The Inefficient Pricing of News Shocks

Antoine Didisheim, Bryan Kelly, Mo Pourmohammadi, Hanqing Tian*

March 14, 2026

Abstract

The stock market fails to efficiently process information in news text (Chen et al., 2024). But news itself is highly predictable by prevailing stock characteristics, which complicates inferences about market efficiency. After purging news of its predictable content, the resulting "news shocks" more than double the monthly return predictive power of raw news, and they continue to significantly predict returns up to 18 months ahead. The magnitude and longevity of the news shock anomaly is larger than every anomaly in the Jensen et al. (2022) universe. Markets overreact or underreact to news depending on the saliency and opacity of the news topic. The news shock anomaly derives almost entirely from opaque news topics to which investors have the strongest tendency to underreact.

JEL Code: G11, G12

Keywords: News text, return prediction, text embeddings, LLM, return anomaly, overreaction, underreaction

*Antoine Didisheim is at University of Melbourne. Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER. Mo Pourmohammadi is at Yale School of Management. Hanqing Tian is at University of Melbourne. AQR Capital Management is a global investment management firm that may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

1 Introduction

Few topics are as central to economics as market efficiency. The extent of efficiency determines how effectively investment capital is allocated for productive uses. It dictates whether savvy investors can earn returns larger than justified by their risk (and whether passive investors suffer lower returns). Efficient prices establish guideposts for corporate decisions and government policy (Hayek, 1945; Baumol, 1965; Fama, 1970; Dow and Gorton, 1997; Feldman and Schmidt, 2003).

The field of behavioral economics has accumulated ample evidence of inefficiencies in financial markets and theories for their origins (Barberis, 2018; Hirshleifer, 2015). Yet the questions of how prices reflect new information—how accurately, how quickly, and which information—remain partially answered at best. In this paper we present a new analysis of price efficiency upon the arrival of financial news articles, using a data set whose coverage, detail, and timeliness are second-to-none in financial research.

The investigation of market behavior in light of news text has a long history in economics, dating at least to Cowles' (1933) manual reading of *The Wall Street Journal*'s editorials to predict returns of the Dow Jones Industrials Index. Advances in computation have made it possible to gradually expand the scope of research relating news text to market prices, including early forms of sentiment analysis (Antweiler and Frank, 2004; Tetlock, 2007) and machine learning analysis of word counts (Jegadeesh and Wu, 2013; Ke et al., 2019; Bybee et al., 2024).

Most recently, large language models (LLMs) refine examinations of price efficiency with far more comprehensive consideration of the information in news text (Lopez-Lira and Tang, 2024; Chen et al., 2024, CKX henceforth). CKX propose a particularly tractable procedure in which news is summarized via its numerical vector representation internal to an LLM—known as an "embedding"—which is then used in otherwise traditional regressions to investigate price responses to news.[1] These papers conclude that stock prices respond to news with a delay of up to a few days and that this inefficiency is large enough for sophisticated investors to earn a small but statistically significant profit after trading costs. The magnitude of this inefficiency appears in line with traditional behavioral theories of limits to arbitrage and limits to attention/cognition.

Building on the approach of CKX, we show that estimates of price responses to news article text are confounded by the fact that much of what we read in news media is not

[1]See CKX and Gentzkow et al. (2019) for a discussion of embeddings and their advantages for econometric analysis of text.

news at all—i.e., not in a statistical or economic sense. Instead, news article content is highly predictable by prevailing economic data. We show that it is only after purging news text of its predictable content that one arrives at a clearer picture of price responsiveness to news. Indeed, it is the *unpredictable* component of news text, what we call the "news shock," that the market responds to gradually and slowly, which gives rise to news-based return predictability.

On one hand, it is reassuring from a rational pricing theory perspective that investors appear to disregard predictable aspects of news and respond primarily to news shocks. However, by isolating the news shock we uncover evidence that markets digest news media far less efficiently than suggested by previous studies. The inefficient pricing of news shocks translates into an asset pricing anomaly that dwarfs other well known informational inefficiencies such as momentum, reversal, and post-earnings announcement drift. In fact, the news shock anomaly is roughly twice as large as any anomaly in the compilation of Jensen et al. (2022), JKP henceforth. News shocks present a new and particularly acute puzzle for asset pricing and behavioral finance.

Financial News is Predictable. Our first contribution is to show that a significant portion of the information contained in news articles can be predicted ahead of time. To demonstrate this, we follow CKX and use an LLM to represent articles as numerical embedding vectors. This embedding summarizes the semantic and contextual information detected by the LLM. State-of-the-art embeddings are such a comprehensive representation of text that an LLM can often recover the original text from the numerical embedding alone (e.g. Ge et al., 2023). Because all articles are represented as a fixed length numerical vector, embeddings are ideal for modeling the information content of text with relatively simple statistical tools such as regression.

Focusing on articles about individual stocks, we show that on average around 10% of variation in article embeddings can be predicted using only standard stock-level characteristics from the literature. This fact implies that prevailing stock-level data can be leveraged to purge news articles of their predictable content and isolate the arrival of truly new information. After regressing article embeddings on stock characteristics, we retrieve the unexpected content of news in the form of a residual embedding, which we call the "news shock." This innovation overcomes a core challenge in investigating price responsiveness to news media: separating old news that has already been incorporated in prices from the new information arrival that is relevant for re-valuing assets.

News Shocks: The Greatest Anomaly? CKX show that raw article embeddings are significant return predictors, and that most of the information in raw embeddings is incorporated in prices within a few days. Expanding on their analysis, we show that raw embeddings can be used to form profitable monthly trading strategies with Sharpe ratios, alphas, and turnover similar in magnitude to many anomalies.

These raw embeddings, however, understate the amount of return predictability contained in news articles. When we split raw embeddings into their predictable component and the news shock, we find that the predictable component has little power to predict returns. The fact that embeddings are predictable by stock characteristics means that return predictions from raw embeddings have significant overlap with well known anomalies. Most existing anomalies are comparatively small in magnitude, therefore the return predictability due to raw embedding is likewise muted.

In contrast, the residual embedding, or news shock, is an especially powerful predictor of returns. This is revealed only after the raw embedding is purged of its "old news" and we isolate the effect of *new* information arrival on prices. The return predictability associated with news shocks appears to be the largest price inefficiency documented to date. A long-short portfolio formed on the basis of news shocks produces an annualized Sharpe ratio of 3.1 over our sample 1996-2022. By comparison, the largest Sharpe ratio among the universe of JKP anomaly factors is 1.4 over the same period.[2] If we restrict analysis to stocks above the NYSE 50^{th} size percentile, the news shock Sharpe ratio is 1.4, versus 0.9 for the best performing JKP factor in this same large stock universe. We consider exhaustive robustness checks that vary the stock universe, the portfolio construction methodology, the LLM that generates embeddings, the news article data set, the stock characteristics used to predict news, the time sample, and so on. In every variation we consider, the conclusion is the same: the news shock anomaly is nearly double the size of the next largest anomaly in the JKP data.

Nature of the News Shocks Anomaly. LLM embeddings are highly efficient compressions of text and make it possible to build powerful text-based return prediction models, but they are uninterpretable to the human eye. We adapt a methodology from the machine learning literature to decode embeddings into interpretable news topics. From hundreds of thousands of distinct interpretable topics in raw text, we identify 12 economic themes whose news is inefficiently incorporated into prices and thus contribute to the news shock

[2]This corresponds to the cash-based operating profits-to-book assets (JKP mnemonic "cop_at") factor of Ball et al. (2016).

anomaly. They are: “Earnings & Financial Results;” “Corporate Guidance & Outlook;” “Analyst Ratings & Sentiment;” “Distress, Bankruptcy & Delisting;” “Momentum & Trading Activity;” “Corporate Actions & Restructuring;” “Leadership & Governance;” “Growth & Demand Trends;” “Biotech, Pharma & Healthcare;” “Regulatory & Legal Actions;” “Sector-Specific Signals;” and “Product Launches & Operations.”

Next, we show that different types of news translate into different price dynamics. Certain news topics are associated with large price reversals, while others lead to strong price momentum. For example, markets appear to significantly underreact to news about “bankruptcy and going-concern risk” and “product launch and approval,” which are economically complex topics that require careful reasoning through legal and regulatory details. Meanwhile, we also show evidence that markets overreact to news such as “deal announcements” and “analyst ratings.” These are high salience topics and fall into “extreme” news categories (e.g. Kwon and Tang, 2025). Our findings suggest that market efficiency is structurally dependent on information type: procedural complexity interacts with inattention to induce slow adjustments (underreaction), while high salience and volatility trigger extrapolation heuristics (overreaction). This dynamic is rationalized by the cognitive constraint model of Ba et al. (2024).

We then investigate which forms of inefficiency are responsible for the highly profitable news shock anomaly. Among all of the themes we identified, 85% of the news shock portfolio’s weight is assigned to underreaction topics. In short, the news shock anomaly is almost entirely dedicated to exploiting the market’s underreaction to news.

Look-ahead Bias in LLMs. A recent literature emphasizes the importance of guarding against look-ahead bias in pre-trained LLMs, particularly for financial applications (He et al., 2025; Sarkar and Vafa, 2024). To prove that our results are not driven by look-ahead bias, we repeat our analysis using “chronologically consistent" LLMs from He et al. (2025) that are trained recursively using only backward-looking data. This ensures that our trading strategy performance is not inflated by information about future realized returns that has been inadvertently stored in the LLMs’ parameters. The findings from this robustness test confirm the qualitative conclusions of our full sample analysis. While the magnitudes are somewhat weaker, the news shock anomaly constructed from chronologically consistent models continues to exceed that of any other anomaly in the JKP dataset. Furthermore, we note that news shocks derived from chronologically consistent models imply a lower bound on the true magnitude of the news shock anomaly because the degradation in performance comes not only from guarding against look-ahead bias, but also arises from the fact that chronologically consistent models are lower quality models than those used in our full sample (they rely on less sophisticated architectures, less training data, and less thorough training computation).

Literature Review Our paper builds upon and contributes to a literature that strives to incorporate textual data in empirical asset pricing. Gentzkow et al. (2019), Loughran and McDonald (2020), and Hoberg and Manela (2025) provide comprehensive surveys of this fast developing field. Our work is particularly related to the literature studying news text in return prediction and asset pricing. In addition to the work referenced earlier, this includes Jiang et al. (2021), Hoberg and Phillips (2018), Wang et al. (2018), Bybee et al. (2023), and Meursault et al. (2023), among others. We contribute to this literature by demonstrating that financial news text is highly predictable, and emphasizing the importance of modeling this predictability in order to isolate genuinely new information—shocks—in news text. We also reinforce the convenient and powerful role that LLM embeddings play in devising text-based asset pricing models (building on CKX).

Second, we contribute to a large literature studying asset pricing anomalies. While far too large to survey here, this literature is evaluated in a number of recent meta-studies including Hou et al. (2020a), Harvey et al. (2016), Novy-Marx and Velikov (2016, 2024), Chen and Zimmermann (2022), and JKP.

Third, we contribute to a literature studying the behavioral foundations of financial market inefficiencies, including Daniel et al. (1998), Hong and Stein (1999), Bordalo et al. (2019), Kwon and Tang (2025), and Ba et al. (2024), among many others.

2 Data

We use the following news and financial markets data for our analysis.

News. News article text data is from the Thomson Reuters Real-time News Feed ("Reuters") over the period January 1996 to December 2022. We clean and filter the data following the procedure of CKX. Most importantly, we restrict our analysis to articles tagged as associated with a single stock (according to the metadata provided by the vendor).

We remove articles with fewer than 100 characters or more than 100,000 characters, and exclude near-duplicate articles. One difference versus CKX is that our main analysis focuses on the higher quality news content from Reuters and excludes their less informative articles

Table 1: Article Summary Statistics

We reduce our Reuters news sample using a sequence of filters. First, we restrict the universe to articles provided by Reuters (NS:RTRS) and exclude all third-party providers. Second, we retain only articles associated with a single ticker to ensure clear firm-level identification. Third, we merge the articles with CRSP using the publication date to obtain the corresponding permno, discarding unmatched observations. Fourth, we remove articles containing fewer than 100 characters or more than 100,000 characters to eliminate noise and abnormal entries. Finally, we address textual redundancy by computing TF–IDF vectors and removing articles whose cosine similarity exceeds 0.8 relative to articles published by the same firm within the preceding five business days.

Raw Articles			Remove 3PTY	Tagged with Single Stock	Linked to Returns	Filter Short & Long	Filter Redundancy
Reuters	3PTY	Total	Total	Total	Total	Total	Total
11,747,260	3,304,518	15,051,778	11,747,260	10,075,064	8,569,207	7,670,692	6,680,550

aggregated from "third-party" (3PTY) news providers.[3] The final Reuters article count after applying all filters is 6.7 million. Robustness analyses in Section 6 show that including third-party news has virtually no effect on our results. While CKX perform a multinational analysis, we focus on US firms.[4] In Section 6 we also explore the robustness of our findings to using an alternative news source of similar quality and expansiveness: the Dow Jones Newswires.

Embeddings. "Embeddings," broadly speaking, refer to numerical vector representations of text. They are translations of human language into a numerical language suitable for statistical modeling. A successful translation of this sort should arrive at similar vectors for two texts that have similar meaning.

Some embeddings can be very simple, such as vectors of counts for certain terms in a text (or even scalar "sentiment" scores). Such simple embeddings, however, are poor representations of information in text (Baroni et al., 2014). With the advent of LLMs, modern embeddings are capable of expressing text meaning with extraordinary efficiency using embedding vectors of only a few hundred or few thousand dimensions (Gentzkow

[3] Article embeddings can vary substantially with differences in information quality, writing style, formatting, and editorial standards. Because we aggregate news embeddings of all articles for a given stock within each month, structural inconsistencies across data providers can accumulate and distort the aggregated embedding. Our choice to exclude third-party news is to maintain the purity of the embedding representation coming from the core Reuters data set.

[4] We intend to expand our analysis internationally in a future draft.

et al., 2019; Tao et al., 2024; Patil et al., 2023; Ash and Hansen, 2023). While the machine learning literature extols LLM embeddings for their efficacy in tasks like web search and multi-modal prediction (Huang et al., 2025), we are particularly influenced by CKX who emphasize the value of embeddings for modeling the link between text and asset returns.

We embed each news article using the *E5-Mistral-7B* model of Wang et al. (2024), which we refer to as simply "Mistral" henceforth. Naturally, there are many embedding models one might adopt and we analyze the sensitivity of our results to different open-weight LLMs in Section 6.[5]

Mistral embeddings represent each token within a document as a 4096-dimensional numerical vector. Following the methodology of CKX, we define an "article embedding" as the equally weighted average of token embeddings within the article. Next, we define a stock-month embedding as the equally weighted average of article embeddings for a given stock within a month. Stocks without any news coverage within the month are assigned a missing value.

A well known property of LLM embeddings is "anisotropy," meaning that embeddings between very different texts nonetheless tend to have a high degree of similarity (Gao et al., 2019). This arises in part from the natural structure of human language—documents often significantly overlap in terms of words and syntax even when referencing different topics. While anisotropy does not necessarily hinder LLMs in text generation and other natural language tasks, it can create a drag on the effectiveness of embeddings for downstream regression models (Gao et al., 2019).

The machine learning literature proposes various ways to mitigate this issue, including re-centering and re-scaling embeddings, which helps dampen their commonalities and amplify their distinct content (e.g. Su et al., 2021; Fei et al., 2021). Following this literature, we apply a Z-score normalization to all stock-month embeddings. Specifically, in each month t, each embedding coordinate is normalized by i) subtracting a pooled mean estimated from all stock-month observations through month $t-1$ and ii) dividing the difference by the pooled expanding standard deviation. Ultimately, we arrive at a stock-month news observation, denoted $E_{i,t}$, that is a 4,096-dimensional vector of averaged and Z-scored article embeddings.

Stock Characteristics and Returns. We use monthly excess returns and characteristics of individual stocks from JKP. We apply coverage filters following Didisheim et al. (2024)

[5]The open-weight nature of Mistral means that our LLM computations are performed privately on our own hardware and thus maintain privacy of our text data per our licensing agreement. This is not possible with popular "closed" models such as GPT-5 and Gemini.

Table 2: News Coverage and Stock Size

This table reports, for each market capitalization decile, the average percentage of firms with at least one news article per month. It also reports the average number of news articles for stocks conditional on having at least one article. The sample covers U.S. stocks from January 1996 to December 2022.

	10	20	30	40	50	60	70	80	90	100	All
% With News	30.0	35.4	39.5	44.0	47.9	53.2	58.2	64.5	70.9	81.5	52.5
# Articles	4.6	4.9	5.3	5.7	6.0	6.5	7.1	7.7	9.4	19.7	8.7

to arrive at the 132 JKP characteristics with the highest coverage over our sample period. All characteristics are then rank-standardized each month following Gu et al. (2020). The remaining missing characteristic observations are imputed with zeros (the cross-sectional mean rank within each month). We also use data for 25 GICS industry groups from JKP. Articles are aligned with stock identifiers based on precise time stamps. Our analysis is conducted on monthly data and all articles that arrive within the month are (conservatively) used only at month-end.

On average, our cross section consists of 4,198 stocks per month. On average 52.5% of our stocks have at least one news article in a given month. Conditional on having at least one article in a month, the average number of news articles per stock-month is 8.7. Table 2 reports more detailed coverage statistics by market capitalization decile (reported coverage percentages are conditional on having at least one article per month). In the largest 10% of stocks, on average 81.5% of stocks have at least one article in a given month, and the average number of articles per stock-month is 19.7. In the smallest decile, 30.0% of stocks have news in a given month and the average stock has 4.6 articles per month. The strong correlation between firm and news coverage naturally skews our sample toward larger and more liquid stocks.

3 Predicting News with Stock Characteristics

From this point forward, our use of text embeddings for return prediction deviates from CKX. Our analysis begins from the premise that substantial content of financial news articles can be anticipated ahead of time and therefore may not induce changes in stock prices. In this section we document the predictability of stock news and introduce an approach to isolating “news shocks" that strip out the predictable component of embeddings.

This first step in purging the common component in news among stocks is to remove the common embedding location and scale shared by all articles in our sample, accomplished with the global Z-score described in Section 2. It is likely, however, that even after adjusting for structure that is common to all news embeddings, a large degree of the remaining heterogeneity in news across stocks is predictable based on stock-level attributes. For example, conditional on a stock belonging to the technology sector, there is a sharp increase in the likelihood of an article discussing computer hardware and software and a decrease in the probability of news about, say, fertilizer. Stocks with high book-to-market ratios are more likely to see articles written about cash flow stability and dividends. Younger stocks are more likely to see news about growth trajectory. Highly levered stocks are more likely to experience news about debt covenants and credit ratings.

We investigate the news predictability hypothesis with regressions of news embeddings on stock characteristics. Let S_t be an $N_t \times L$ matrix of L stock characteristics for a universe of N_t stocks in month t (including a constant characteristic). Let E_t be the $N_t \times D$ matrix stacking at time t embeddings for all stocks in month t (for Mistral, $D = 4{,}096$). Each month we run the cross-sectional regression:

$$E_t = S_t\beta_t + \varepsilon_t, \tag{1}$$

with $L \times D$ coefficient matrix $\beta_t = (S_t'S_t)^{-1}S_t'E_t$. The fitted value $S_t\beta_t$ is the "predictable news" component. We refer to the residual embedding that is orthogonal to S_t, denoted ε_t, as the "news shock."

How much "news" is explained by prevailing asset characteristics? The regression in (1) produces an adjusted R^2 for each of the $4,096$ embedding coordinate dependent variables. Let $E_{i,t}(c)$ denote the c^{th} coordinate of the embedding vector for stock i at time t (and likewise for $\varepsilon_{i,t}(c)$). The coordinate-specific R^2 for c is defined as

$$R^2(c) = 1 - \frac{\sum_{i,t} \varepsilon_{i,t}(c)^2/(N-k)}{\sum_{i,t} E_{i,t}(c)^2/N}, \text{ where } N = \sum_t N_t, k = \sum_t L. \tag{2}$$

Because the embeddings are element-wise Z-scores, each coordinate has nearly zero mean,[6] thus we leave the denominator of the R^2 uncentered. Each coordinate also has approximately unit standard deviation, thus it is reasonable to calculate a pooled R^2 over all embedding

[6] This is not exact because the Z-score moments are calculated recursively in an expanding sample.

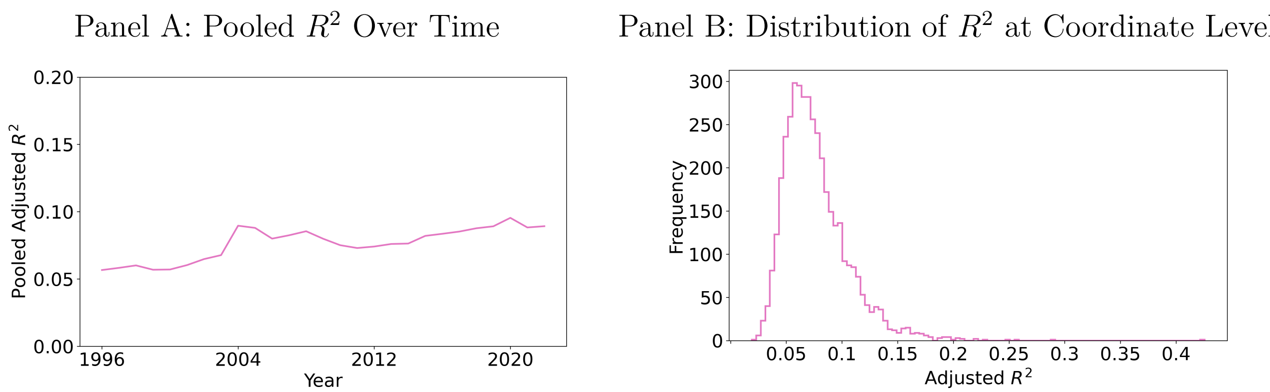


Figure 1: Adjusted R^2 of News Embeddings Predictability Based on Stock Characteristics

Note. This figure reports the adjusted R^2 for predicting stock-month embeddings using stock-level JKP characteristics. Panel A reports the monthly R^2 pooled over all coordinate dimensions. Panel B reports the distribution of average coordinate-level R^2 values.

coordinates to summarize the overall predictability of news:

$$R^2 = 1 - \frac{\sum_{i,t,c} \varepsilon_{i,t}(c)^2/(N-k)}{\sum_{i,t,c}\Big(E_{i,t}(c)\Big)^2/N}. \tag{3}$$

On average, we find that JKP stock characteristics predict news in our sample with a pooled R^2 of roughly 7.5%. Panel A of Figure 1 shows that the predictability of news is fairly consistent over time ranging from 5-10% and gradually increasing over time. Panel B shows a histogram of R^2's for individual embedding coordinates. Around 90% of coordinates have an average adjusted R^2 between 4% and 12%. This distribution shows that there are some aspects of news text that are more strongly predictable with stock characteristics, corresponding to a few embedding coordinates having R^2 over 25%.

Figure 2 shows the predictive gains from gradually adding characteristics to the model in terms of their effect on overall R^2. The first bar shows the effect of controlling only for market beta (denoted “CAPM"), which explains news with an R^2 of less than 1%. Next, we add market capitalization and book-to-market (denoted “FF3"), which increase the predicted variation in news to 2%. Adding profitability, investment, and momentum characteristics (denoted “FF6") raises the R^2 to 2.6%. The next bar shows the effect of adding 25 GICS industry indicators, raising the R^2 to 8%. The 132 JKP characteristics predict news with an R^2 of 7.6%, and JKP together with industry indicators culminate in an R^2 of 10.2%. The main takeaway from this figure is that both industry variables and JKP characteristics give

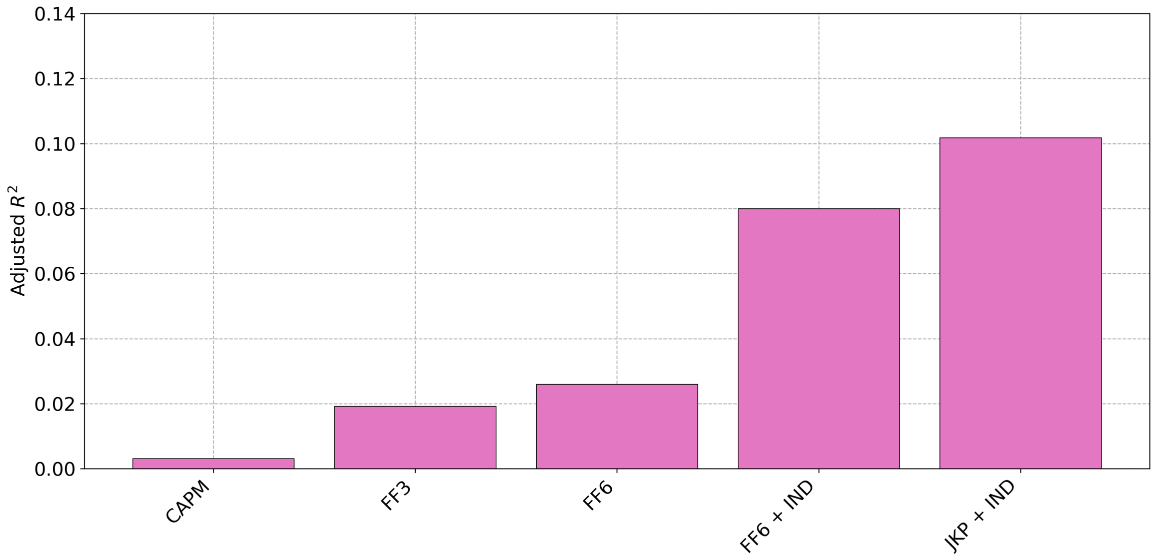


Figure 2: Adjusted R^2 by Group of Stock Characteristics

Note. This figure shows the pooled adjusted R^2 from equation 3 using different groups of stock characteristics. "CAPM" uses a stock's beta, "FF3" adds market capitalization and book-to-market, and "FF6" adds profitability, investment, and momentum characteristics. "IND" uses 25 GICS industry indicators, and "JKP" uses 132 anomaly characteristics.

a large boost in news predictability, and there appears to be a significant degree of shared information in these two predictor sets.

Table 3 further decomposes news predictability according to the 13 characteristic themes from JKP by running separate embedding regressions using only characteristics within a given theme. The results reveal a clear dichotomy between characteristics that describe a stock's fundamental identity and those that capture transient price dynamics. Return-based characteristics, such as momentum and short-term reversal, have low explanatory power for news. In other words, price trends are rarely an important driver of stock-level news. In contrast, characteristics based on persistent fundamental attributes (such as value, quality, leverage, risk, and profitability) are much stronger predictors of news. Their variable importances hover around 3% on average, triple that of momentum. These findings align with our hypothesis that embeddings encode structural economic narratives about firms.

The regressions in (1) explain realized news with contemporaneous stock characteristics. In Figure 3 we explore how news predictability is impacted if we instead regress embeddings

Table 3: Characteristic Importance for News Prediction

This table reports the distribution of adjusted R^2 from regressing embedding coordinates on subsets of firm characteristics. Characteristics are grouped into 13 themes defined in JKP. For each theme, we estimate Equation (1) for every embedding coordinate and report the mean, median, and key percentiles of the resulting R^2 distribution across coordinates. Themes are sorted by their mean adjusted R^2.

		Adjusted R^2 Percentiles				
Theme	Mean	25%	50%	75%	95%	Max
Value	0.037	0.023	0.032	0.045	0.074	0.320
Quality	0.034	0.021	0.029	0.042	0.068	0.277
Low Leverage	0.031	0.019	0.027	0.039	0.065	0.294
Low Risk	0.027	0.018	0.024	0.032	0.050	0.162
Profitability	0.026	0.016	0.022	0.032	0.051	0.135
Investment	0.025	0.016	0.022	0.031	0.050	0.178
Size	0.019	0.012	0.017	0.024	0.039	0.124
Seasonality	0.015	0.009	0.012	0.018	0.031	0.108
Momentum	0.014	0.009	0.012	0.017	0.027	0.114
Accruals	0.009	0.005	0.007	0.011	0.023	0.113
Debt Issuance	0.009	0.004	0.007	0.011	0.022	0.122
Profit Growth	0.009	0.006	0.008	0.011	0.016	0.063
ShortTerm Reversal	0.006	0.004	0.005	0.007	0.010	0.034

on past stock characteristics:

$$E_t = S_{t-j}\beta_t + \varepsilon_t \tag{4}$$

for lags $j \in \{1, 3, 6, 12\}$ months. Figure 3 plots the distribution of R^2 across embedding coordinates for the contemporaneous model ($j = 0$) versus the lagged specifications. The distributions of adjusted R^2 across individual embedding coordinates are more or less indistinguishable for lags up to 12 months. Evidently, news predictability is not driven by contemporaneous innovations in stock characteristics but instead by their long-lived attributes, consistent with the evidence in Table 3 that fundamental characteristics are the strongest predictors of news content.

In summary, the news shock ($\varepsilon_{i,t}$) collects what is left in news after filtering out expected content based on well established stock characteristics. The strongest news predictors are

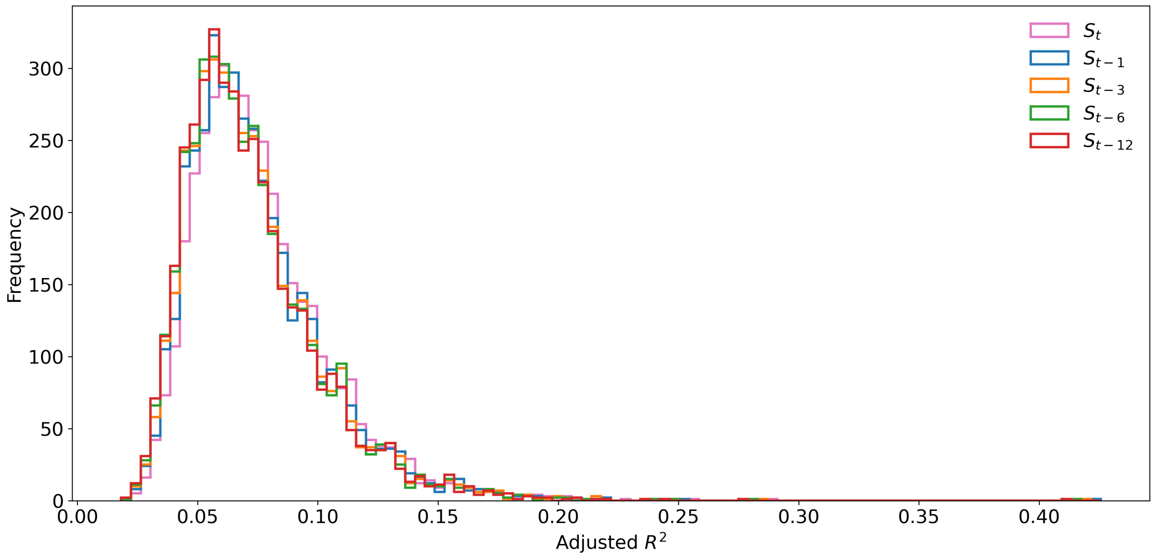


Figure 3: News Predictability With Lagged Stock Characteristics

Note. This figure reports the distribution of adjusted R^2 across embedding coordinates (E_t) when using lagged rather than contemporaneous JKP characteristics S_{t-j} for prediction. We consider lags of $j \in \{0, 1, 3, 6, 12\}$ months.

persistent characteristics associated with stock fundamentals. In contrast, news shocks represent the irregular, event-driven component of news.

4 News Shocks and Return Predictability

In this section we investigate stock return predictability with news text. News, however, is not a single predictor because—effective summarization of text requires a high-dimensional representation. LLMs achieve this by refracting news into thousands of signals that constitute the embedding. Therefore, when working with news text, the standard predictive analysis in the asset pricing literature of sorting on a scalar stock-level predictor requires modification.

Our main approach directly trains a long-short investment portfolio that leverages news embeddings via maximum Sharpe ratio regression (or "MSRR," see Kelly and Xiu, 2023). In Section 6, we show that our findings are robust to an alternative MSE-based approach that first predicts returns with embeddings then performs traditional portfolio sorts based on predicted returns, following CKX.

4.1 The MSRR Approach

Let $x_{i,t}$ be a generic D-dimensional vector of predictors for the one-month-ahead excess return of stock i, $R_{i,t+1}$. The $N \times D$ matrix of signals for all stocks is denoted X_t and the N-vector of returns is R_{t+1}. Our MSRR approach models stock-level portfolio weights as a linear function of stock characteristics:

$$w_{i,t} = x_{i,t}'b \quad \text{ or, in vector form, } \quad w_t = X_t b. \tag{5}$$

This portfolio rule is trained to optimize the unconditional Sharpe ratio (potentially subject to shrinkage) according to the objective function[7]

$$\min_b \sum_t \left(1 - b'X_t'R_{t+1}\right)^2 + \lambda b'b. \tag{6}$$

Kelly and Xiu (2023) explain that when $\lambda = 0$ this problem is equivalent to estimating the tangency portfolio of the characteristic-managed factors, $F_{t+1} = X_t'R_{t+1}$, and the MSRR estimator $\hat{b}$ that optimizes (6) are the tangency weights. The ridge penalty term $\lambda b'b$ shrinks the portfolio weights to stabilize portfolio estimates. The MSRR portfolio return is $F_{t+1}'\hat{b}$.

In our analysis we use news embeddings—either raw embeddings $E_{i,t}$ or residual "news shocks" $\varepsilon_{i,t}$—to stand in for the predictor vector $x_{i,t}$. Raw embeddings factors (denoted $F^E_{t+1} = E_t'R_{t+1}$) and news shock factors (denoted $F^\varepsilon_{t+1} = \varepsilon_t'R_{t+1}$) are linked through regression (1), and their difference equates to factors that trade "predictable news" (denoted $F^{E|S}_{t+1} = (S_t\hat{\beta}_t)'R_{t+1}$):

$$\begin{aligned} & E_t'R_{t+1} = (S_t\hat{\beta}_t)'R_{t+1} + \varepsilon_t'R_{t+1} \\ \Leftrightarrow \quad & F^E_{t+1} \quad = F^{E|S}_{t+1} \qquad + F^\varepsilon_{t+1}. \end{aligned} \tag{7}$$

All news factors, F^E, $F^{E|S}$, and F^ε, consist of D different news-managed portfolios that treat each individual embedding coordinate as a signal (viewing each month of stock news through the D-dimensional embedding prism referenced earlier). The MSRR estimator $\hat{b}$ aggregates factors for each embedding coordinate into a single comprehensive news portfolio, which we denote as $F^{\star E} = {F^E_{t+1}}'\hat{b}$ (likewise for $F^{\star E|S}$ and $F^{\star\varepsilon}$).

In our empirical analysis, we recursively estimate the MSRR portfolio weights at time t using an expanding rolling window through t, and then use these to construct the out-

[7] See Kelly and Xiu (2023), Brandt et al. (2009), and Britten-Jones (1999).

of-sample news portfolio return at $t + 1$ (we use an initial 12-month training window at the beginning of the sample). In each training sample we select the ridge penalty, λ, via leave-one-out cross-validation.

4.2 News Shock Anomaly Performance

Figure 4 reports performance of news-based strategies. Panel A shows the Sharpe ratio of the total embedding portfolio, $F^{\star E}$, in the green dashed line. This news portfolio has a Sharpe ratio of 1.1 (this portfolio is not dollar neutral because the embeddings do not have a zero mean in the cross section). The first set of bars shows the effect of cross-sectionally de-meaning the embeddings, which corresponds to running regression (1) while setting S_t to be a constant. This ensures that $F^{\star \varepsilon}$ is a dollar-neutral long-short portfolio, which raises the Sharpe ratio to 1.7. The remaining bars show the effect of residualizing embeddings with respect to an expanding set of predictors. First we consider stock beta along with the constant (denoted "CAPM"), then adding further market capitalization and book-to-market ratio ("FF3"), followed by the addition of investment, profitability, and momentum ("FF6"), and finally including all JKP characteristics.

As we add more predictors to S_t, the Sharpe ratio of the news shock strategy $F^{\star \varepsilon}$ gradually climbs, eventually reaching 3.1 when controlling for the full suite of JKP characteristics. The interpretation of this surprisingly strong news shock portfolio is that the unexpected component of news is poorly reflected in prices at the time it arrives. Trading on news shocks produces large and significant excess returns as prices respond to this information with a delay. Panel C of Figure 4 plots the cumulative return on the news shock strategy (based on the full set of JKP predictors). The strategy does not suffer any major drawdowns and does not appear to decay late in the sample, like many other anomalies (McLean and Pontiff, 2016; Pénasse, 2022).

In contrast, news that can be predicted ahead of time indeed appears priced-in ahead of time. Trading on "predictable news" in the form of $F^{\star E|S}$ produces smaller Sharpe ratios. Because $F^{\star E}$ and $F^{\star E|S}$ are not dollar neutral, both raw and fitted embeddings possess non-zero cross-sectional means, it is important to evaluate their performance in excess of the market. Thus Panel B reports alphas and their confidence intervals for each news strategy. Predictable news has insignificant alpha versus the market except in the JKP case. News shocks, on the other hand, have large and significant alpha in all cases. The muted (though non-negligible and statistically significant) alpha of the raw news portfolio $F^{\star E}$ arises from mixing highly informative news shocks with much less informative predictable news content.

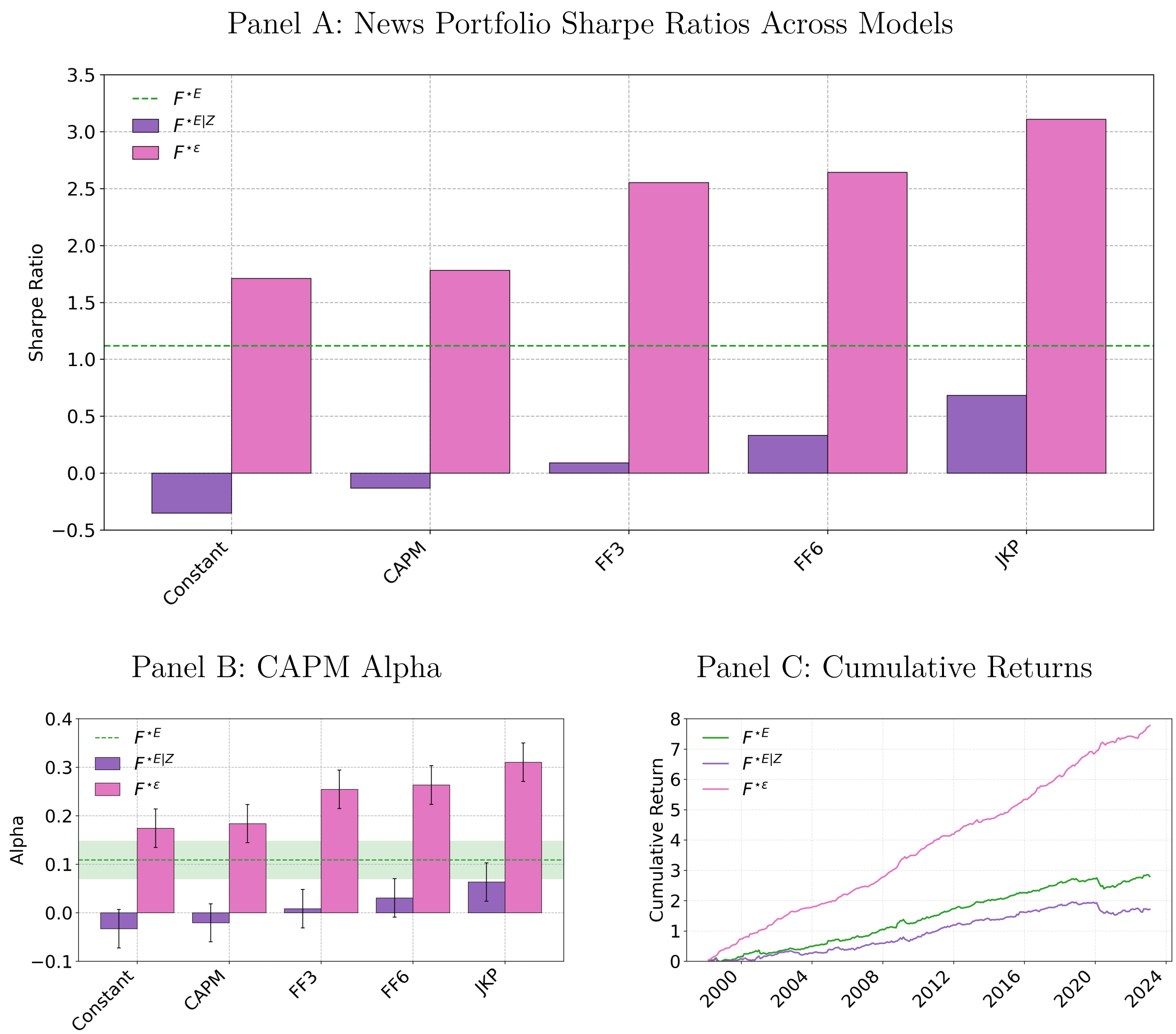


Figure 4: News Portfolio Performance

Note. Panel A reports the annualized out-of-sample Sharpe ratios for MSRR news portfolios. The dashed green line represents the raw embeddings portfolios, $F^{\star E}$. Purple bars correspond to the "predictable news" portfolios $F^{\star E|S}$ and pink bars represent news shock portfolios $F^{\star \varepsilon}$. Panel B reports the annualized CAPM alpha and its 95% confidence interval for each model. Panel C reports cumulative returns of news portfolios when JKP factors are used to residualize embeddings. All portfolios are (ex post) standardized to 10% annual volatility to aid in interpretation of alphas and cumulative returns.

How many conditioning characteristics are necessary to purge the predictable component of news and isolate news shocks? To examine this, we fix a grid of $k = [5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ characteristics, and construct S_t as a random sample of k predictors from the set of 132 JKP characteristics. We use these k signals to construct residual embeddings and repeat our MSRR procedure to arrive at a final news shock portfolio.

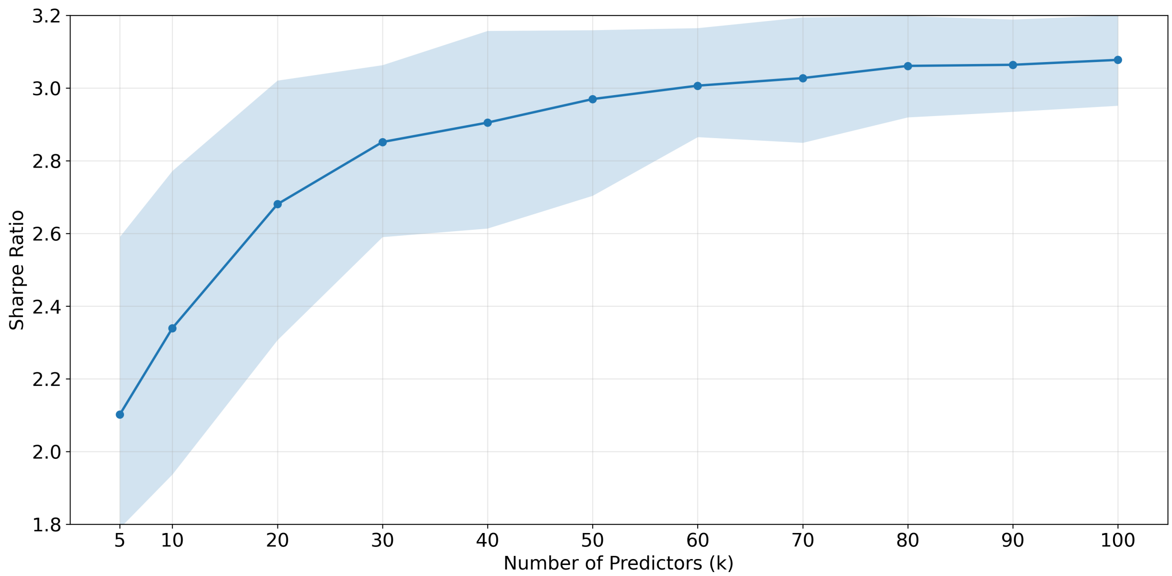


Figure 5: How Many Predictors Are Necessary to Construct News Shocks?

Note. This figure reports Sharpe ratios of MSRR news shock portfolios when news embeddings are residualized against a gradually expanding set of stock characteristics. For each k, we randomly select k characteristics from the full set of 132 characteristics, use this subset to construct news shocks, and then apply MSRR to arrive at a final news shock portfolio. To abstract from the effects of predictor ordering, for each k we repeat this analysis 100 times randomizing the set of anomaly predictors in S_t and report the average Sharpe ratio. The shaded region shows the 5^{th} to 95^{th} percentile range of Sharpe ratios across repetitions.

To abstract from the effects of predictor ordering, for each k we repeat this analysis 100 times randomizing the set of anomaly predictors in S_t.

Figure 5 reports the average Sharpe ratio of the news shock portfolio at each value of k. The last point on the curve corresponds to our main analysis above with all 132 characteristics used in our main analysis. The figure also displays 95% confidence intervals derived from the 100 repetitions. There is a monotonically increasing and concave effect to adding signals that capture the predictable content of news. The first few characteristics are extremely valuable for stripping away predictable news. By the time 50 predictors are used, most of the improvement from purging old news is realized, and after 80 predictors the marginal benefit of another predictor is close to zero. This pattern is consistent with the significant correlation among stock characteristics documented in JKP. There is relatively little variation in news shock portfolio performance as we vary the set of stock characteristics, even when $k = 5$.

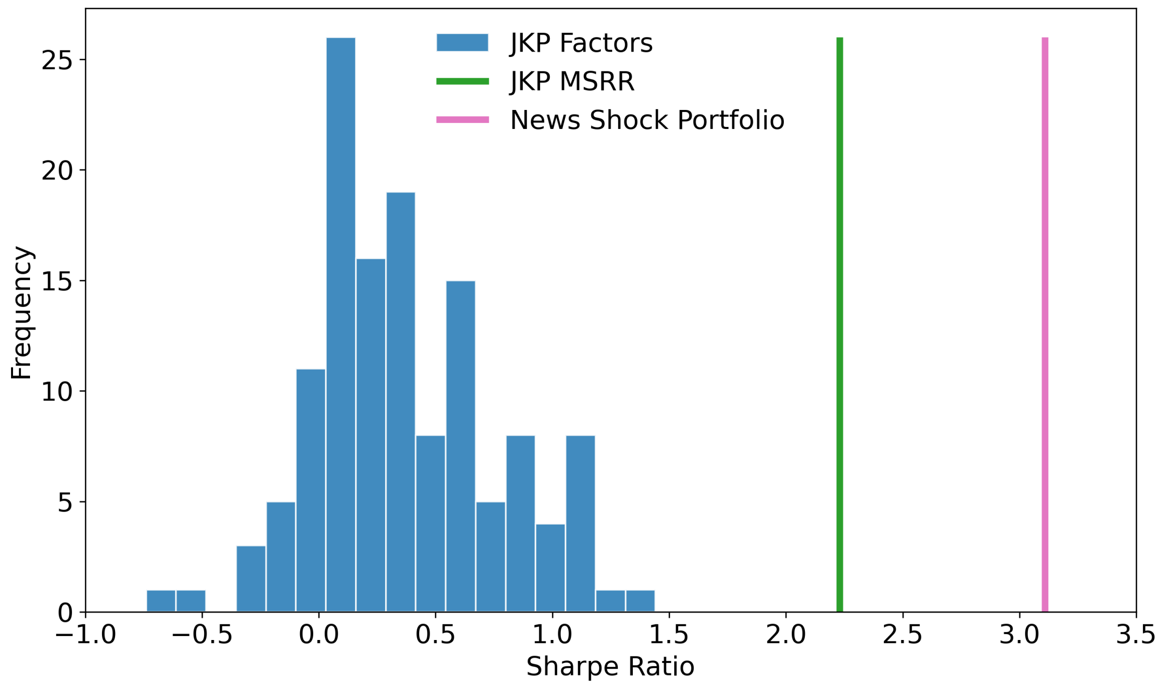


Figure 6: News Shocks and the Broader Anomaly Universe

Note. This Figure shows the Sharpe ratio distribution of individual JKP factors, the Sharpe ratio of the MSRR portfolio of JKP factors, and the Sharpe ratio of the news shock portfolio.

4.3 News Shocks In the Broader Anomaly Universe

In Figure 6 we compare the news shock anomaly to the full universe of 132 anomalies studied by JKP. As described in Section 2 the JKP characteristics are cross-sectionally rank-standardized, and the corresponding factors are defined as $S_t' R_{t+1}$. Thus the JKP factors are directly comparable to the way embeddings factors are formed within the MSRR procedure. The maximum Sharpe ratio among the JKP factor universe is 1.41,[8] roughly half that of the news shock portfolio.

MSRR can also be used to aggregate the JKP anomaly factors into a single optimal portfolio in real time, following the same methodology we use to aggregate individual embedding factors.[9] The MSRR portfolio of JKP factors is shown in the green bar in Panel

[8] This corresponds to the cash-based operating profits-to-book assets (JKP mnemonic "cop_at") factor of Ball et al. (2016).

[9] The distribution of Sharpe ratios for individual embedding coordinates—these are the 4,096 underlying embedding portfolios that are aggregated into the news portfolio via MSRR—ranges between roughly ±1.5 and is centered near zero. MSRR combines these factors via a recursive, backward-looking portfolio

A. The news shock anomaly remains larger in magnitude than the mean-variance efficient combination of all JKP anomalies taken together.

Building on Figure 6, Table 4 studies similarities and differences of the news shock anomaly versus previously documented anomalies. The table reports time series regressions of news portfolios on the 13 anomaly theme portfolios studied in JKP. Controlling for 13 anomalies is a particularly stringent alpha test,[10] while the betas allow us to assess whether any previously documented patterns lurk beneath the news shock anomaly.

The columns of Table 4 correspond to different notions of news shocks, beginning with de-meaned embeddings (“Constant") and then working with residualized embeddings from a growing set of characteristic predictors as we move to the right of the table (following the same progression reported in Figure 4). When accounting for only a constant or a constant plus beta, nearly 40% of news portfolio variation is explained by well known anomalies. The largest overlap shows up in the form of large and significant exposure to the momentum and quality themes. Despite this, the news shock anomaly has a highly significant alpha of about 15% per year (news shock portfolio is normalized 10% per year).

As more and more of the predictable content is purged from news, the R^2 versus other anomalies gradually drops and the alpha gradually rises. In the last column, only 11% of the variation in news shock anomaly returns is explained by the theme factors, and the alpha reaches 29% per year. Exposures to momentum and quality weaken and change sign, while a notable exposure to low risk and a small but statistically significant exposure to short-term reversal emerge.

4.4 Large Versus Small Stocks

Informational inefficiencies are typically most pronounced in market segments where limits to arbitrage are binding, such as small and illiquid stocks (e.g. Hou et al., 2020b). Motivated by this, we compare anomaly behavior across size groups. We restrict analysis to stocks in the JKP “mega” and “large” size categories (those above the 50^{th} percentile of the NYSE

optimization. In other words, the 3.1 Sharpe ratio for the news shock portfolio reflects that the predictive power of individual embedding coordinates can be harvested in real time to build a high performance news shock portfolio.

[10] For interested readers, Appendix Table 6 reports this analysis for other leading factor models such as the Fama and French (2015) five-factor model and the Hou et al. (2021) Q5-factor model. Compared to Table 4, these models explain a smaller fraction of performance of news shock strategies.

Table 4: News Portfolio Exposures

This table reports regressions of the news shock portfolio constructed from various news predictor sets (Constant, CAPM, FF3, FF6, and JKP) on anomaly theme portfolios from JKP. All portfolios are (ex post) standardized to have 10% annual volatility to aid interpretation of alpha and beta coefficients, and alphas are reported in annualized terms.

	Constant	CAPM	FF3	FF6	JKP
Alpha	0.146***	0.157***	0.218***	0.243***	0.294***
	(7.58)	(8.32)	(9.88)	(10.65)	(12.74)
Market	0.128*	0.146*	0.145*	0.107	0.094
	(1.69)	(1.96)	(1.66)	(1.19)	(1.03)
Accruals	0.022	0.021	0.123	0.069	0.082
	(0.32)	(0.32)	(1.54)	(0.84)	(0.98)
Debt Issuance	0.026	-0.000	0.056	0.004	0.024
	(0.26)	(-0.00)	(0.48)	(0.03)	(0.20)
Investment	0.100	0.058	-0.304	-0.143	0.110
	(0.50)	(0.30)	(-1.32)	(-0.61)	(0.46)
Low Leverage	-0.366	-0.304	0.004	0.205	0.532
	(-1.13)	(-0.96)	(0.01)	(0.53)	(1.37)
Low Risk	-0.211	0.015	0.063	0.271	0.476**
	(-1.07)	(0.08)	(0.28)	(1.15)	(2.00)
Momentum	0.241***	0.262***	0.282***	0.072	-0.053
	(2.69)	(2.99)	(2.74)	(0.68)	(-0.49)
Profit Growth	0.026	-0.012	0.008	-0.119	-0.047
	(0.36)	(-0.18)	(0.10)	(-1.40)	(-0.54)
Profitability	0.237	0.262	-0.010	0.171	0.205
	(1.21)	(1.36)	(-0.04)	(0.73)	(0.87)
Quality	0.275***	0.244**	0.222*	0.127	-0.062
	(2.67)	(2.42)	(1.87)	(1.04)	(-0.50)
Seasonality	-0.048	-0.065	0.036	0.047	0.035
	(-0.80)	(-1.09)	(0.51)	(0.65)	(0.49)
Short-Term Reversal	-0.014	-0.046	0.072	0.045	0.151**
	(-0.25)	(-0.85)	(1.13)	(0.69)	(2.26)
Size	-0.187**	-0.126	-0.013	0.068	0.007
	(-2.27)	(-1.56)	(-0.14)	(0.69)	(0.07)
Value	-0.496	-0.588*	0.194	-0.240	-0.353
	(-1.43)	(-1.73)	(0.49)	(-0.58)	(-0.85)
R^2	0.384	0.409	0.184	0.131	0.11

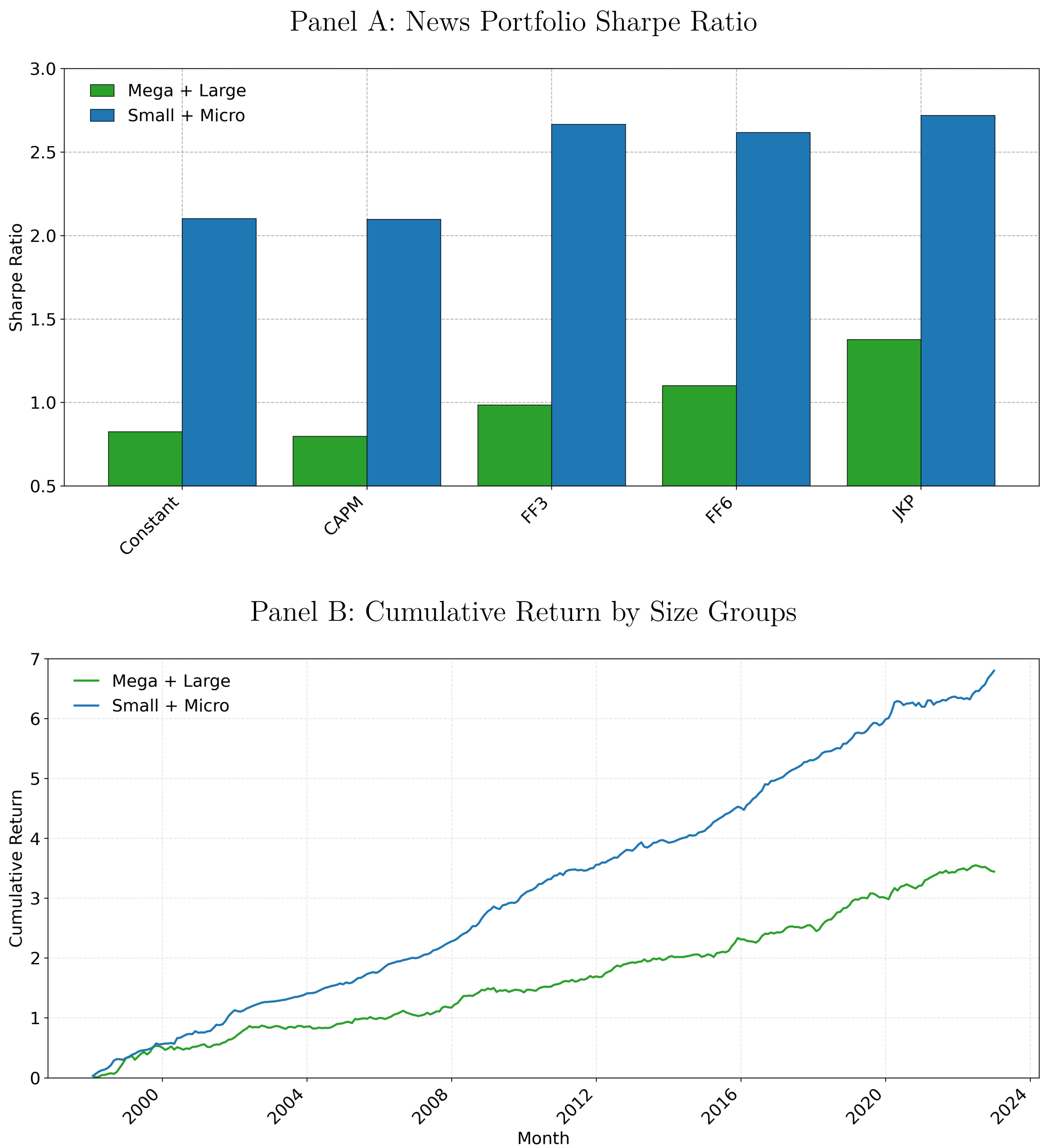


Figure 7: News Portfolio Size Group Analysis

Note. Panel A repeats the analysis of Panel A in Figure 4 but uses only stocks above/below the 50^{th} percentile of the NYSE size distribution each month (JKP "mega" and "large" / "small" and "micro" size categories) to construct the news shock factor. Panel B reports cumulative news shock returns for each size group.

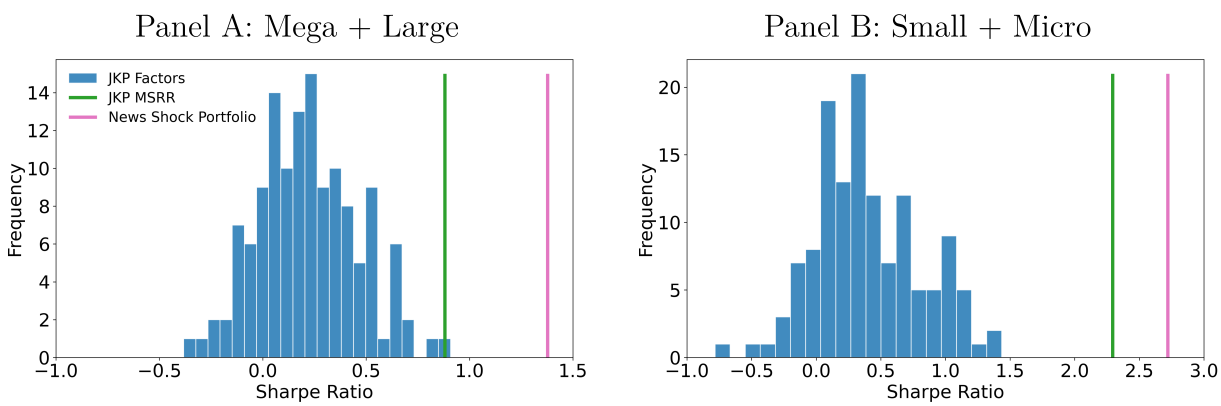


Figure 8: The Broader Anomaly Universe (By Size)

Note. This figure repeats the analysis of Figure 6 but only uses stocks above/below the 50$^{\text{th}}$ percentile of the NYSE size distribution each month (JKP "mega" and "large" / "small" and "micro" size categories) to construct JKP factors and the news shock factor.

size distribution each month) versus those in the "small" and "micro" (those below the 50th but above the 1st percentile of the NYSE size distribution).[11]

Figure 7 repeats the analysis of Figure 4 Panel A but reports performance within each size group. The Sharpe ratio of the news shock portfolio is substantially stronger for small stocks. When news is residualized versus the JKP characteristics (right-most bars), the news shock portfolio earns a Sharpe ratio of 2.8 for small stocks and 1.4 for large stocks. The pattern that news shock anomaly becomes stronger when embeddings are residualized to larger characteristics sets is preserved in both size groups. Panel B of Figure 7 plots the cumulative return of the news shock portfolio separately for large stocks and small stocks. The relative performance between the two portfolios appears stable over time and neither appears to decay much in the latter part of the sample.

Figure 8 shows that the comparison between the news shock anomaly and the broader anomaly universe is also robust across size groups. The news shock portfolio's Sharpe ratio of 1.4 for large stocks compares to 0.9 for the best performing JKP large stock factor (and versus 0.9 for the out-of-sample mean-variance portfolio of JKP factors). Among small stocks, the news shock portfolio Sharpe ratio is 2.7 versus 1.5 for the best performing JKP small stock factor and 2.3 for mean-variance portfolio of JKP small stock factors. The basic conclusion

[11]Our requirement that firms have at least one news article significantly reduces our sample and tilts it toward larger firms. We adopt this split to ensure each group has sufficient breadth for performance assessment.

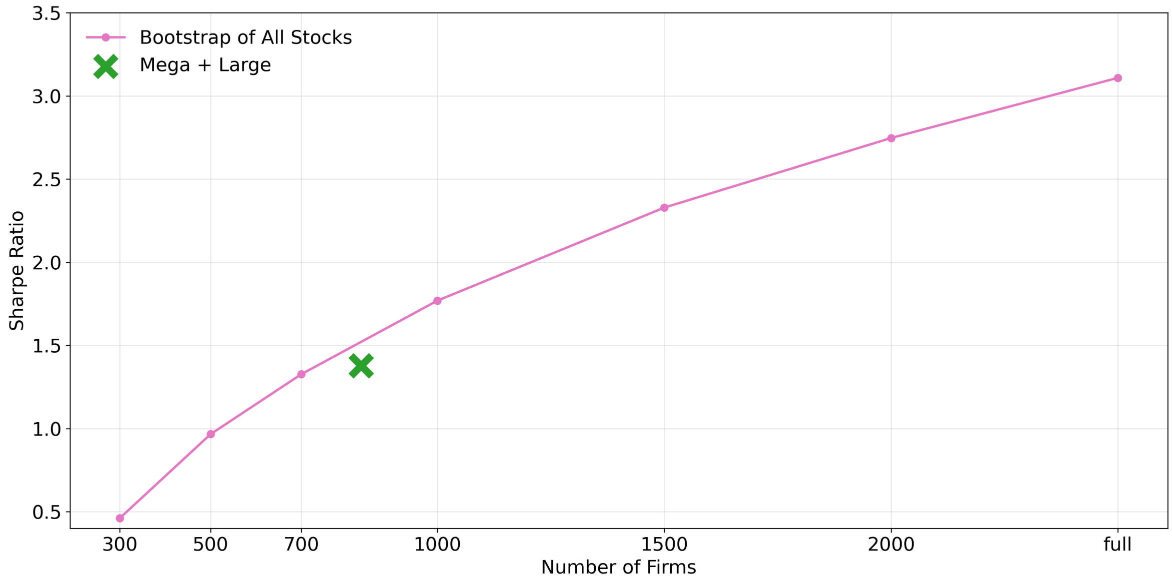


Figure 9: News Shock Anomaly and Number of Stocks in the Cross Section

Note. This figure shows the Sharpe ratio of bootstrapped news shock portfolios for different numbers of stocks N sampled from the full cross section of stocks each month. All other aspects of the portfolio construction follow the baseline specification. We repeat the bootstrap 100 times for each N and plot the average Sharpe ratio. The green point shows the news shock Sharpe ratio in the large stock subsample.

from Figure 8 is that the outperformance of news shocks relative to other anomalies is not driven by small stocks.

Interestingly, we find that the discrepancy between the Sharpe ratios of the small and large stock samples is not entirely due to differences in the strength of the anomaly, but is in large part attributable to the large stock sample having fewer stocks in the cross section. This leads to less diversification in the large stock version of the anomaly and therefore to a lower Sharpe ratio. To show this, we design a bootstrapping experiment that isolates the effect of the number of stocks in the cross section (N) while holding sample composition in terms of market capitalization fixed. First, we fix a sample size N ranging from 300 stocks to as many as 2,500 stocks (representing the average number of stocks in the cross section in our full sample). We randomly sample N firms per month from the full universe and re-construct the news shock strategy.[12] For each N we repeat this with 100 bootstrap

[12] We re-estimate all MSRR weights on each random draw of N stocks. This ensures that the reported Sharpe ratios reflect the performance achievable by an investor who observes only that sub-universe, and it avoids artificially penalizing smaller or more homogeneous universes by forcing them to use weights calibrated on the full cross-section.

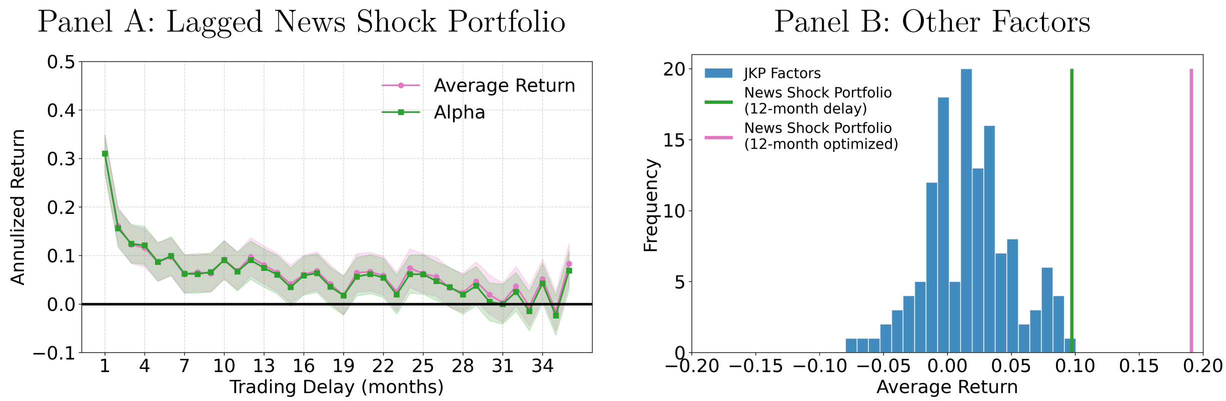


Figure 10: Decay of the News Shocks Anomaly

Note. Panel A shows annualized average returns and CAPM alpha of news shock portfolios $\varepsilon'_{t-\tau}R_{t+1}$ constructed with various trading delays $\tau = 1, \ldots, 36$ months. Panel B reports mean returns of JKP anomalies constructed with a 12-month trading delay alongside the news shock portfolio with a 12-month delay, computed using either the 1-month optimized MSRR weights or weights re-optimized specifically for forecasting 12-month-ahead returns. Shaded areas indicate 95% confidence intervals for means and alphas.

samples and report the average Sharpe ratio across bootstraps. Our bootstrap design ensures that, for any N, the composition of the sample in terms market capitalization matches (on average) that of our full sample. In this way, we isolate the impact of diversification on our subsample analysis.

Figure 9 reports the results. The performance of the bootstrapped news shock strategy shows large variation across N. When $N = 300$, there is comparatively little diversification, and the news shock Sharpe ratio is 1.1. Raising N gradually improves performance until we reach the full sample result of 3.1 for $N = 2,500$. Plotted alongside the bootstrap curve is the Sharpe ratio for the large stock subsample. On average, the subsample of large stocks has a cross section of 832 stocks. The Sharpe ratio of 1.4 when using large stocks alone is very close to the bootstrap Sharpe ratio of 1.5 that is realized when the size distribution matches the full sample but the cross section size is restricted to 832 stocks. The conclusion from this analysis is that the lower Sharpe ratio among large caps is not because large caps are more efficient in pricing news shocks. To the contrary, it appears that large and small stocks have a similar degree of news shock inefficiency, and the difference in their portfolio performance is merely an artifact of differential diversification.

4.5 Persistence, Turnover, and Net-of-cost Performance

How long does it take for news shocks to become fully incorporated in prices? Said another way, how persistent are news shock mispricings and how quickly do returns to the news shock anomaly decay? We investigate this by introducing a time delay before the news shock signal is allowed to be used in the portfolio. While our main analysis forms news shock factors as $F^{\varepsilon}_{t+1} = \varepsilon_t' R_{t+1}$, we modify these factors as $\varepsilon_{t-\tau}' R_{t+1}$ with a trading delay of $\tau = 1, ..., 36$ months.

Figure 10 shows the persistence of the news shocks anomaly. The strategy with no delays (our main specification) earns an average return of about 30% per year on a 10% volatility (essentially all of which is alpha versus the CAPM, which is also plotted). While the return predictability of news shocks drops by roughly half after one month, it takes at least a year and a half before the predictive content of news becomes insignificant. Contrast this with other well known anomalies whose levels and longevity are much less than the news shock anomaly.

Our main news shock strategy leverages unanticipated information arrival each month, which is likely to generate significant portfolio turnover. We define one-sided portfolio turnover as one-half the fraction of total positions that are replaced in a given period:

$$\text{Turnover}_t = \frac{1}{2} \sum_i \left| \tilde{w}_{i,t} - g_{i,t} \tilde{w}_{i,t-1} \right|, \tag{8}$$

where $\tilde{w}_{i,t} = w_{i,t}/G_t$, $G_t = \sum_i |w_{i,t}|$ is the total absolute position size, and $g_{i,t}$ is the return-based growth rate of the position in stock i. Panel A of Figure 11 reports the turnover of the news shock portfolio relative to turnover of JKP anomaly portfolios. The turnover of the news shock portfolio is indeed high at 75%, which exceeds that of all JKP factors (the highest turnover JKP factor is short-term reversal with turnover of 0.67). Panel B of Figure 11 reports the net-of-cost Sharpe ratio of the JKP factors and the news shock portfolio, assuming a 10 basis point trading cost per dollar traded (following Frazzini et al., 2018). Adjusting for trading costs narrows the performance gap between the news shock strategy and the JKP factors.

However, the relatively long-horizon predictability of news shocks documented in Figure 10 suggests that much of the predictability in news shocks can be leveraged with less turnover by averaging embeddings over longer lookback windows. In our main specification, embeddings are averaged over articles from the most recent month. We now consider the

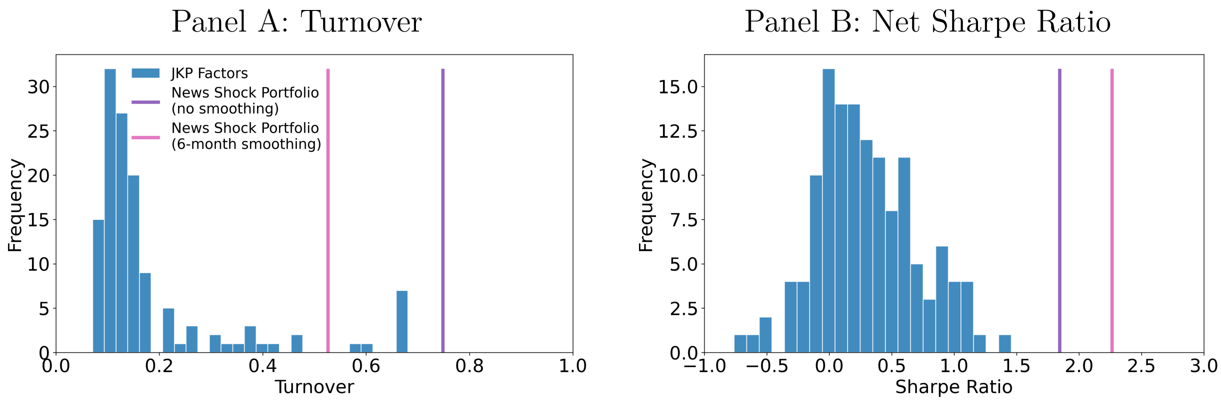


Figure 11: Turnover and Net Sharpe Ratio of the News Shock Anomaly

Note. Panel A reports the one-sided turnover of the JKP factors, the news shock factor using one-month average embeddings, and the news shock factor using average embeddings over the most recent six months. Panel B reports the net Sharpe ratio of each portfolio assuming a 10 basis point trading cost per dollar traded.

effect of averaging article embeddings over the most recent $j = 1, ..., 24$ months in order to smooth out the signal and reduce turnover.

Panel A of Figure 12 shows how the strategy's Sharpe ratio and CAPM alpha are affected by aggregating embeddings in the past $j = 1, ..., 24$ months. The Sharpe ratio remains near 3.0 for lookback windows of up to 6 months. When embeddings are averaged over 24 months the Sharpe ratio is 2.4—smaller but still large relative to the distribution of JKP anomalies. Panel B of Figure 12 shows how turnover of the news shock strategy drops when embeddings are averaged over longer lookback windows. Panel B also plots the net Sharpe ratio and shows that the optimal tradeoff between news timeliness and trading costs occurs when embeddings are averaged over the most recent 6 months. The version of the news shock portfolio based on 6-month average embeddings is also shown in Figure 11 to compare it with the broader anomaly universe. The basic conclusion from Figures 11 and 12 is that the news shock anomaly is not an artifact of unrealistic trading costs. It can be implemented with similar trading costs as many anomalies in the literature by averaging embeddings over longer lookback windows and its net performance still exceeds that of all JKP anomalies.

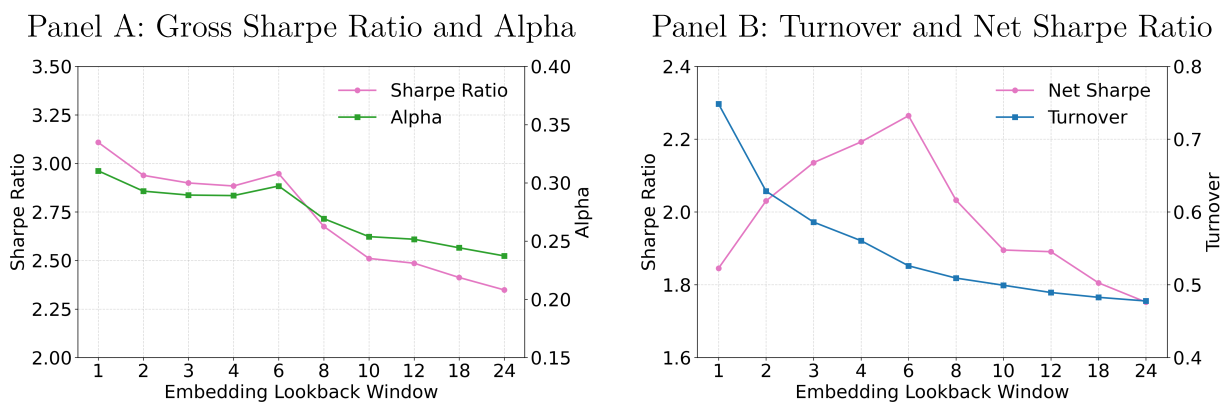


Figure 12: Rolling Average Embeddings

Note. Panel A reports the news shock portfolio gross Sharpe ratio and CAPM alpha when embeddings are averaged over the most recent $1, \ldots, 24$ months. Panel B reports the corresponding portfolio turnover and net Sharpe ratio.

5 Interpreting the News Shock Anomaly

5.1 The Embeddings Interpretation Problem and the SAE Solution

Embeddings strive to accurately represent a vast number of concepts (all concepts conceivable with human language) in a relatively low-dimensional vector space. Typical embeddings of dimension D (say a few thousand) thus tend to be dense, as all D coordinates in the small vector space must work together to convey the much larger variety of distinct concepts. Concepts therefore cannot be traced to individual coordinates but are instead distributed across thousands of dimensions in a complex, overlapping pattern (commonly described as "superposition," Elhage et al., 2022). This "polysemantic" property of embeddings is part and parcel of the highly efficient compression achieved in an LLM, the benefits of which are exemplified by our findings above: Compressing rich information in news text into a 4,096-dimensional vector makes it possible to build powerful return prediction models with a reasonably parsimonious regression. The downside of efficient compression is that polysemantic embeddings are, to the human eye, uninterpretable.

It would be easier to interpret embeddings if they were somehow sparse; that is, if just one or a few coordinates corresponded to distinct concepts. This is possible by decompressing the embeddings into a higher-dimension vector space. Decompression disentangles the small number of mixed, polysemantic embedding coordinates into a much larger number of disjoint, monosemantic coordinates. Imagine this process as a prism that splits a single mixed ray

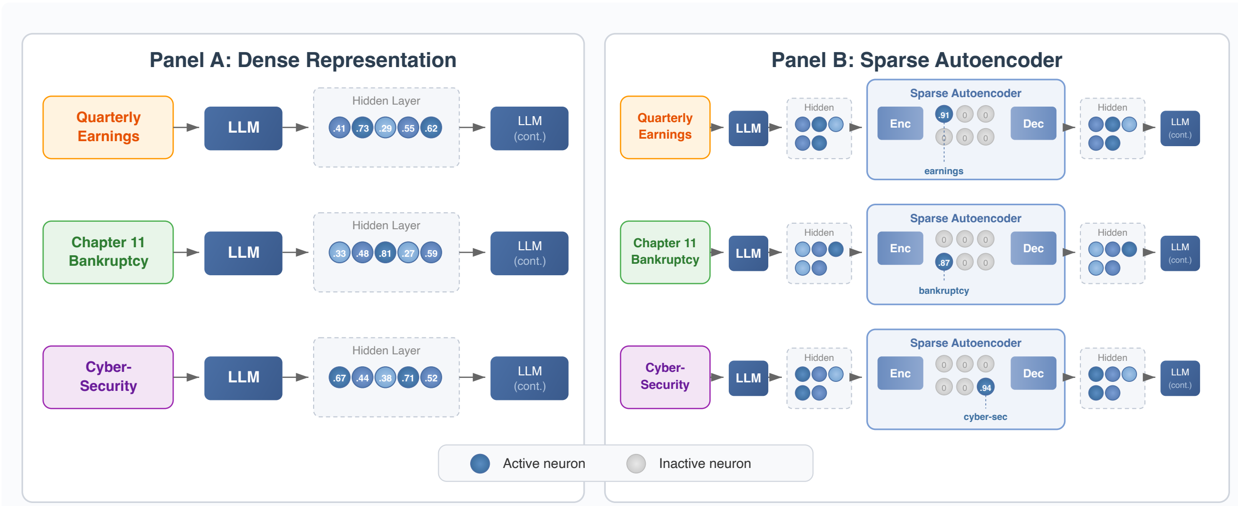


Figure 13: Comparison of Dense vs. Sparse Autoencoder Representations. Panel A illustrates a standard dense representation in which concepts are entangled across dimensions. Panel B illustrates the SAE representation, where narratives (e.g., Quarterly Earnings, Chapter 11 Bankruptcy, Cyber-security) correspond to sparse and more localized patterns of latent feature activation.

of light into its constituent spectral colors. With sparse, monosemantic embeddings it is possible to inspect which embedding coordinates are activated by distinct concepts, thereby assigning an interpretation to individual coordinates.

We construct interpretable embeddings using an LLM sparse autoencoder (or "SAE," Cunningham et al., 2023; Bricken et al., 2023). In most architectures, the dense embedding is extracted from the final hidden layer of the LLM. An SAE is an interpretability layer appended to the LLM that receives the dense embedding and outputs a much larger sparse embedding. The SAE is trained to retain as much information as possible from the dense embedding but is constrained by a lasso penalty that forces most coordinates in the new layer to be zero most of the time.

Figure 13 illustrates this schematically. The original dense embedding is shown as the "hidden layer" in Panel A, with semantic content broadly distributed across its coordinates. In Panel B the SAE is inserted as a new and larger sparse layer whose coordinates specialize in individual concepts. This layer is structured as an autoencoder that takes the original dense embedding as both input and output, illustrating that the goal of the SAE is to preserve information. But the SAE's encoder and decoder parameters are trained to achieve sparsity (and thus interpretability) in the new intermediate layer. The machine learning literature demonstrates the efficacy of SAEs for interpreting LLMs (see the survey of Shu et al., 2025, and references therein).

5.2 From Sparse Embeddings To Interpretation

Successfully disentangling polysemantic embeddings requires an extremely high-dimensional sparse layer. For this purpose, we use a pre-trained SAE with an embedding dimension of 131,000 (from the 9 billion parameters Gemma2 model of Lieberum et al., 2024). From a practical perspective, the SAE can be treated as just another LLM that ingests raw articles and outputs (very large) embedding vectors.

Rather than studying all 131,000 SAE embedding coordinates, we focus on a subset of the most "financially important" coordinates identified by Chen et al. (2025).[13] The Chen et al. (2025) coordinate selection is based on a full sample return prediction objective. At first blush, one may wonder if this introduces some kind of look-ahead bias in our analysis. To the contrary, our objective in this section is an interpretable description of the forces underlying the news shock anomaly. We use the selected feature set to decompose already-established predictive performance into interpretable components, rather than to assess out-of-sample model performance.

We assign interpretable labels to each of the 5,000 SAE embedding coordinates with the following procedure. First, we select the top 100 articles exhibiting the highest activation values for each coordinate. Next, we prompt an LLM with those articles and with instructions to generate a short, descriptive label that summarizes their common theme. Finally, we manually audit the generated labels to verify that they align with the content of top articles.[14]

We focus our exposition on the coordinates that are most important for understanding the news shock anomaly. To this end, we identify the SAE embedding coordinates that are most prominently represented in the news shock portfolio. To re-construct the news shock anomaly, interpretable sparse embeddings (denoted $\tilde{E}_t$) can be used in exactly the same manner as the original dense embeddings (E_t) from our main analysis. We orthogonalize $\tilde{E}_t$ with respect to the JKP factors to construct news shocks, denoted $\tilde{\epsilon}_t$. Then, we use news shocks to form interpretable news-managed portfolios $F^{\tilde{\varepsilon}} = \tilde{\varepsilon}_t' R_{t+1}$. Finally, we aggregate these news-managed portfolios into a single news shock portfolio $F^{\star\tilde{\varepsilon}}$ via MSRR. We train MSRR with a lasso penalty in order to identify elements of $F^{\tilde{\varepsilon}}$ that are the most important

[13]The vast majority of Lieberum et al. (2024) SAE embedding coordinates are non-financial in nature.

[14]The full prompt, including specific system constraints and output formatting requirements used to generate labels, is detailed in Figure 25 of Appendix B.

drivers of the news shock anomaly. We tune the lasso penalty parameter to select exactly 30 interpretable coordinates with non-zero MSRR weights.

5.3 News Themes Underlying the News Shock Anomaly

By the end of the sample, the union of interpretable news-managed portfolios that receive positive weight in the news shock portfolio amounts to 144 unique elements of $F^{\tilde{\varepsilon}}$. We focus our interpretation on these 144 SAE embedding coordinates. The interpretable labels for every coordinate are listed in Table 7 of Appendix B. Examples of coordinate labels include concepts such as "Chapter 11 bankruptcy and going-concern distress," "Executive appointments and leadership changes," and "Cryptocurrency regulatory scrutiny and adoption."

To achieve a manageable dimension for exposition, we further cluster these 144 interpretable coordinates into 12 broad economic themes: "Earnings & Financial Results;" "Corporate Guidance & Outlook;" "Analyst Ratings & Sentiment;" "Distress, Bankruptcy & Delisting;" "Momentum & Trading Activity;" "Corporate Actions & Restructuring;" "Leadership & Governance;" "Growth & Demand Trends;" "Biotech, Pharma & Healthcare;" "Regulatory & Legal Actions;" "Sector-Specific Signals;" and "Product Launches & Operations." To illustrate the semantic coherence of the 12 themes, Figure 14 displays word clouds constructed from the labels of coordinates within each theme.

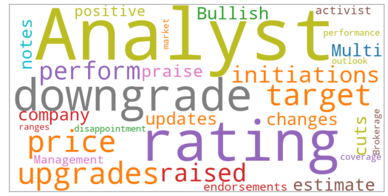
Next, we quantify the economic contribution of these 12 themes (and in turn the 144 interpretable embedding coordinates) to the news shock anomaly by studying the MSRR weights over the 144 corresponding news-managed portfolios. Theme importance is calculated as the sum of absolute portfolio weights allocated to each theme on a monthly basis. Figure 15 illustrates how the news shock portfolio allocation to themes evolves over time. News about "Corporate Actions & Restructuring" remains a fixture of the news shock anomaly throughout our sample. Early in the sample, news about "Momentum & Trading Activity" constitutes a large share of the anomaly early in the sample but dies off after the tech bubble in the early 2000's. In contrast, "Corporate Guidance & Outlook" has little role in the anomaly early on but steadily grows to become the most important theme by the end of the sample.

While Figures 14 and 15 provide a top-down perspective on news themes that are important sources of news shock returns, Figure 16 provides a bottom-up perspective with a few detailed examples of interpretable news-managed portfolios. It presents five SAE coordinates that capture either long-term economic shifts or short-term event-driven shocks.

Earnings &
Financial Results

Corporate Guidance
& Outlook

Analyst Ratings
& Sentiment



Distress, Bankruptcy
& Delisting

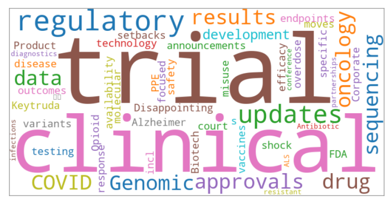
Momentum &
Trading Activity

Corporate Actions
& Restructuring

Leadership &
Governance

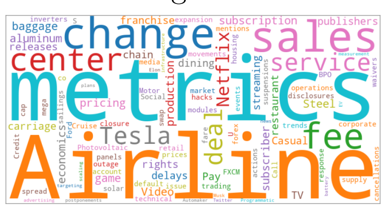
Growth &
Demand Trends

Biotech, Pharma
& Healthcare



Regulatory &
Legal Actions

Sector-Specific
Signals



Product Launches
& Operations

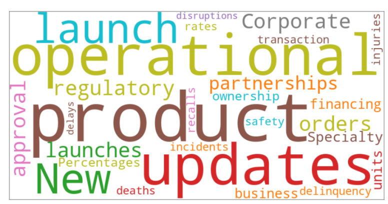


Figure 14: MSRR Factor Theme Word Clouds

Note. Each panel displays a word cloud of the factor names assigned to the corresponding theme. Word size is proportional to frequency. The 144 classified factors are grouped into 12 themes based on the interpretable labels from the SAE dictionary. Word clouds are generated from the human-readable factor labels within each theme. Larger words indicate factors whose name tokens appear more frequently within the theme grouping.

We report the interpretable labels of each coordinate and the cumulative return of the news-managed portfolio (corresponding to $F_t^{\tilde{\varepsilon}}$) for each coordinate.

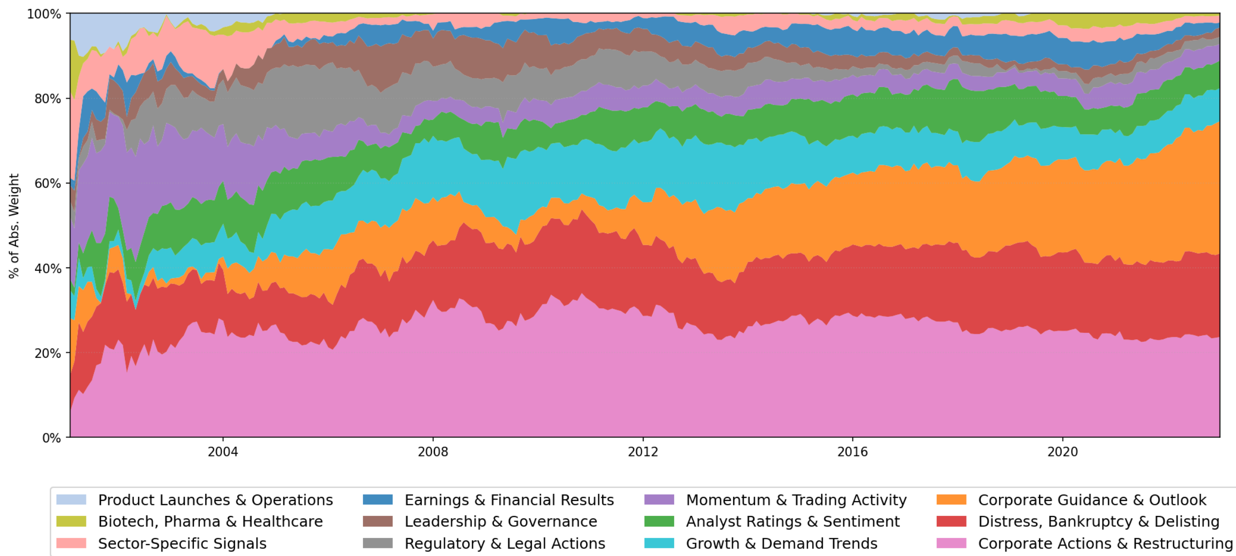


Figure 15: Dynamic Theme Allocation in the News Shock Portfolio

Note. Stacked area chart showing the relative importance of the 12 interpretable economic themes over time. The vertical axis represents the sum of absolute portfolio weights for each theme as a percentage of the total portfolio weight.

Panel A focuses on three macroeconomic coordinates: "E-commerce expansion", "Systemic crisis contagion", and "Corporate COVID response". The "E-commerce expansion" portfolio rises and falls with the Dot-com bubble. The "Corporate COVID response" lies dormant for the majority of the sample and spikes at the onset of the pandemic in 2020. "Systemic crisis contagion" has a consistent negative slope and posts sharp declines during recessions.

Panels B and C illustrate how our news portfolios can capture market movements associated with punctuated events. Panel B displays the "Meme stock short squeeze" portfolio, which is dominated by a discrete jump coinciding with this historic short-squeeze (as a frame of reference, the secondary axis in Panel B tracks the share price of GameStop). Similarly, Panel C plots the "Crypto" news portfolio alongside the price of Bitcoin. Although our interpretable factors trade solely in stocks based on firm-level news coverage, the cumulative return of the "crypto" factor appears to slightly presage the Bitcoin run-up of 2021. We present further detail for these examples in Appendix B, where Table 8 reports specific headlines tied to the meme stock and crypto portfolios. The composition of the crypto portfolio is dominated by six firms that account for an average of 50% of its absolute weights, including Riot Blockchain, Silvergate Capital, MicroStrategy, International Money

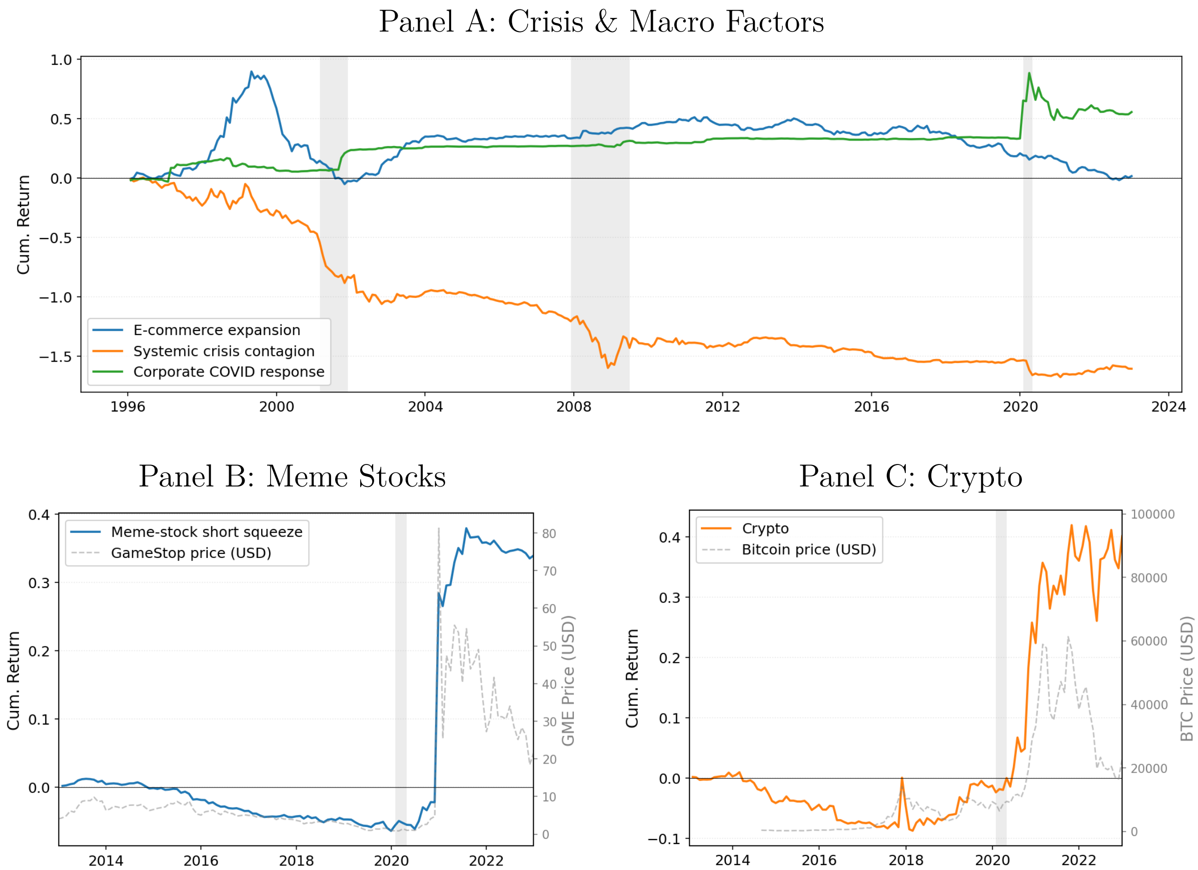


Figure 16: Evolution of Individual Interpretable Features

Note. This figure presents the cumulative returns of five select sparse features, standardized to 10% annualized volatility. Panel A displays three macro-oriented features ("E-commerce expansion," "Systemic crisis contagion," and "Corporate COVID response") overlaid with grey shaded areas indicating NBER recessions. Panel B plots the "Meme Stock" feature alongside the GameStop share price (right axis). Panel C plots the "Crypto-correlated" feature alongside the Bitcoin price (right axis).

Express, WSFS Financial, and MercadoLibre, all of whom have business models tightly linked to the cryptocurrency ecosystem.

5.4 Underreaction and Overreaction to News

Section 5.3 describes the thematic content of news underlying the news shock anomaly. But it does not explain how or why this news translates into a profitable trading strategy. In this section, we investigate the nature of return predictability associated with different news topics to derive a more detailed interpretation of the news shock anomaly. Inspired by the frequent organization of behavioral finance literature into underreaction or overreaction phenomena,[15] we use our interpretable SAE embeddings to isolate news topics that are associated with price continuation versus those that demonstrate price reversal, and in turn to understand the most profitable forms of predictability that underly the news shock anomaly.

We define contemporaneous (and hence non-tradeable) news-managed portfolios as

$$F^{\tilde{\epsilon}}_{t,t} = \tilde{\epsilon}_t' R_t, \tag{9}$$

which capture the immediate price impact of the news shock. We then define the tradeable post-news return as

$$F^{\tilde{\epsilon}}_{t,t+1} = \tilde{\epsilon}_t' R_{t+1}, \tag{10}$$

which measures subsequent price response to news shocks. The tradable returns use the same news-managed portfolio construction from our main analysis, with additional subscripts to draw a clear distinction between $F^{\tilde{\epsilon}}_{t,t}$ and $F^{\tilde{\epsilon}}_{t,t+1}$.

For each SAE embedding coordinate c, we measure the correlation between the contemporaneous return and return in the subsequent month

$$\rho(c) = \text{corr}\left(F^{\tilde{\epsilon}}_{t,t}(c), F^{\tilde{\epsilon}}_{t,t+1}(c)\right).$$

When $\rho(c) > 0$ the subsequent return continues in the same direction as the initial response, thus we infer that the market underreacts to news about topic c. When $\rho(c) < 0$ the market walks back its initial price response indicating overreaction to news about the topic.

Figure 17 illustrates the returns patterns of the five SAE embedding coordinates that exhibit significant underreaction (those with the large positive values of $\rho(c)$, shown in Panel A) and the five coordinates that exhibit significant overreaction (Panel B). We plot the Sharpe ratios of the contemporaneous return $F^{\tilde{\epsilon}}_{t,t}$ in gray and the subsequent return $F^{\tilde{\epsilon}}_{t,t+1}$ in blue.

Markets demonstrate significant underreaction to news about “corporate financing,” “bankruptcy and going-concern risk,” “enforcements and delistings,” “financial restatements” and “product launch and approval,” which are economically complex topics that require careful reasoning through legal and regulatory details.

On the other hand, markets overreact to news about “deal announcements,” “pre-market/intraday moves,” “analyst ratings,” “small-cap stock surges,” and “SPAC reverse

[15] For example, Daniel et al. (1998), Hong and Stein (1999), chapter 5 of Shleifer (2000), Hirshleifer (2015), and Ba et al. (2024), among many others.

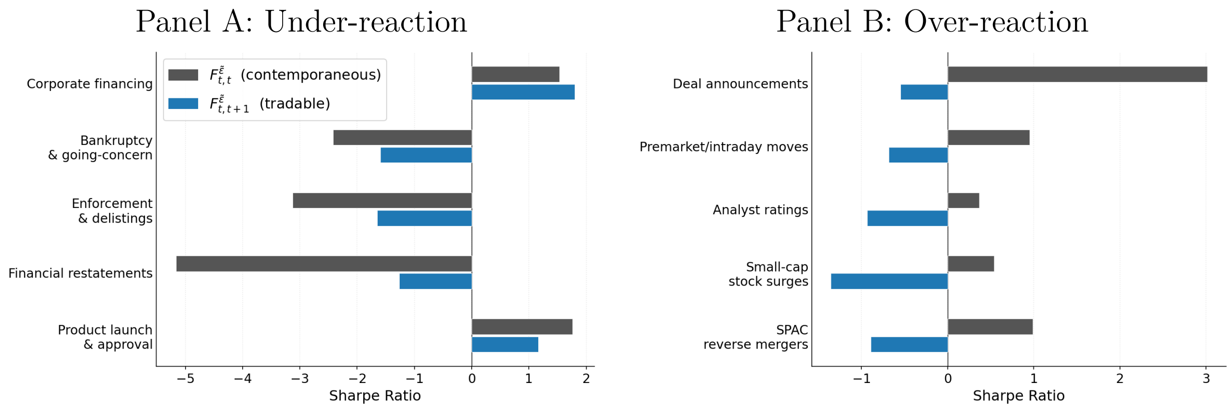


Figure 17: Underreaction and Overreaction by News Topic

Note. This figure reports Sharpe ratios of the infeasible contemporaneous news portfolio $F^{\tilde{\epsilon}}_{t,t}(c)$ and the feasible portfolio $F^{\tilde{\epsilon}}_{t,t+1}(c)$ for select SAE embedding coordinates, c. Panel A reports the coordinates and their interpretable labels associated with the most pronounced market underreaction and Panel B reports coordinates exhibiting the strongest overreaction.

merges." Deal announcements, analyst ratings, and large price swing narratives represent high-salience information that falls into "extreme" categories defined by Kwon and Tang (2025). Consistent with their premise that attention-grabbing signals trigger extrapolative expectations, investors appear to initially push prices too far in response to these news topics and later correct course.

Our findings suggest that market efficiency is structurally dependent on information type: procedural complexity interacts with inattention to induce slow diffusion (underreaction), while high salience and volatility trigger extrapolation heuristics (overreaction). This dynamic is rationalized by the cognitive constraint model of Ba et al. (2024).

In the universe of 5,000 interpretable SAE coordinates, the distribution of reaction types ($\rho(c)$) is slightly dominated by overreactions with 41.7% of news topics classified as overreaction. However, in a portfolio return sense, the MSRR procedure finds that underreaction is by far the more profitable form of inefficiency in the data. Among the subset of 144 news-managed factors with non-zero MSRR weight, 66.9% of these correspond to underreaction topics. When we look at the detailed weight allocation, this skew becomes even more pronounced, with 84.6% of the strategy's total weight assigned to underreaction topics. In short, the news shock anomaly is almost entirely dedicated to exploiting the market's underreaction to news.

6 LLM Lookahead Bias and Other Robustness

6.1 Chronologically Consistent LLMs

Due to their extremely large number of parameters, LLMs have a tendency to "memorize" data that they were trained on (Carlini et al., 2022). When using LLMs output for financial time series modeling, there is potential for look-ahead bias when the LLM has been trained on data that was not available at the time when a so-called out-of-sample financial model forecast is being produced (see, e.g. Glasserman and Lin, 2023; Sarkar and Vafa, 2024).

Perhaps the most direct way to ensure that our results are not being driven by look-ahead bias in the pre-trained Mistral model is to replicate our analysis using a sequence of LLMs that are trained only on data that would have been available to the forecaster in real-time. This idea lies behind the "chronologically consistent" LLMs (or CCLLMs) developed by He et al. (2025). CCLLMs are trained on carefully curated datasets where the training corpus is restricted to information available up to specific timestamps.

We design a dual experiment to isolate the impact of look-ahead bias. First, we construct a clean "point-in-time" specification where, for each month t, we generate embeddings using the CCLLM[16] trained only on data available up to t, then use these embeddings to construct MSRR news portfolios. In parallel, we also construct a biased "foresight" counterfactual portfolio where we generate embeddings for the entire sample using a single CCLLM that is trained on all data up to 2022, thereby voluntarily introducing look-ahead bias. The only difference between the point-in-time and foresight models is the training data set; their architectures are identical. If LLM look-ahead bias is a major issue, then we should see the point-in-time model significantly underperform the foresight model.

Figure 18 compares the performance of the CCLLM point-in-time model versus the same model trained with foresight. Two key facts emerge from this analysis. First, and most importantly, the performance is virtually identical whether we use point-in-time or foresight models, indicating that look-ahead bias is not a driver of portfolio performance. In fact, in the case of JKP-based news residuals (the right-most bars in the figure), the foresight model has slightly worse performance (Sharpe ratio of 1.61) than the point-in-time model (1.63).

The literature documenting LLM look-ahead bias focuses on biases in *text generation* (e.g. Sarkar and Vafa, 2024). To the best of our knowledge, it has not been shown that applications using the "embeddings for downstream regression" workflow are subject to the

[16] In particular, we use the point-in-time version of the 1.5 billion parameter GPT2 model trained by He et al. (2025), which they refer to as "ChronoGPT".

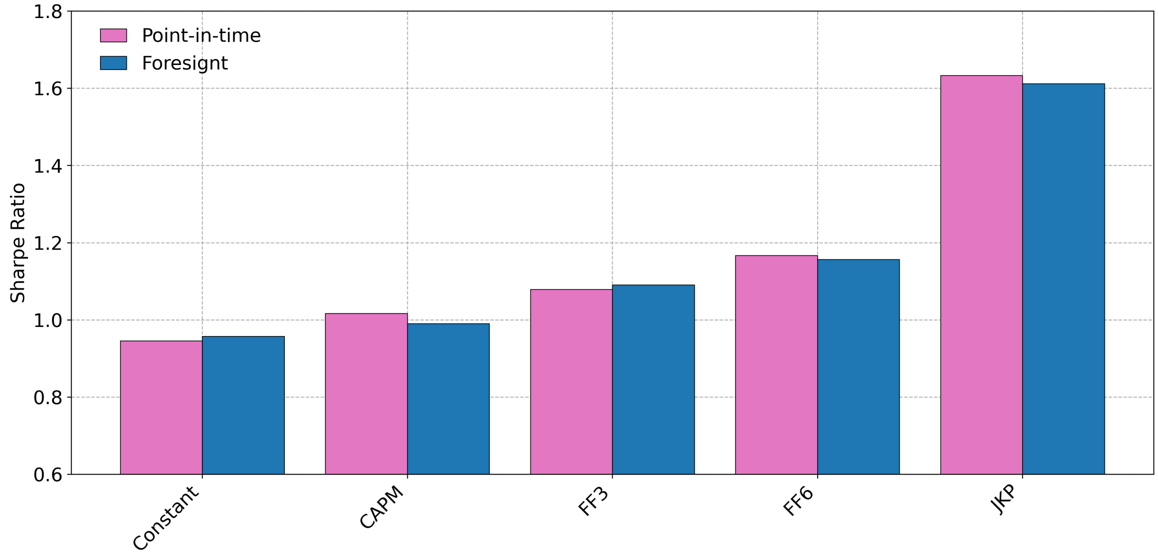


Figure 18: News Portfolio Performance with Chronologically Consistent LLMs

Note. This figure reports Sharpe ratios of news shock portfolios constructed using ChronoGPT embeddings from He et al. (2025). The pink bars ("point-in-time") use embeddings generated by models with expanding training windows that exclude future data. The blue bars ("foresight") use embeddings generated by a single ChronoGPT model trained on data up to 2022 and applied to the full historical sample.

same look-ahead bias. Indeed, He et al. (2025) note the "somewhat surprising finding is that look-ahead bias appears to be modest in this [news-embedding-based] stock return forecasting application." Consistent with He et al. (2025), Figure 18 suggests that look-ahead is likely to be a minor concern in our setting. This is because, in the "embeddings for downstream regression" workflow, the downstream objective (portfolio performance/return prediction) is decoupled from the LLMs' training objective (token prediction). Look-ahead bias is ultimately a form of small-sample overfit. It has a larger impact on tasks like text generation that derive directly from the LLM's training objective, than on a downstream task like return prediction that has no influence on the LLM's training (this logic is formalized by Wolpert, 1992).

The second main fact from Figure 18 is that replacing the industrial-scale Mistral model with the smaller academic CCLLM model leads to a decline in out-of-sample Sharpe ratio from 3.1 to 1.6 (in the case of JKP residualization). This is true even when the CCLLM is allowed "foresight." While the performance of the CCLLM-based news portfolio nonetheless exceeds the best performing anomalies in the prior literature, it is important to recognize that this attenuated performance is a conservative lower bound on the news shock anomaly.

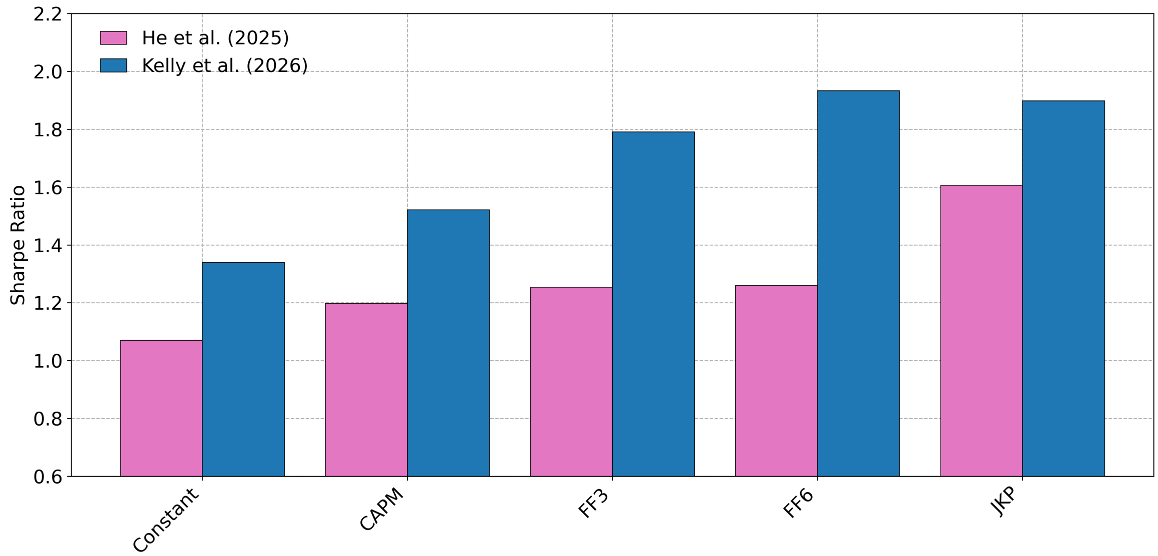


Figure 19: Comparison of Point-in-Time GPT-1.5B

Note. This figure shows the Sharpe ratios of the chronologically consistent GPT-2 model of He et al. (2025) and an identical model architecture trained on a larger dataset by Kelly et al. (2026). The sample period for this figure is 2014-2022.

This is simply due to the fact that Mistral is a far more sophisticated language model than the point-in-time CCLLM. It is well accepted that LLM behavior is accurately described by "scaling laws" (Kaplan et al., 2020; Hoffmann et al., 2022). In particular, language model performance follows a power law that increases with the number of parameters, the number of training observations, and the amount of compute used for training. In terms of parameter scale, the CCLLM is the 1.5 billion parameter GPT2 model, while the Mistral model uses 7 billion parameters. In terms of training data scale, the initial vintage CCLLM is trained on 71 billion tokens; in contrast, the Mistral model is trained on seven trillion tokens (then fine-tuned on another 1.8 million tokens). In terms of compute scale, compute details for the training of Mistral are undisclosed, but it is well understood that a few private institutions like Mistral possess computing resources far in excess of typical academic research teams (this is the so-called "compute divide" emphasized by Ahmed and Wahed, 2020).

Recent developments in the literature show that the chronologically consistent LLMs of He et al. (2025) can be improved on by training with more tokens and by training models with more parameters. For example, Kelly et al. (2026) retrain the same ChronoGPT (GPT-2) model of He et al. (2025), using approximately 140 times more training tokens, scaling the model from 1.5B to 4B parameters, and extending the context length from 1792 tokens to 2048 tokens. Kelly et al. (2026) show that changes in training alone improve point-in-time model performance by 30.6 % on standard LLM benchmarks such as HellaSwag (Zellers et al., 2019). Here we show that improved CCLLM training also gives a more positive read on real-time return predictive power of news shocks, as demonstrated in Figure 19. In particular, the more extensively trained point-in-time CCLLM improves the news shock portfolio Sharpe ratio from 1.6 (with the He et al., 2025) to 1.9 with the Kelly et al. (2026) model. Presumably, inferences about the point-in-time news shock anomaly can be improved even further when the improved model training of Kelly et al. (2026) is combined with larger LLM architectures.

6.2 Other Embedding Models

In our baseline specification, we employ the *E5-Mistral-7B* embedding model. We select this architecture for three reasons. First, it consistently achieves state-of-the-art results on common LLM benchmark tasks (Wang et al., 2024). Second, it is an open-weight model, so we perform data computations privately on our own hardware and maintain privacy of the news data per our licensing agreement (this is in contrast to "closed" models such as GPT-5 and Gemini that typically require that data be transferred to OpenAI or Google servers). Third, with only 7 billion parameters, it is sufficiently compact to run on a single A100 GPU, facilitating large-scale experimentation.

Naturally, there are alternatives to Mistral that we may consider. CKX show that their results are fairly similar across a variety of state-of-the-art LLMs. In this section, we examine the sensitivity of our results to the choice of LLM using the Llama3 family of open-weight models developed by Meta (Grattafiori et al., 2024). Unlike *E5-Mistral-7B*, which is explicitly fine-tuned for embedding tasks, the Llama3 models are generative LLMs trained to predict the probability distribution of next tokens. Nevertheless, following the methodology of Chen et al. (2024), we extract high-quality embeddings from these models by averaging the final-layer hidden states across all tokens in a sequence. A key advantage of the Llama3 suite is the simultaneous release of three models differing primarily in parameter count with either 8 billion, 70 billion, or 405 billion parameters.[17] We also consider older and smaller models such as BERT with 110 million parameters and GPT2 with 1.5 billion parameters (this is the OpenAI pre-trained version of GPT2 and not the chronologically trained model discussed

[17] Given the substantial memory requirements of Llama3-405B, we employ 4-bit (int4) quantization to ensure that model inference remains computationally feasible within our available resources.

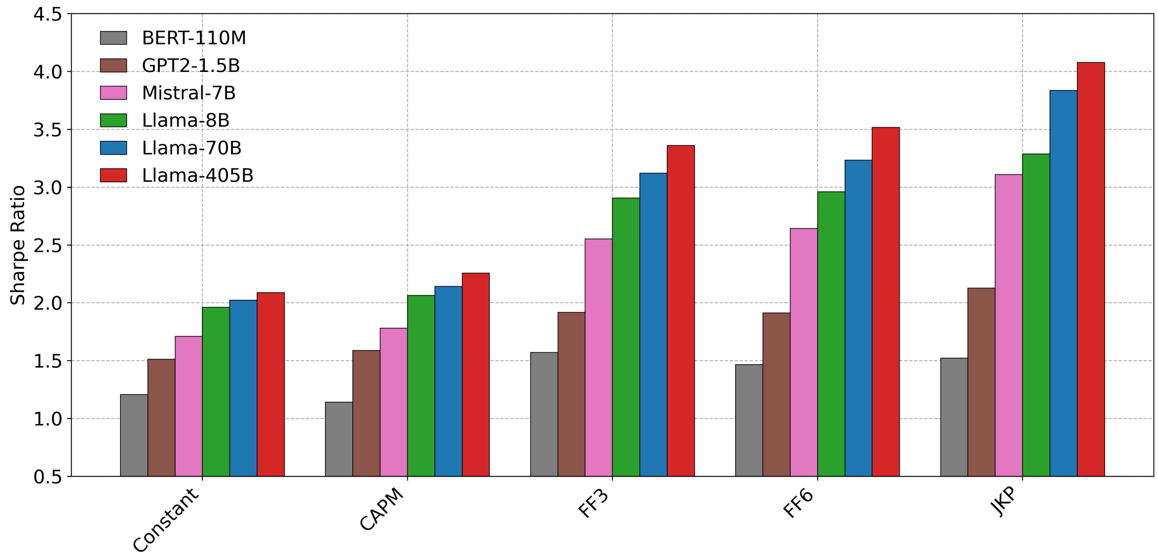


Figure 20: News Shock Portfolios From Various Embedding Models

Note. This figure reports the Sharpe ratios of news shock portfolios constructed using alternative embedding models. The smallest model is BERT-110M and the largest model is Llama3-405B. Mistral-7B is the model used in the main analysis.

above). This setup provides a controlled setting to isolate the effect of model size on the performance of news shock portfolios.

Figure 20 reports Sharpe ratios for the news shock portfolios derived from each LLM. The results show a monotonic relationship between model size and investment performance. As we increase parameter count from the 110 million parameter BERT model to the 405 billion Llama3 model, the Sharpe ratios rise. Focusing on the full JKP specification, the Sharpe ratio rises from 1.5 with BERT to 3.1 for our baseline Mistral model, to 3.3 for Llama3-8B, then to 3.8 for Llama3-70B, and eventually to 4.1 for the massive Llama3-405B model.

6.3 Other News Text Data Sources

Our main analysis uses news text from the Reuters news service. In this section we investigate the robustness of our findings to using other sources of news. The first is the Dow Jones Newswires, analyzed previously by Ke et al. (2019). This sample spans the period 1996–

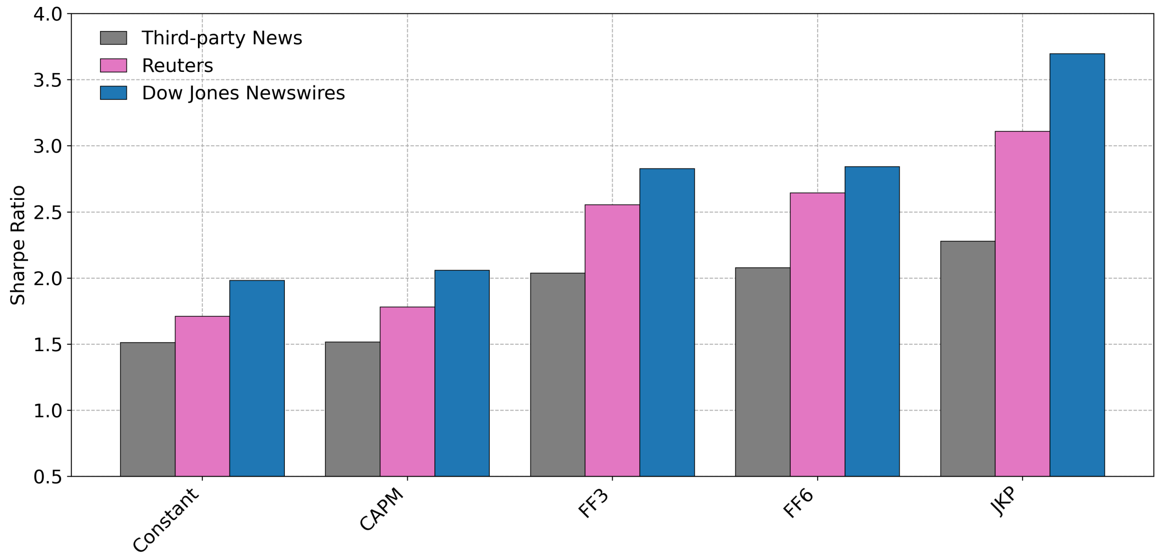


Figure 21: News Shock Portfolios From Alternative Data Sources

Note. This figure reports Sharpe ratios of news shock portfolios constructed using alternative news data sources: Reuters (main analysis), Third-party News, and Dow Jones Newswires. We consider the same embedding predictor sets as in Figure 4.

2021. We apply the same article filters to Dow Jones that we applied to the Reuters data,[18] arriving at a final article count of 5,378,838.

The second data set is the set of third-party news aggregated by Reuters, which we filtered out from our main analysis. This amounts to 2,265,171 articles over the period 1996-2022 after filtering.

We repeat our main analysis for each data set using the same procedure described in Section 4.1. In particular, we produce article embeddings via Mistral, compute stock-month average embeddings, residualize average embeddings versus JKP factors, and use MSRR to construct the news shock portfolio. Figure 21 reports the results. Third-party news shows weaker return prediction performance than the main Reuters data set but nonetheless achieves a Sharpe ratio as high as 2.3 (in the JKP case). On the other hand, the news shock portfolio derived from Dow Jones Newswires earns a Sharpe ratio as high as 3.7, exceeding the performance of the baseline Reuters data set. The conclusion from this analysis is that the news shock anomaly is robust to using alternative news data sources.

[18] Reuters filters include retaining articles tagged to single stocks, imposing minimum and maximum article lengths, etc. In addition, we follow the cleaning procedure of Ke et al. (2019) to handle articles that share the same accession number.

6.4 The "MSE Approach" and Portfolio Sorts

The MSRR approach in Section 4.1 estimates the stock-level portfolio weights as a function of (residual) embeddings to directly maximize the portfolio's Sharpe ratio. In this section we investigate an alternative approach to evaluating the news shock anomaly that takes a more conventional tack of first predicting stock-level returns using news embeddings, following CKX, then conducting portfolio sorts based on the predicted returns.

In particular, we modify the MSRR problem in (5)–(6) to a standard return predictive regression for minimizing squared errors ("MSE"). Referring once again to a generic D-vector of stock-level predictors $x_{i,t}$, the MSE model is

$$R_{i,t+1} = x'_{i,t}b + u_{i,t+1},$$

with estimation objective[19]

$$\min_b \sum_t \sum_i \left(R_{i,t+1} - x'_{i,t}b\right)^2 + \lambda b'b. \tag{11}$$

Given the estimator $\hat{b}$, we denote the predicted values from this regression as

$$\hat{\mu}_{i,t} = \hat{E}[R_{i,t+1}|x_{i,t}] = x'_{i,t}\hat{b}.$$

In other words, the MSE approach condenses the high-dimensional representation of news into a scalar expected return "characteristic." Given the condensed news characteristic $\hat{\mu}_{i,t}$, we conduct traditional quintile portfolio sorts following the asset pricing literature.

In the interest of space, we focus on the case where the return predictors $x_{i,t}$ are defined as residual embeddings $\epsilon_{i,t}$ derived from the full set of JKP characteristics. We form zero-net-investment quintile-spread portfolios that are long the 20% of stocks with the highest news-based expected return $\hat{\mu}_{i,t}$ and short those with the lowest expected returns. Long and short quintile portfolios are either equally weighted or value weighted (for value weights we use the "capped-value-weight" approach of JKP).

Panel A of Table 5 reports the performance of news shock portfolios using the MSE approach. The basic patterns in Table 5 align with the results documented in earlier sections that news shocks are significant predictors of returns. In the case of equal-weight portfolios, the quintile spread portfolio returns 12.8% per year (the alpha versus the FF6 model is also

[19]We estimate this panel predictive regression model in an expanding window and select λ with leave-one-out cross validation.

Table 5: News Shock Portfolio Sorts

This table reports the performance of decile portfolios sorted by MSE news signals. For each model specification, firms are sorted each month into ten deciles based on the MSE/MSRR predictions. We report the average monthly excess return (Avg), monthly standard deviation (Std), annualized Alpha with t-stat (Alpha/t-stat) and Sharpe ratio (SR).

Panel A: MSE

	Equal-weighted					Capped value-weighted				
	Avg	Std	Alpha	t-stat	SR	Avg	Std	Alpha	t-stat	SR
Low(L)	0.025	0.226	-0.067	-3.5	0.11	0.059	0.188	-0.021	-1.9	0.32
2	0.071	0.220	-0.020	-1.1	0.32	0.070	0.183	-0.010	-1.1	0.38
3	0.090	0.220	-0.002	-0.1	0.41	0.086	0.185	0.005	0.6	0.47
4	0.116	0.218	0.026	1.5	0.53	0.104	0.184	0.024	2.6	0.57
High(H)	0.153	0.226	0.061	3.2	0.68	0.127	0.187	0.045	4.6	0.68
H-L	0.128	0.048	0.128	13.3	2.68	0.067	0.054	0.067	6.1	1.25

Panel B: MSRR

	Equal-weighted					Capped value-weighted				
	Avg	Std	Alpha	t-stat	SR	Avg	Std	Alpha	t-stat	SR
Low(L)	0.033	0.221	-0.059	-3.3	0.15	0.062	0.187	-0.019	-1.8	0.33
2	0.075	0.224	-0.018	-1.0	0.34	0.078	0.186	-0.004	-0.4	0.42
3	0.094	0.220	0.003	0.2	0.43	0.088	0.183	0.008	0.9	0.48
4	0.114	0.222	0.023	1.2	0.51	0.100	0.184	0.020	2.2	0.54
High(H)	0.139	0.222	0.048	2.7	0.63	0.118	0.185	0.037	3.9	0.64
H-L	0.107	0.037	0.107	14.3	2.89	0.056	0.041	0.056	6.8	1.36

12.8%). For capped value-weight portfolios the quintile spread portfolio returns 6.7% per year (with an alpha of 6.7% and t-statistic of 6.1).

We can also sort stocks into portfolios based on the estimated portfolio weights $w_{i,t}$ from the MSRR specification in equation (5). This provides a direct comparison of the MSE and MSRR approaches. When sorting on MSRR weights from our main JKP specification we find an equal-weight news shock anomaly Sharpe ratio of 2.9, compared to 3.1 for the main MSRR analysis in Figure 4 and 2.7 for the MSE approach (and Sharpe ratios of 1.4 versus 1.3 for the value-weight versions of MSRR and MSE sorts, respectively). MSE and MSRR methods are based on the same data and differ only in their estimation objective.

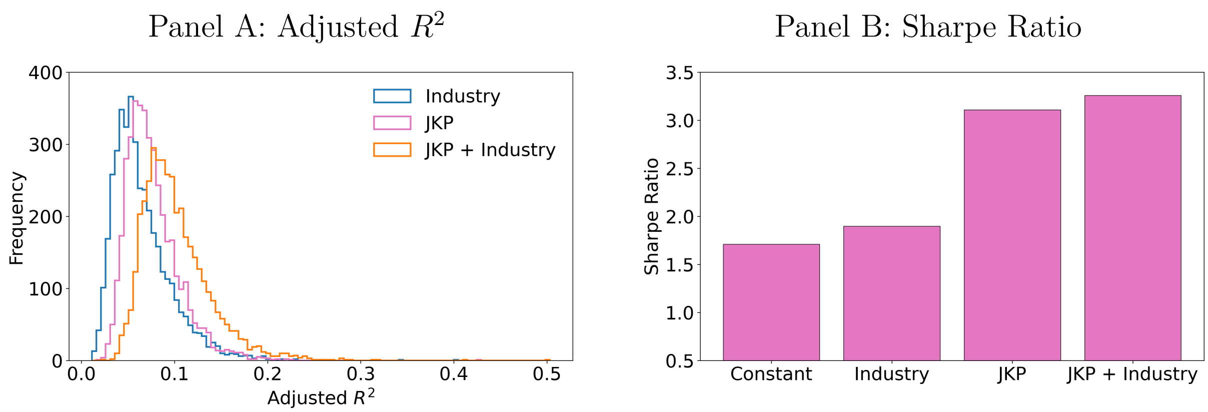


Figure 22: Adjusted R^2 and Sharpe Ratio with Industry Fixed Effects

Note. Panel A shows the distribution of adjusted R^2 across embedding coordinates for three specifications: "Industry," which predicts embeddings using 25 GICS industry indicators; "JKP," which uses 132 stock characteristics; and "JKP+Industry," which combines the two. Panel B reports the Sharpe ratios of news shock portfolios constructed using each specification.

They both deliver strong performance and their returns are highly correlated (67%). The relative outperformance of MSRR indicates that a Sharpe ratio-based training objective more precisely identifies the investment-relevant aspects of price inefficiency associated with news shocks.

The main conclusion from the analysis in Table 5 is that our conclusions do not depend on the methodology used to assess news-based predictability. Whether using direct MSRR portfolios, using traditional portfolio sorts based on MSRR estimates, or using sorts based on predictive regression estimates, we find evidence for inefficient pricing of news shock with a magnitude that stands against the backdrop of the broader anomaly literature.

6.5 The Role of Industry Information

Section 4.2 demonstrates that industry indicators are marginally useful predictors of news after controlling for JKP characteristics. The lexicon describing a biotechnology firm (e.g., "clinical trials," "FDA approval") is structurally distinct from that describing an energy producer (e.g., "drilling," "crude reserves"). In this section, we investigate the incremental portfolio performance impact of using industry information to construct news shocks.

Panel A of Figure 22 builds on the earlier analysis of Figure 2 to compare the power of industry indicators for predicting news. The distribution of predictive R^2 across embedding coordinates has a slightly lower mean R^2 based on industry dummies (6.5%) than that based on JKP characteristics (mean R^2 of 7.7%). Stock characteristics and industry dummies contain some distinct information from one another, as evidenced by the orange curve showing the effects of combining both predictor sets together in S_t. The combination increases the average news prediction R^2 to 10.2%, and increases the maximum R^2 from 42.6% for JKP alone to 50.3% when including industry effects.

While industry indicators have incremental explanatory power for news, they have little impact on the performance of the news shock portfolio. Panel B of Figure 22 reports MSRR results for the industry-only residualization, as well as the case when industry dummies are combined with JKP factors. The improvement from a constant-only embedding prediction model to an industry-based model improves the Sharpe ratio from 1.7 to 1.9. As previously shown, there is a much larger improvement from the JKP-based (to a Sharpe ratio of 3.1). Including both JKP and industry predictors raises the news shock Sharpe ratio further to 3.3. While this gain is small in magnitude, it is statistically significant, as the alpha t-statistic from regressing the JKP+Industry model on the JKP-only model is 3.8.

In summary, while news text has a high degree of industry specificity as shown in Panel A, much of this can be captured by controlling for stock characteristics. This fact is explained in part by the fact that stock characteristics tend to cluster by industry (e.g., tech stocks tend to have low book-to-market ratios and high volatility; utility stocks have low betas and high dividend-payout ratios). But non-industry characteristics also help purge other predictable aspects of news in a manner that significantly enhances our ability to detect news-based price inefficiencies.

6.6 Subsample Analysis

Section 4.4 explores robustness of the news shock anomaly in different subsamples of the cross section. In this section, we investigate the robustness of news shock portfolios in different time subsamples. In particular, we compute rolling 5-year Sharpe ratios of the main news shock portfolio (based on embeddings residualized against JKP characteristics). Figure 23 reports the results. Subsample Sharpe ratios range between 2.1 and 4.5. The advent of LLMs corresponds loosely to the post 2018-sample (after BERT is introduced), where we see a level shift in Sharpe ratio to the range of 2.0 to 2.5. This likely arises from LLMs aiding the integration of news-based information into asset managers' portfolios, thus increasing competition around the news shock anomaly and decreasing its returns.

Figure 23: News Shock Portfolio 5-Year Rolling Sharpe Ratio

Note. This figure plots the rolling 5-year Sharpe ratio of the main news shock portfolio based on embeddings residualized against JKP characteristics.

6.7 Alternative Training Windows

An important modeling decision in machine learning applications for asset pricing is the training window size (Gu et al., 2020). In our baseline MSRR specification, we use an expanding window (with a minimum window of twelve months early in the sample). The expanding window maximizes the amount of data used in training. In this section, we assess the sensitivity of using a shorter rolling training window; we consider windows as short as the most recent six months. Figure 24 reports Sharpe ratios of the main news shock portfolio when MSRR is trained in rolling windows of $[6, 12, 36, 60, 120]$, with the last point on the x-axis representing the expanding window specification used in our baseline analysis. The conclusions from Figure 24 are i) that performance of the news shock portfolio increases with the length of the training window and ii) even with very short windows of a year or less, the portfolio Sharpe ratio remains above 2.1.

7 Conclusions

We show that the content of stock-level news articles is predictable based on a stock's prevailing characteristics. We use this insight to purge articles of "old news" and isolate

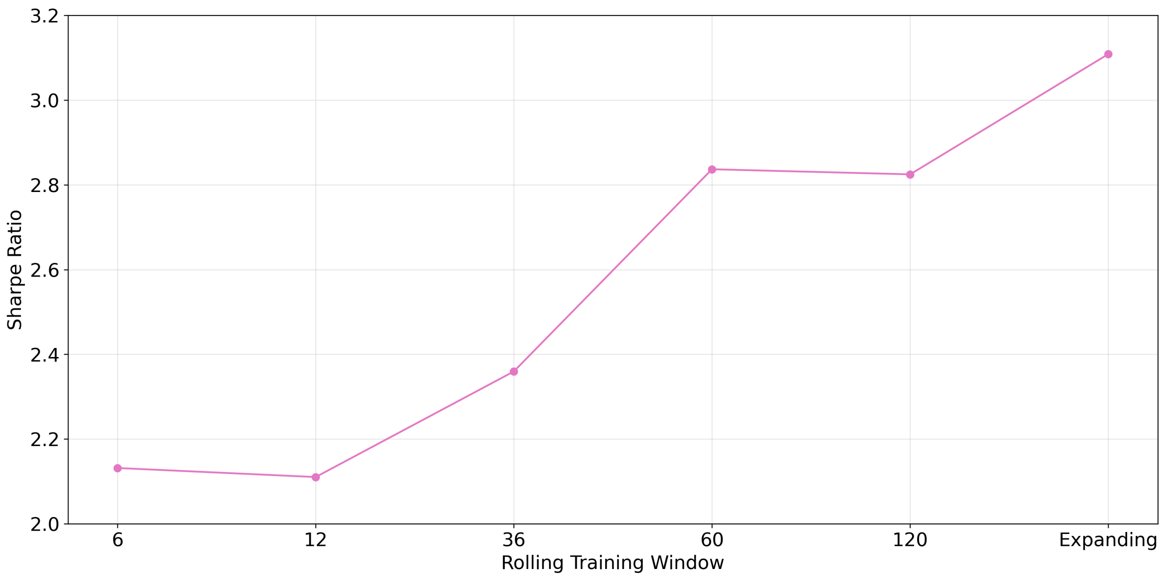


Figure 24: News Shock Portfolio Performance With Alternative Training Windows

Note. This figure shows the Sharpe ratio of the MSRR news portfolio for different choices of the maximum lookback window. All other aspects of the model follow the baseline specification. The x-axis reports the maximum window length (in months) used when estimating the news portfolio, with "Expanding" indicating an unrestricted expanding window.

new information arrival which we refer to as "news shocks." News shocks demonstrate pronounced and prolonged return predictability with greater magnitude than many other anomalies in the literature. The inefficient pricing of news shocks is highly robust. It exists in a variety of samples (both different asset universes and different time periods), with embeddings from a variety of different LLMs, with multiple news data sources, and with different predictive modeling approaches. Our results are not driven by look-ahead bias in pre-trained LLMs—repeating our analysis using "chronologically consistent" LLMs confirms the qualitative conclusions of our main analysis.

We identify 12 economically interpretable themes whose news is inefficiently incorporated into prices and give rise to the news shock anomaly. We also show that different news types possess different predictable and profitable price dynamics. Some news produces large price reversals, while others lead to strong price momentum. However, the profitability of the news shock anomaly derives predominantly from its ability to exploit the market's underreaction to news.

References

Ahmed, N. and Wahed, M. (2020). The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15:659–688.

Ba, C., Bohren, J. A., and Imas, A. (2024). Over-and underreaction to information. *Working Paper*.

Ball, R., Gerakos, J., Linnainmaa, J. T., and Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of financial economics*, 121(1):28–45.

Barberis, N. (2018). Psychology-based models of asset prices and returns. *The Oxford Handbook of Behavioral Economics and Behavioral Finance*, pages 1–44.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Baumol, W. J. (1965). *The Stock Market and Economic Efficiency*. Fordham University Press, New York.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671.

Bybee, L., Kelly, B., and Su, Y. (2023). Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2024). The structure of economic news. *The Review of Economic Studies*, 91(2):689–737.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2022). Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *The Critical Finance Review*, 11(2):207–264.

Chen, H., Didisheim, A., Somoza, L., and Tian, H. (2025). A financial brain scan of the llm. *arXiv preprint arXiv:2508.21285*.

Chen, Y., Kelly, B. T., and Xiu, D. (2024). Expected returns and large language models. Working Paper 4416687, Social Science Research Network. https://ssrn.com/abstract=4416687.

Cowles, A. (1933). Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, 1(3):309–324.

Cunningham, H. et al. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *The Journal of Finance*, 53(6):1839–1885.

Didisheim, A., Ke, S. B., Kelly, B. T., and Malamud, S. (2024). Apt or "aipt"? the surprising dominance of large factor models. Technical report, National Bureau of Economic Research.

Dow, J. and Gorton, G. (1997). Stock market efficiency and economic efficiency: Is there a connection? *The Journal of Finance*, 52(3):1087–1129.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Fei, N., Gao, Y., Lu, Z., and Xiang, T. (2021). Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 142–151.

Feldman, R. J. and Schmidt, J. E. (2003). Supervisory disclosures and financial system resilience: An empirical analysis. *The B.E. Journal of Macroeconomics*, 3(1):1–24.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2018). Trading costs. *Available at SSRN 3229719*.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. (2019). Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Ge, T., Hu, J., Wang, L., Wang, X., Chen, S.-Q., and Wei, F. (2023). In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Glasserman, P. and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. *arXiv preprint arXiv:2309.17322*.

Grattafiori, A. et al. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783. github.com/meta-llama/llama3.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically consistent large language models. *arXiv preprint arXiv:2502.21206*.

Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial Economics*, 7(1):133–159.

Hoberg, G. and Manela, A. (2025). The natural language of finance. *Foundations and Trends in Finance*, 14(4):244–365.

Hoberg, G. and Phillips, G. M. (2018). Text-based industry momentum. *Journal of Financial and Quantitative Analysis*, 53(6):2355–2388.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184.

Hou, K., Mo, H., Xue, C., and Zhang, L. (2021). An augmented q-factor model with expected growth. *Review of Finance*, 25(1):1–41.

Hou, K., Xue, C., and Zhang, L. (2020a). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Hou, K., Xue, C., and Zhang, L. (2020b). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Huang, H., LeCun, Y., and Balestriero, R. (2025). Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2022). Is there a replication crisis in finance? *The Journal of Finance*.

Jiang, H., Li, S. Z., and Wang, H. (2021). Pervasive underreaction: Evidence from high-frequency data. *Journal of Financial Economics*, 141(2):573–599.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. *The Journal of Finance*, 74(6):2987–3035.

Kelly, B., Malamud, S., Schwab, J., and Xu, A. (2026). Improvements in Point-in-time LLMs. Yale working paper, Social Science Research Network.

Kelly, B. T. and Xiu, D. (2023). Financial machine learning. *Available at SSRN*.

Kwon, S. Y. and Tang, J. (2025). Extreme categories and overreaction to news. *Review of Economic Studies*, page rdaf037.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. (2024). Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300.

Lopez-Lira, A. and Tang, Y. (2024). Can chatgpt forecast stock price movements? return predictability and large language models. *Journal of Financial Economics*. Forthcoming.

Loughran, T. and McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12(1):357–375.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Meursault, V., Liang, P. J., Routledge, B. R., and Scanlon, M. M. (2023). Pead.txt: post-earnings-announcement drift using text. *Journal of Financial and Quantitative Analysis*, 58(6):2299–2326.

Novy-Marx, R. and Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, 29(1):104–147.

Novy-Marx, R. and Velikov, M. (2024). Assaying anomalies. *Available at SSRN 4338007*.

Patil, R., Boit, S., Gudivada, V., and Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Pénasse, J. (2022). Understanding alpha decay. *Management Science*, 68(5):3966–3973.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. SSRN Working Paper No. 4754678.

Shleifer, A. (2000). *Inefficient Markets: An Introduction to Behavioral Finance*. Clarendon Lectures in Economics. Oxford University Press, Oxford.

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., and Du, M. (2025). A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1690–1712. Association for Computational Linguistics.

Su, J., Cao, J., Liu, W., and Ou, Y. (2021). Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Tao, C., Shen, T., Gao, S., Zhang, J., Li, Z., Hua, K., Hu, W., Tao, Z., and Ma, S. (2024). Llms are also effective embedding models: An in-depth overview. *arXiv preprint arXiv:2412.12591*.

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

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Improving text embeddings with large language models.

Wang, Y., Zhang, B., and Zhu, X. (2018). The momentum of news. *Available at SSRN 3267337*.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.

A Additional Results

Table 6: News Shock Portfolio Exposures—FF5 and Q5 Models

This table reports regressions of the news shock portfolio constructed from various news predictor sets (Constant, CAPM, FF3, FF6, and JKP) on the Fama and French (2015) five-factor model (Panel A) and the Hou et al. (2021) Q5-factor model (Panel B). All portfolios are (ex post) standardized to have 10% annual volatility to aid interpretation of alpha and beta coefficients, and alphas are reported in annualized terms.

Panel A: FF5 Model Exposures

	Constant	CAPM	FF3	FF6	JKP
Alpha	0.153***	0.165***	0.237***	0.256***	0.305***
	(7.96)	(8.66)	(12.08)	(13.05)	(15.14)
Market	0.048	-0.016	0.064	0.018	0.002
	(0.79)	(-0.26)	(1.03)	(0.30)	(0.03)
SMB	0.046	0.037	0.212***	0.178***	0.142**
	(0.75)	(0.60)	(3.35)	(2.82)	(2.18)
HML	-0.511***	-0.511***	-0.372***	-0.363***	-0.326***
	(-6.78)	(-6.84)	(-4.82)	(-4.72)	(-4.12)
RMW	0.359***	0.348***	0.370***	0.266***	0.169**
	(5.17)	(5.06)	(5.20)	(3.75)	(2.32)
CMA	0.136*	0.110	-0.019	-0.080	-0.023
	(1.87)	(1.52)	(-0.25)	(-1.08)	(-0.30)
R^2	0.179	0.196	0.141	0.145	0.097

Panel B: Q5 Model Exposures

	Constant	CAPM	FF3	FF6	JKP
Alpha	0.154***	0.166***	0.248***	0.267***	0.310***
	(7.76)	(8.52)	(12.16)	(13.05)	(14.59)
Market	0.076	0.014	0.049	-0.022	-0.014
	(1.16)	(0.22)	(0.73)	(-0.33)	(-0.20)
Size	0.076	0.073	0.160***	0.100	0.069
	(1.28)	(1.25)	(2.62)	(1.64)	(1.09)
Investment	-0.221***	-0.250***	-0.280***	-0.345***	-0.214***
	(-3.91)	(-4.50)	(-4.83)	(-5.94)	(-3.55)
ROE	0.394***	0.427***	0.313***	0.098	0.013
	(5.33)	(5.87)	(4.12)	(1.29)	(0.17)
Expected Growth	0.073	0.041	-0.011	0.059	0.093
	(0.99)	(0.56)	(-0.15)	(0.77)	(1.17)
R^2	0.17	0.197	0.124	0.12	0.052

B Interpretable Factors Additional Results

Table 7: Classification of Interpretable News Factors into Economic Themes
This table presents the 12 broad economic themes derived from the manual clustering of the 144 unique sparse features selected by the MSRR procedure. The second column provides representative examples of the LLM-generated semantic labels corresponding to features within each cluster.

Theme	**Example Factors**
Earnings & Financial Results	Quarterly sales/revenue change announcements; Quarterly earnings results: EPS and revenue; Fourth-quarter and full-year results metrics; GAAP earnings or loss per share; Record/strong earnings with raised guidance; Strong year-over-year growth metrics; Quantitative growth and increases (percent); Quarterly company operating metrics changes; Quarterly comps/subscriber metrics and guidance updates; Operational/financial metrics with numeric guidance *(+12 more)*
Corporate Guidance & Outlook	Company growth metrics and forward guidance; Corporate forward-looking guidance and forecasts; Cautious corporate guidance amid uncertainty; Company press-release style corporate actions and guidance; Corporate operational outlook and risk commentary; Company-specific operational/financial update headlines; Company operational milestones and forward guidance; Company financial guidance and outlook statements; Disappointing performance and underperformance signals; Negative outlook: weak performance, losses, distress *(+7 more)*

Theme	**Example Factors**
Analyst Ratings & Sentiment	Analyst downgrades and estimate cuts; Analyst upgrades and raised price targets; Analyst rating and price target changes; Analyst upgrades and raised price targets; Analyst rating downgrade to "perform"; Multi-company analyst notes and updates; Bullish praise and positive endorsements; Management/activist disappointment with performance; Analyst downgrades to market perform; Analyst ratings and outlook ranges *(+2 more)*
Distress, Bankruptcy & Delisting	Chapter 11 bankruptcy and going-concern distress; Distressed debt refinancing and bankruptcy risk; Severe distress: downgrade, default, bankruptcy risk; Credit downgrades and severe financial distress; Severe business distress and downside risk; Severe cost-cutting and distress measures; Bankruptcy and business shutdown announcements; Regulatory enforcement, lawsuits, delistings, bankruptcies; Delisting risk and contract termination; Late SEC filings and delisting risk *(+7 more)*
Momentum & Trading Activity	Turnaround signals: improving profits, momentum; Sharp market selloffs and volatility spikes; Reversal or slowing of momentum; Small-cap catalyst-driven stock surges; Meme-stock short squeeze retail frenzy; Unusually high trading volume spikes; Unusually high stock trading volume spikes; Premarket/intraday stock price moves; Stock volatility spikes on company news; Stock splits and reverse splits
Corporate Actions & Restructuring	Corporate financing and capital markets actions; Uncertain potential deals or breakthroughs; SPAC reverse-merger going-public deals; Regulatory or deal intent indications; Deal terminations and corporate wind-downs; Strategic alternatives and financial restructuring; Corporate actions and company announcements; Miscellaneous corporate event disclosures; Corporate agreements and order imbalances; Corporate agreements and litigation announcements *(+10 more)*

Theme	**Example Factors**
Leadership & Governance	Executive appointments and leadership changes; Executive turnover: CEO/CFO resignations and interim appointments; Executive transitions to chair/advisor roles; CEO appointment, succession, or resignation news; Corporate executive leadership transitions and successions; Executive/Founder leadership and board changes; CEO succession planning and executive search; Shareholder fiduciary duty breach investigations; Executive/board changes and financing events; Retaining outside firms for searches/investigations
Growth & Demand Trends	Rising scale metrics and guidance increases; Strong demand driving record profits/outlooks; Moderating growth and demand slowdown; Earnings headwinds and negative guidance; Worsening losses, provisions, guidance cuts; Structural decline in legacy industries demand; Decline of legacy physical media industries; Corporate financial outperformance and shareholder returns; Highly specific company operational disclosures; U.S. domestic market metrics and outlook *(+2 more)*
Biotech, Pharma & Healthcare	Disappointing outcomes: trial/regulatory/court setbacks; Clinical trial endpoints and efficacy results; Alzheimer's disease clinical trials results; Clinical trial updates and FDA approvals; Keytruda-focused oncology trial updates; Product-specific regulatory/availability announcements (incl. COVID variants); Genomics and sequencing technology developments; Opioid misuse, overdose, drug safety; Corporate COVID-19 response: PPE, testing, vaccines; Biotech clinical trial data shock moves *(+4 more)*

Theme	**Example Factors**
Regulatory & Legal Actions	Government ministry/regulatory approvals and licenses; Upcoming hearings/committee meetings; pending decisions; Consumer protection enforcement, refunds, penalties; E-cigarette/vaping tobacco regulatory actions; Cryptocurrency regulatory scrutiny and adoption; Suspensions, halts, and crisis disruptions; Corporate newswire headlines mixing financial results with fraud/conviction headlines; Investigation findings on incidents/attacks; Canada-focused government and policy news; Major corporate scandals and litigation settlements *(+1 more)*
Sector-Specific Signals	Netflix subscription pricing and subscriber metrics; Steel and aluminum production economics; Airline baggage and change fees; Video game publishers' franchise releases and delays; Pay-TV streaming carriage rights deals; Casual dining chain restaurant operations; Photovoltaic solar panels modules inverters supply; Airline fare sales and change-fee waivers; FXCM retail forex trading metrics; Call center/BPO service center expansion or closure *(+10 more)*
Product Launches & Operations	New product launch and regulatory approval; Corporate operational updates: orders, launches, partnerships; Specialty business units and financing updates; Percentages and ownership/delinquency rates; Product safety incidents, recalls, injuries/deaths; Operational/transaction disruptions and delays

System Prompt

```
You are analyzing a single latent feature of a finance model. For this feature, you will receive multiple news headlines where the feature's activation is high. The headlines are ordered from MOST activation (first) to LEAST activation (last).
Your task:

• Infer what concept, pattern, or property this feature is detecting.

• If relevant, the concept should be associated to risk factors.

• Use ALL the headlines to form your judgment, pay more attention to the earlier examples.

• Focus on what the main concepts headlines have in common, not on rare or incidental details.

Output requirement:

• Output ONLY a short human-readable description of the concept (around 5 words).

• Do NOT add any prefixes such as 'This feature detects...'.

• Do NOT output anything else: no explanations, lists, reasoning, or formatting.
```

User Content

```
Below are {N} news headlines where this feature has high activation.
Headlines:
{List of Headlines}

Based on these headlines, describe what this feature most likely represents.
```

Figure 25: The exact system and user prompts used to generate semantic labels for the SAE features. The model is provided with the top-activating headlines for a specific feature index and instructed to synthesize a short, finance-related concept label.

Table 8: Selected High-Activation Headlines

Representative headlines selected from the highest-scoring activations of each feature, ranked by cosine similarity to the learned SAE direction. Headlines are drawn from the Reuters/Dow Jones news corpus (1996–2022).

Panel A: Cryptocurrency	
1	U.S. Treasury official says any cryptocurrency project, including Libra, operating in whole or substantial parts of U.S. will clearly have to satisfy U.S. regulatory standards regardless of base
2	Coinbase says it has received regulatory approval in Ireland to operate as a virtual asset service provider (VASP)
3	Google ends ban on cryptocurrency-related ads, plans to allow regulated crypto exchanges to buy ads in U.S., Japan; new policy starts in October
4	BNY Mellon will hold, transfer and issue Bitcoin and other cryptocurrencies on behalf of its asset-management clients
5	Musk says Bitcoin paid to Tesla will be retained as Bitcoin, not converted to fiat currency
Panel B: Meme-Stock Short Squeeze Retail Frenzy	
1	AMC Entertainment Holdings Inc — more than 80% of AMC shares are held by a broad base of retail investors with an average holding of around 120 shares
2	Popular Reddit "WallStreetBets" forum which investors said helped drive surge in GameStop appears to go private
3	Citadel CEO Ken Griffin says "I think the GameStop situation is incredibly unique in that it was such a heavily shorted stock"
4	Shares of GameStop reverse loss, last up 2% after Keith Gill, known as 'Roaring Kitty', gives congressional testimony
5	Robinhood must face market manipulation claims over trading restrictions during last year's "meme stock" rally — U.S. judge