

Figure 13: Figure 3 presents the cumulative return of the LLM-based strategy over the out-of-sample period.

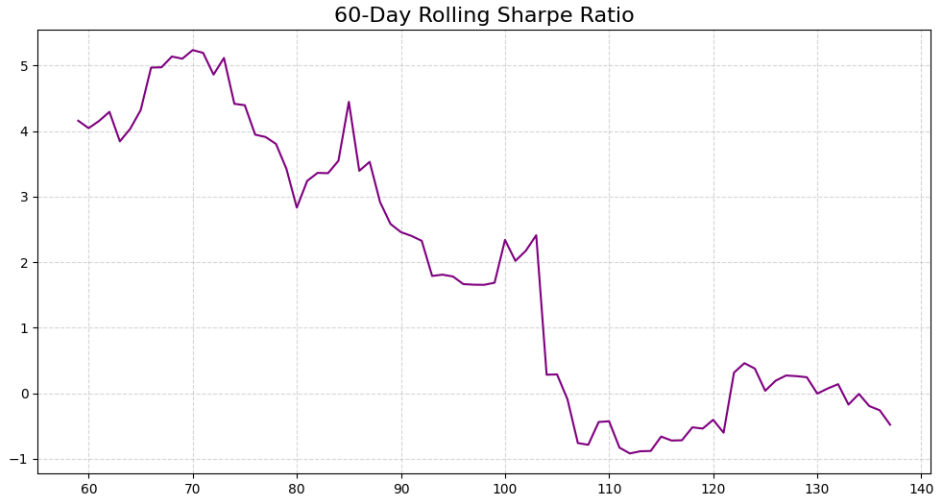

Figure 14: Figure 4 reports a rolling Sharpe ratio, highlighting the episodic nature of event-driven performance.

6 Comparison with the Fama-French Five Factor Model

A FF5 benchmark is constructed over the same period. As expected, the FF5 strategy delivers higher cumulative and risk-adjusted returns. However, regression analysis indicates minimal correlation between the LLM-based strategy and FF5 factors. This orthogonality suggests that textual disclosures capture a distinct information channel related to corporate communication rather than structural risk premia. Consequently, the LLM-based strategy may offer diversification benefits when combined with traditional factor models.

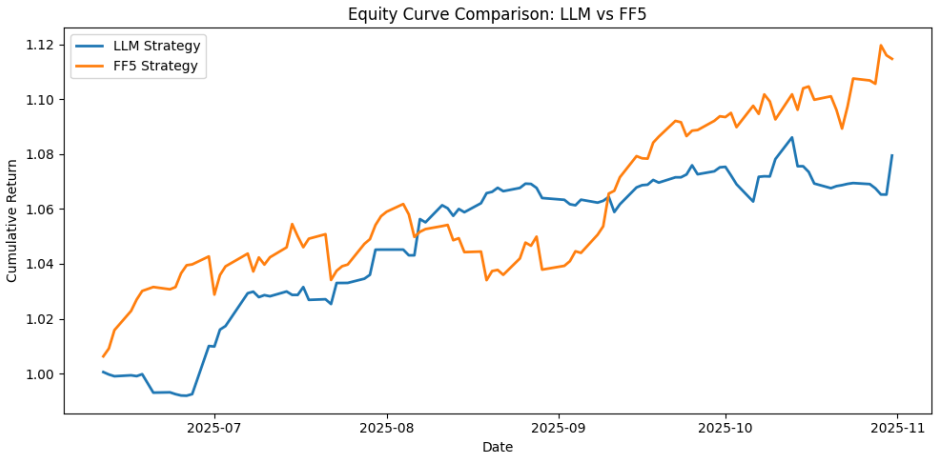

Figure 15: The above figure presents the cumulative return of the FF5-based strategy against the LLM strategy over the out-of-sample period.

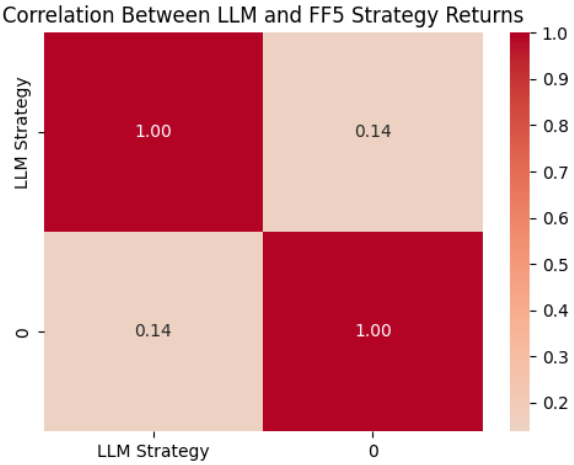

Figure 16: Corrlation plot between FF5 and SEC Filings models illustrates orthogonality of strategies.

Strategy	FF5 Signal	LLM Signal
Cumulative Return	11.47%	6.32%
Annualized Return	31.48%	12.44%
Annualized Volatility	7.86%	7.40%
Sharpe Ratio	4.00	1.55
Max Drawdown	2.78%	4.51%

Table 2: Signal Strategy Metrics.

7 Comparison of Signal Strategy Against S&P 500

7.1 Motivation

Including the S&P 500 as a benchmark serves a critical role in contextualizing the performance of the proposed LLM-based trading strategy. In empirical asset pricing, absolute returns alone are rarely informative unless evaluated relative to the prevailing market environment. Periods of strong market performance can inflate returns across a wide range of strategies, creating the risk of attributing success to model skill when it may instead reflect broad market exposure. As commonly noted in the investment literature, it is essential not to confuse genuine predictive ability with favorable market conditions.

The S&P500 represents the most widely accepted proxy for aggregate market performance and investor opportunity cost. By evaluating the LLM strategy alongside the S&P 500 over the same out-of-sample window, the analysis explicitly controls for market-wide tailwinds or headwinds. This comparison allows the reader to distinguish between returns generated through exposure to systematic market risk and returns arising from information extracted from corporate disclosures. Without such a benchmark, it would be difficult to assess whether the LLM signal provides any value beyond passive market participation.

Importantly, the purpose of this comparison is not to position the LLM strategy as a replacement for market exposure. Rather, it serves to highlight the distinct risk profile and informational content of the signal. While the S&P500 delivers higher cumulative returns over the same period, it does so with materially greater volatility and deeper drawdowns. In contrast, the LLM-based strategy exhibits smaller drawdowns, lower return variance, and more stable risk-adjusted performance. These characteristics are particularly relevant in professional portfolio management, where capital preservation and drawdown control are often prioritized over maximizing nominal returns.

From a portfolio construction perspective, low-volatility and low-drawdown strategies are highly desirable even when their standalone returns are modest. Such strategies can materially improved portfolio efficiency when combined with higher-risk assets. The empirical results suggest that the LLM signal functions precisely in this capacity: as a defensive, information-driven alpha source that performs consistently across market conditions rather than amplifying exposure during bull markets. The comparison with the S&P500 therefore strengthens the argument that the LLM strategy captures genuine informational value rather than market beta.

In summary, benchmarking against the S&P 500 enhances the credibility of the empirical analysis by anchoring the results in the broader market context. It demonstrates that the LLM-based signal is not merely a byproduct of favorable market conditions but instead reflects a distinct return-generating mechanism. This positioning aligns the strategy with practical investment objectives, where the goal is not to outperform the market at all times, but to construct robust, diversified portfolios that balance return generation with effective risk management.

7.2 Results

Although the S&P500 index delivers the highest cumulative return over the evaluation period, the comparative analysis reveals several important distinctions in the risk and return dynamics of the three strategies.

First, the equity curve comparison shows that the LLM-based strategy produces a more gradual and stable accumulation of returns. Unlike the S&P 500 and FF5 strategies, whihc experience pronounced fluctuations, the LLM strategy evolves through smaller, incremental gains tied to discrete information

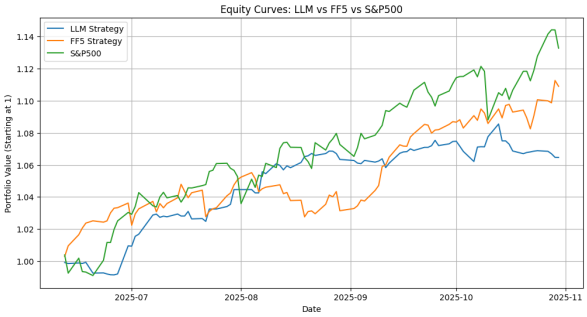

Figure 17: Cumulative equity curves for the LLM-based filing strategy, the Fama-French Five-Factor strategy, and the S&P 500 benchmark during the out-of-sample evaluation period.

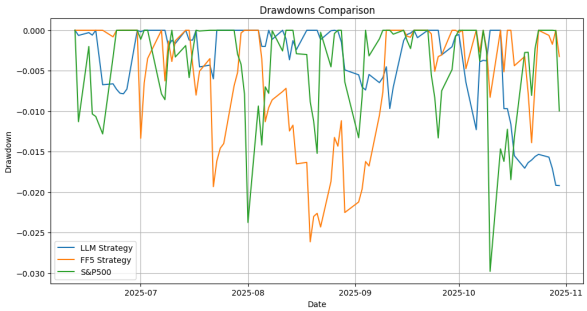

Figure 18: Comparison of the drawdown of the LLM-based filing strategy, FF5 based strategy and S&P 500 benchmark over the out-of-sample period.

events. This behavior is consistent with the event-driven nature of SEC filings, where predictive opportunities arise sporadically rather than continuously.

Second, the drawdown analysis provides particularly strong evidence in favor of the LLM signal. While both the S&P 500 and FF5 strategies experience deeper and more volatile drawdowns, the LLM strategy exhibits materially shallower drawdowns and faster recoveries. This indicates superior downside protection and suggests that the LLM signal avoids prolonged exposure during adverse market conditions. In practice, this characteristic is highly valuable to portfolio managers, as controlling drawdowns is often more important than maximizing raw returns.

Third, the rolling 60-day Sharpe ratio highlights the risk-adjusted strength of the LLM approach. Over the evaluation window, the LLM strategy consistently achieves the highest rolling Sharpe ratio among the three strategies, indicating that its returns are generated more efficiently per unit of risk. This result reinforces the idea that the LLM model extracts meaningful non-random information from corporate disclosures rather than simply loading onto broad market risk factors.

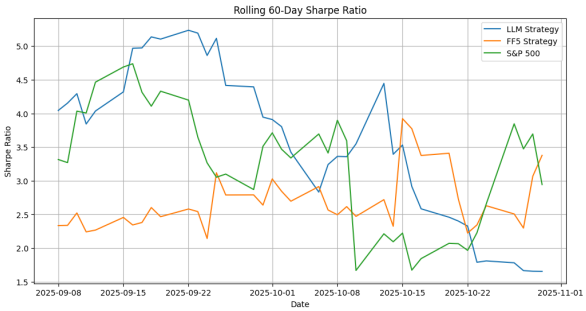

Figure 19: Rolling 60-day Sharpe ratios for the LLM-based strategy, the Fama-French Five-Factor strategy, and the S&P 500 benchmark.

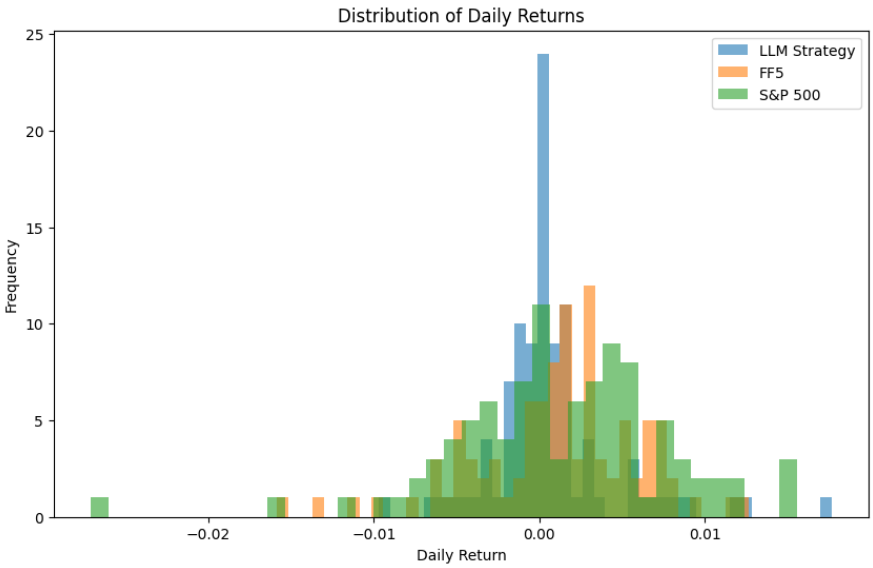

Figure 20: Distribution of daily returns for the LLM-based strategy, the Fama–French Five-Factor strategy, and the S&P 500 benchmark.

Finally, the distribution of daily returns further illustrates the distinct behavior of the LLM strategy. The return distribution is tightly centered around zero with relatively low variance, suggesting that the model produces frequent small signals rather than rare extreme outcomes. In contrast, the S&P 500 displays substantially higher variance and heavier tails, reflecting exposure to market-wide shocks. This difference implies that the LLM strategy behaves more like a controlled alpha signal than a directional market bet.

Taken together, these results suggest that while the LLM-based strategy does not dominate traditional benchmarks on a standalone cumulative-return basis, it offers meaningful advantages in terms of drawdown control, risk adjusted performance, and return stability. Crucially, the low correlation with both FF5 and market returns implies that the LLM signal provides an independent source of information, making it a strong candidate for inclusion within a diversified multi-factor portfolio.

8 An Extension: Text-Based Prediction of Factor-Adjusted Returns

8.1 Methodological Extension: Integrating LLM-Based Signals with the Fama-French Framework

The previous sections of this project established that large language models can extract meaningful signals from SEC filings and that these signals generate statistically and economically relevant return patterns when evaluated in isolation. The LLM-based strategy exhibited attractive risk-adjusted performance characteristics, including relatively shallow drawdowns and low correlation with traditional factor-based strategies. At the same time, comparison with Fama-French Five Factor model demonstrated that a substantial portion of realized equity returns is attributable to well-documented systematic risk premia rather than firm-specific informational content.

While these results provide important insights, they also raise a natural follow-up question: does the information embedded in corporate disclosure language provide predictive value beyond what is already explained by standard asset pricing factors? Addressing this question requires a methodology that explicitly disentangles systematic factor-driven returns from idiosyncratic, information-driven price movements. Motivated by this consideration, the present section extends the earlier analysis by combining the LLM signaling framework with the FF5 model in a unified, two- stage approach.

The extension builds directly on the methodological foundations of the earlier chapters. Rather than replacing either the LLM strategy or the FF5 framework, the objective is to assess whether textual

signals extracted from SEC filings explain residual return variation after controlling for established risk factors. This approach aligns closely with the asset pricing literature, where new predictors are typically evaluated based on their ability to explain abnormal returns than total returns.

8.2 Two-Stage Modeling Framework

The combined methodology adopts a sequential modeling structure. In the first stage, expected returns are estimated using the Fam-French Five-Factor model. In the second stage, a machine-learning model trained on LLM-derived textual features is used to predict the residual component of returns that remains unexplained by the factor model.

8.2.1 Stage 1: Factor-Based Return Decomposition

Consistent with the earlier benchmarking analysis, stock returns are first decomposed using the Fama–French Five-Factor model. For each firm i and filing date t, excess returns are modeled as:

$$r_{i,t} - r_{f,t} = \alpha + \beta_{\text{MKT}}\,\text{MKT}_t + \beta_{\text{SMB}}\,\text{SMB}_t + \beta_{\text{HML}}\,\text{HML}_t + \beta_{\text{RMW}}\,\text{RMW}_t + \beta_{\text{CMA}}\,\text{CMA}_t + \varepsilon_{i,t}, \quad (3)$$

where $r_{i,t}$ denotes the return on stock i, $r_{f,t}$ is the risk-free rate, and MKT_t, SMB_t, HML_t, RMW_t, and CMA_t represent the market, size, value, profitability, and investment factors, respectively. The estimated residual term $\varepsilon_{i,t}$ represents the portion of returns not explained by systematic risk premia and is interpreted as abnormal or idiosyncratic return.

The motivation for this decomposition follows directly from the findings of the earlier sections. While the LLM-based strategy demonstrated promising performance, much of the variation in raw returns can still be attributed to factor exposure and general market conditions. By explicitly removing these components, the analysis ensures that any remaining predictability is attributable to firm-specific information rather than implicit factor loading.

8.2.2 Stage 2: Text-Based Prediction of Residual Returns

In the second stage, the residual returns obtained from the FF5 model serve as the prediction for the LLM-based machine-learning model. Rather than forecasting total returns, the model is tasked with predicting the direction of abnormal returns following a filing event.

Specifically, residual returns are converted into a binary classification target, indicating whether the abnormal return over the subsequent five trading days is positive or negative. This formulation emphasizes directional predictability and avoids the instability associated with modeling the magnitude of residual returns directly, which are often noisy and highly variable.

The feature set consists exclusively of LLM-derived linguistic indicators extracted from SEC filings, including sentiment measures, risk-related language, and categorical tone classifications. A Light Gradient Boosting Machine (LightGBM) classifier is employed to capture nonlinear relationships and interactions between these textual features. The choice of LightGBM reflects its strong performance in financial prediction tasks, particularly in environment characterized by low signal-to-noise ratios and heterogeneous feature effects.

Importantly, this modeling stage builds directly on the LLM framework developed earlier in the dissertation. The same feature extraction process, preprocessing steps, and time-based train-test splits are retained to ensure comparability with prior results.

8.3 Out-of-Sample Evaluation and Portfolio Construction

As in the standalone LLM strategy, model evaluation is conducted using a strict out-of-sample framework. The dataset is sorted chronologically by filing date and split into training and testing periods, ensuring that no future information is used during model estimation.

Out-of-sample predictions are converted into a trading signal by ranking firms based on predicted probabilities of positive residual returns. On each filing date, an event-driven long-short portfolio is formed by taking long positions in firms with the highest predicted residual performance and short positions in firms with the lowest. Portfolio returns are computed using equal weighting and are evaluated exclusively on filing days.

This portfolio construction methodology mirrors the approach used in the earlier LLM analysis, allowing for direct comparison between raw-return based signals and residual based signals.

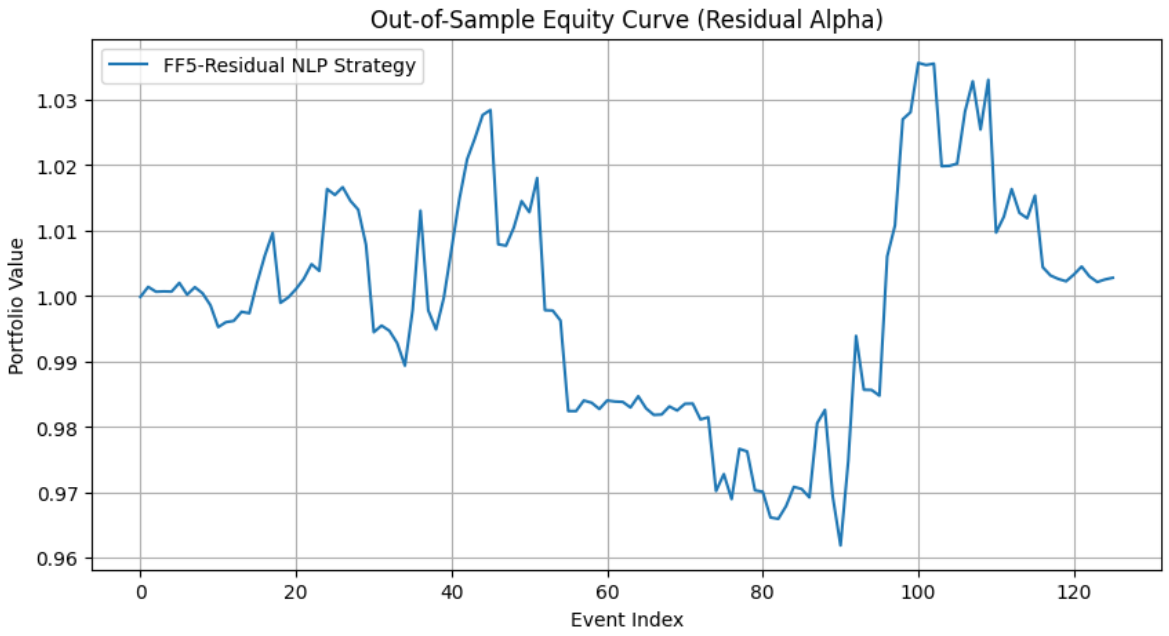

Figure 21: Out-of-sample cumulative residual returns of the SEC filing–based strategy after adjusting for Fama–French Five-Factor exposures.

8.4 Interpretation and Motivation for the Extended Approach

The combined LLM-FF5 methodology represents a deliberately conservative test of textual predictability. By removing expected returns implied by a widely accepted asset pricing model, the analysis isolates a narrow but economically meaningful component of return variation. Positive performance under this framework should not interpreted as a replacement for factor-based investing, but rather as evidence that corporate disclosure language contains incremental information not captured by traditional models.

The modest magnitude of residual-based returns is consistent with the empirical asset pricing literature, where most returns are explained by systematic risk premia and abnormal returns tend to be small. However, even small residual predictability is economically significant, particularly when it exhibits low correlation with existing strategies. In this sense, the extended methodology complements the earlier findings by demonstrating that the LLM signal is not merely a proxy for factor exposure or market trends.

This extension therefore strengthens the overall contribution of the dissertation. While the standalone LLM strategy highlights the practical feasibility of text-based trading signals, the combined approach situates these signals firmly within the asset pricing framework. Together, the results suggest that LLM-derived textual features represent a distinct and additive source of information, with potential applications in multi-factor portfolio construction and risk management.

Finally, to further evaluate the economic content of the LLM-based signal, this section examines its ability to rank firms cross-sectionally based on expected abnormal returns. Rather than focusing solely on portfolio-level performance, the analysis groups filing events into quantiles according to the model's predicted signal and computes average forward residual returns within each group. This approach provides a direct test of whether higher predicted signal values correspond to systematically different return outcomes after controlling for standard factor exposures. By abstracting from specific portfolio construction choices, the quantile analysis offers a transparent assessment of the model's ranking power and helps distinguish genuine informational content from performance driven by a small number of trades or market-wide effects.

Figure 14 presents average FF5 adjusted forward returns by signal quantile in the out-of-sample period. While the relationship is not strictly monotonic, the distribution shows substantial dispersion in residual returns across quantiles. Several signal buckets exhibit economically meaningful positive or negative residual performance, indicating that the LLM captures nonlinear patterns in disclosure

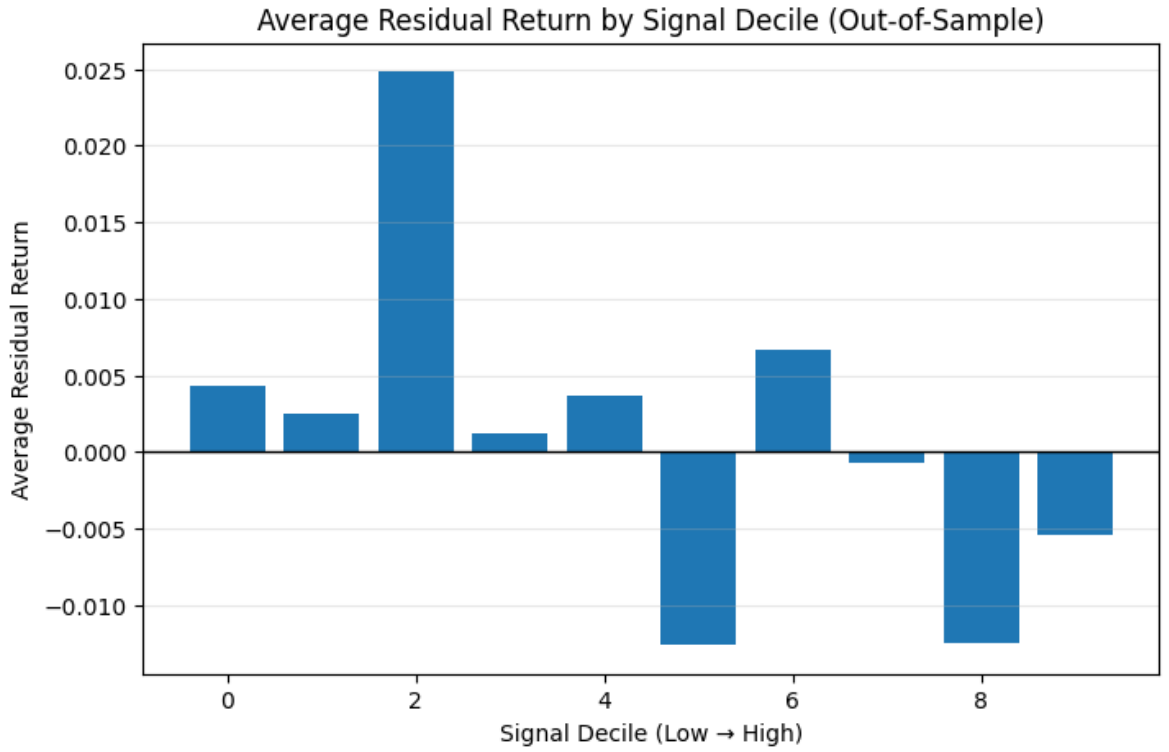

Figure 22: Average five-day forward residual returns by signal decile in the out-of-sample period.

language that are not explained by traditional factor exposures. This behavior is consistent with an event-driven signal whose predictive power is concentrated in specific linguistic regimes rather uniformly distributed across ranks.

9 Bridging the Two Pipelines: A Unified Text-Based Return Framework

Although both components of this project rely on large language models to extract signals from unstructured text, they differ fundamentally in information timing, persistence, and economic interpretation. Financial news articles and regulatory filings represent two distinct channels through which information enters markets, and their predictive value naturally operates at different horizons.

Financial news is characterized by high frequency, rapid dissemination, and strong short-term market impact. As demonstrated in the news sentiment analysis, sentiment signals extracted from headlines and summaries exhibit their strongest correlations with returns at very short horizons and decay quickly thereafter.This behavior is consistent with an information flow model in which news acts as a short-lived shock to investor beliefs, leading to brief price adjustments that are quickly arbitraged away.

In contrast, SEC filings are infrequent, information-dense, and legally constrained disclosures. The filing based LLM pipeline shows that linguistic features extracted from 10-K, 10-Q, and 8-K filings generate event-driven return predictability over multi-day horizons and exhibit low correlation with traditional risk factors. Rather than capturing immediate sentiment reactions, filing language reflects managerial tone, risk disclosure, and strategic positioning; information that the market may process more slowly due to its complexity.

These differences suggest that the two text sources should not be viewed as substitutes, but as complements operating at different points along the return horizon spectrum. News sentiment captures transient deviations in expectations, while filing language captures deeper firm-level information shocks. Importantly, neither channel is designed to explain the persistent component of returns driven by systematic risk premia, which motivates the inclusion of a factor-based framework.

10 Limitations and Future Research

10.1 News Sentiment Analysis Model

10.1.1 Success and Financial Intuition

Across the FINSPID out of sample window, the most consistent positive results came from strategies built on daily net sentiment and implemented as a cross sectional long short rule with frequent rebalancing and relatively extreme quantile cutoffs. Two pieces of intermediate evidence support this design choice. First, the horizon correlation diagnostics indicate that the relationship between sentiment features and forward returns is strongest at very short horizons and fades as the horizon length increases. Second, the parameter sweep concentrates its top ranked strategies around daily rebalancing and extreme quantiles, suggesting that the signal to noise ratio is highest when the strategy focuses on the most informative tail observations rather than trading marginal sentiment changes. From an implementation standpoint, the pipeline itself was a success. We were able to move from raw news and prices to an aligned daily panel with reproducible features, and then apply the exact same modeling and strategy functions to both the historical FINSPID panel and the 2025 rebuilt panel. This consistency is important because it means differences in results are less likely to be caused by changes in methodology.

The strategy construction reflects a simple hypothesis about information arrival. News sentiment should matter most when it represents a relatively new shock to beliefs and when prices have not fully incorporated it. This fits a short horizon design where the strategy refreshes positions frequently and relies on cross sectional ranking, because the key question is whether sentiment helps identify relative winners and losers among comparable large cap names on the same day. Using a long short structure also reduces market direction exposure, which makes the evaluation more about the signal than the overall market trend during the backtest window. The poor transfer to the 2025 window is also interpretable in financial terms. If sentiment effects are regime dependent, then a period with different macro conditions, different market concentration, or different information processing speed could weaken the relationship. In addition, when daily sentiment is sparse, the cross sectional ranking becomes less representative of the whole universe and the strategy can end up reflecting noise from a small subset of news heavy tickers rather than broad information.

10.1.2 Practical Constraints and Limitations

A central limitation is data coverage and comparability between sources. FINSPID provides broad and dense historical coverage, which supports stable daily aggregation and consistent cross sectional ranking. In contrast, the Finnhub based 2025 panel has much lower coverage, meaning many ticker days have no news and sentiment features are missing or weakly estimated. This directly affects strategy performance because quantile selection becomes noisier and the effective number of tradable names per rebalance date shrinks. There are also methodological limitations that matter for interpretation. The backtests do not include a full set of trading frictions such as bid ask spreads, borrow costs for shorts, and realistic turnover constraints, and frequent rebalancing in particular would likely reduce realized performance in live trading. The strategy family is intentionally simple, but that simplicity also implies that we are not conditioning on other known drivers of returns such as momentum, size, or sector exposures. Finally, the parameter sweep introduces selection risk, and while we rank strategies by performance averaged across two out of sample windows, the 2025 window is short and therefore has higher estimation noise.

10.1.3 Areas for Improvement

A key lesson is that strong historical results can be tightly linked to data quality and coverage rather than purely to model sophistication. The FINSPID results show that when sentiment features are available for most ticker days, a simple cross sectional rule can appear economically meaningful. The 2025 robustness check shows that the same logic can break down when coverage is sparse or when market dynamics differ. A second lesson is that intermediate diagnostics are essential. The horizon correlation plots provided a concrete reason to focus on short horizons and frequent rebalancing, and the coverage statistics explained why the later period behaves differently. Without these intermediate

checks, it would be easy to treat the strategy as universally effective or universally ineffective, when the evidence instead points to conditional effectiveness.

10.2 SEC Filings LLM Model

While the empirical results provide encouraging evidence that large language models can extract economically meaningful information from corporate disclosures, several limitations of the present study should be acknowledged.

First the textual processing of SEC filings is necessarily incomplete. To control computational cost and ensure tractability, only a truncated portion of each filing is analyzed. As a result, potentially informative sections, such as detailed Management Discussion and Analysis (MD&A), risk factor disclosures, and footnotes are excluded from the model's input. These sections often contain nuanced discussions of firm-specific risks, strategic shifts, and forward-looking uncertainties that may carry additional predictive value. Future research could address this limitation through full-document ingestion using chunking strategies or retrieval-augmented approaches that allow large documents to be processed efficiently without exceeding model constraints.

Second, the linguistic feature set employed in this study remains relatively coarse. Although sentiment, tone, and risk-related indicators capture broad patterns in disclosure language, they do not fully exploit the representational power of modern language models. Embedding-based representations, which encode semantic relationships across sentences and documents, may provide a richer and more flexible feature space. Incorporating such representations could improve the model's ability to detect subtle shifts in narrative framing, changes in managerial confidence, or deviations from historical disclosure patterns.

Third, while LightGBM offers a strong and computationally efficient baseline for modeling nonlinear relationships, it is not necessarily optimal for all aspects of textual prediction. Alternative machine-learning approaches, such as XGBoost, ensemble architectures, or hybrid models combining tree-based methods with neural embeddings, may yield improvements in predictive stability and generalization. Additionally, the current classification framework focuses on directional prediction over a fixed five-day horizon. Extending the analysis to multiple horizons or continuous returns targets could provide further insight into the temporal dynamics of disclosure-driven price adjustments.

Finally, the portfolio construction and evaluation framework remains intentionally simplified. The analysis does not account for transaction costs, liquidity constraints, or overlapping holding periods, all of which are relevant for assessing real-world implementability. Incorporating these elements would allow for a more comprehensive evaluation of economic viability and risk-adjusted performance.

Taken together, these limitations highlight substantial opportunities for future research. Enhancing document coverage, expanding feature representations, adopting more expressive modeling architectures, and refining portfolio construction techniques may materially improve the performance and robustness of LLM-based trading strategies.

11 An Integrated Multi-Horizon Strategy Framework

The empirical findings from both pipelines motivate a unified return prediction framework that explicitly separates long-term expected returns, medium-term disclosure-driven alpha, and short-term sentiment effects.

At the foundation of the strategy lies the Fama-French Five Factor (FF5) model, which captures well documented sources of long-run expected returns related to market risk, size, value, profitability, and investment. As shown in the filing-based analysis, FF5 explains a substantial portion of realized returns but leaves economically meaningful residual variation unexplained. This residual component provides a natural target for text-based signals.

Building on this structure, SEC filing-based LLM signals are used as a medium-horizon alpha layer. By predicting the direction of factor-adjusted residual returns following filing events, the strategy isolates firm-specific information embedded in corporate disclosures rather than systematic risk exposure. Since filings occur at at discrete intervals, this component is implemented as an event-driven overlay with relatively low turnover.

Finally, financial news sentiment is incorporated as a short-horizon tactical adjustment. Given the rapid decay documented in the news sentiment results, these signals are best suited for short

holding periods or for dynamically tilting positions around existing exposures. Rather than serving as a standalone return driver, news sentiment can refine timing; scaling exposure up or down in response to unusually strong positive or negative sentiment shocks.

Conceptually, the combined strategy operated as follows: 1. Baseline allocation determined by FF5 expected returns (long term). 2. Event-driven alpha overlay activated around SEC filing dates using LLM-extracted disclosure features. 3. Short term sentiment filter or tilt applied using news-based sentiment signals to manage timing and near-term risk. This layered approach aligns each signal with the horizon at which it is most informative, reducing overfitting risk and improving interpretability. By construction, the framework also mitigates the fragility observed when short-horizon sentiment strategies are evaluated in isolation, while preserving their informational value when used selectively.

12 Conclusion

This study examines whether large language models can be used to extract economically meaningful predictive financial signals from unstructured text and how much signals relate to established asset pricing frameworks.Focusing on two complementary information channels, corporate filings and firm-specific financial news, we demonstrate that LLM-derived textual features contain predictive information that can be transformed into systematic trading signals when aligned with appropriate investment horizons. Using SEC filings from S&P500 firms, the analysis demonstrates that linguistic features capturing sentiment, tone, and perceived risk generate statistically meaningful return predictability at short-to-medium horizons. A standalone filing-based strategy delivers positive risk-adjusted performance with relatively shallow drawdowns and return dynamics that differ from traditional factor-based strategies. Benchmarking against the Fama-French Five Factor (FF5) model confirms that a substantial portion of realized equity returns is driven by systematic risk premia, yet the filing-based signals exhibit low correlation with these factors. An integrated two-stage approach that combines FF5 residualization with LLM-based prediction further provides evidence that corporate disclosure language contains incremental information beyond standard factor exposures, albeit with modest magnitude.

In parallel, the analysis of financial news sentiment demonstrates that high-frequency textual signals can be informative over very short horizons. Using a reproducible pipeline to convert raw news into daily sentiment measures, we find that cross-sectional strategies based on extreme sentiment can generate strong out-of-sample performance in historical data with dense coverage. However, these results prove fragile when applied to a later period constructed from independent news sources, underscoring the importance of data coverage, regime dependence, and execution constraints. In contrast to filing-based strategies, news sentiment signals are highly sensitive to turnover and implementation design, with lower-turnover applications such as sentiment-based filtering exhibiting greater robustness.

Taken together, the findings suggest that LLM-based text signals are best viewed not as substitutes for traditional asset pricing models, but as complementary sources of information that operate at different horizons. Regulatory filings convey structured, firm-specific information that unfolds over multiple trading days and is largely orthogonal to factor risk premia, while financial news captures fast-moving sentiment shocks with short-lived predictive power. Factor models such as FF5 anchor long-run expected returns, providing a natural baseline against which text-based signals can be evaluated and integrated.

Several extensions represent natural next steps for future research. First, expanding and merging news data sources could improve coverage and allow for more robust evaluation of short-horizon sentiment strategies. Second, incorporating realistic trading frictions, turnover penalties, and liquidity constraints would provide a more complete assessment of economic viability, particularly for high-frequency strategies. Third, exposure-neutral implementations, such as sector-neutral or beta neutral constructions, could help isolate pure informational effects. Fourth, exploring alternative labeling horizons and event-based windows may better align textual signals with the timing of information release. Finally, extending the analysis across additional years and market regimes would allow for a more comprehensive assessment of robustness and stability.

Overall, this study highlights the potential role of large language models in modern asset pricing research. By bridging unstructured textual data with systematic financial modeling, LLM-based approaches offer a promising avenue for incorporating qualitative information into quantitative investment processes, particularly when combined thoughtfully with established factor-based frameworks.

LLM-derived textual features contain economically meaningful information that can be transformed into tradable signals.

Three complementary strategies are evaluated. First, a standalone LLM-based strategy generates positive risk-adjusted performance, exhibiting relatively shallow drawdowns and distinct return dynamics compared to traditional factor models. Second, benchmarking against the Fama-French model confirms that much of the observed equity returns is driven by systematic risk premia rather than firm-specific information. Finally, an integrated two-stage approach, that combines FF5 residualization with LLM-based prediction, provides evidence that disclosure language contains incremental information beyond standard factor exposures, albeit with modest magnitude.

While the LLM-based strategies do not outperform established factor models in isolation, they capture a largely orthogonal source of return variation linked to corporate communication rather than price-based risk factors. This independence is arguably the most important contribution of the analysis. It suggests that text-based signals derived from large language models can serve as valuable complements to traditional quantitative strategies, particularly within multi-factor or multi-source portfolio frameworks.

Overall, the findings highlight the potential role of large language models in modern asset pricing research. By bridging unstructured textual data and systematic financial modeling, LLM-based approaches offer a promising avenue for incorporating qualitative information into quantitative investment processes. Continued methodological refinement and integration with established frameworks may further enhance their relevance for both academic research and practical portfolio management.

References

References

[1] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models (arXiv:1908.10063). *arXiv*.

[2] Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Journal of Finance*, forthcoming.

[3] Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

[4] Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3), 1327–1370.

[5] Finnhub. (n.d.). Company news endpoint (API documentation).

[6] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

[7] Koc, U., Karakaya, A., & others. (2024). FINSPID: A financial news sentiment dataset (arXiv:2402.06698). *arXiv*.

[8] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

[9] Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230.

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[11] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.