

Reference-Dependent Sentiment^{*}

Litong Zhang[†]

October 20, 2026

Abstract

Under what circumstances is a given paper gain/loss more relevant to trading decisions? I examine whether reference-dependent thinking among investors is associated with the strength of disposition-effect trading. Using 185.1 million StockTwits posts from 2013 to 2024 and an LLM-supervised classification, I construct a daily measure of reference-dependent sentiment, capturing whether investors frame their positions as gains or losses relative to past prices. In the cross-section, talking about gains (losses) is associated with more (less) retail selling of stocks with embedded gains (losses). At the market level, when investors talk more about gains (losses), turnover is higher (lower) on up-market (down-market) days. Around earnings announcements, buying good-news stocks after investors have been talking about gains and selling bad-news stocks after they have been talking about losses earns a Fama-French five-factor alpha of 1.03% per month, or 12.4% annualized, while the reverse strategy does not. Exploiting the introduction of StockTwits' leaderboard as a plausible shock to framing legitimacy, I find that public reinforcement of gain-loss framing strengthens the relationship between reference-dependent sentiment and trading. Overall, reference-dependent sentiment identifies when the disposition effect shows up more strongly in trading and in prices around earnings news.

JEL: G12, G14, G41.

Keywords: reference-dependent thinking, disposition effect, social media, capital gain overhang, LLM-supervised classification

^{*}I especially thank my advisor, David Hirshleifer, for his invaluable advice and guidance. I also thank Odilon Câmara, Ricardo De la O, Cary Frydman, Gerard Hoberg, Scott Joslin, Mete Kilic, Yaron Levi, and Miao Ben Zhang for their suggestions and comments.

[†]University of Southern California, Marshall School of Business, 701 Exposition Blvd., Ste. 231, Los Angeles, CA 90089. Email: Litong.Zhang@marshall.usc.edu.

1 Introduction

Prospect theory and realization utility both rely fundamentally on comparisons between current outcomes and reference points (Kahneman and Tversky 1979; Barberis and Xiong 2012). Perhaps the most robust empirical manifestation of such reference dependence in financial markets is the disposition effect: the tendency of investors to realize winning positions more readily than losing ones (Shefrin and Statman 1985; Odean 1998).

Standard explanations for the disposition effect center on preference-based interpretations. Barberis and Xiong (2009, 2012) and Ingersoll and Jin (2013) show that an investor will predictably exhibit a disposition effect if she has preferences defined over realized gains and losses.

At the same time, realization utility need not be equally important for all investors or in all circumstances (Barberis and Xiong 2012). For example, the shareholders of a stock can sit on similar paper gains at different moments but think about them differently. At one moment, they may evaluate their positions relative to what they paid and think of them as gains. At another moment, they may hold the same positions but evaluate them in terms of the firm's future prospects rather than their purchase prices. At a third, they may not even be paying attention to the gain at all. Among these, the first way of thinking is an example of *reference-dependent thinking* and is most likely to generate disposition-effect selling; applied to paper losses, the same way of thinking is most likely to generate a reluctance to sell.

This raises a simple question: under what circumstances is a given paper gain more likely to generate widespread gain realization? The previous example suggests that the same paper gain need not generate the same degree of disposition-effect trading; what also matters is whether and when reference-dependent thinking is active among investors.

While existing experimental and natural-experiment evidence shows that making a purchase price more salient strengthens disposition-effect trading among individual investors (Frydman and Rangel 2014; Frydman and Wang 2020), much less is known about whether individual reference-dependent thinking aggregates across investors, and whether it is associated with more selling where investors have gains to realize and helps predict subsequent returns.

Empirically, this question is difficult to answer. Reference-dependent thinking is inherently difficult to observe because it describes how investors think about their positions, rather than an outcome recorded in standard financial data (Cookson and Niessner 2020). Brokerage holdings and transaction records reveal investors' positions and trades, but not whether they are thinking about those positions in gain and loss terms at a given moment. Inferring this way of thinking from observed trading behavior alone is also difficult, because the same observed behavior can be consistent with different underlying mechanisms (Choi and Robertson 2020; Liu et al. 2022).

In this paper, I explore linguistic evidence of how investors express this reference-dependent thinking in their own language. This way of thinking does not necessarily remain unspoken. On investor social media, investors routinely communicate investing episodes to peers in gain and loss terms as their positions evolve (Figure 1).¹ I aggregate these individual expressions and define the tendency of such gain/loss framing in the market as *reference-dependent sentiment*: an observable measure of reference-dependent thinking among investors.

I implement this measure using 185.1 million posts from StockTwits between 2013 and 2024. StockTwits is a financial social media platform organized around investment discussion and has been widely used in the literature to study investor sentiment. Existing work shows that StockTwits activity has a significant predictive association with market variables such as trading volume, volatility, and returns.²

Crucially for studying reference-dependent sentiment, posts on the platform are publicly organized by cashtags: any investor looking up a given security encounters the gain/loss framings that others have recently posted, so investors' gain/loss framings are visible to others rather than confined to private experience. Unlike general-purpose social media, the platform draws a self-selected audience of active investors whose market commentary frequently reflects their own gain and loss exposure. Because this discourse occurs alongside ongoing market dynamics, I can track reference-dependent sentiment in real time.

A central challenge in constructing a measure of reference-dependent sentiment is to identify gain and loss language in StockTwits posts. The same word can carry very

¹Illustrative StockTwits posts include: *\$AAPL up 18% from my entry* and *Closed \$AMZN for a nice gain*. Each post describes a position as a gain or a loss relative to a past reference point.

²See Cookson and Niessner (2020, 2023); Renault (2017) and Gu, Hirshleifer, Teoh, and Wu (2025).

different economic meanings depending on the comparison being made. For example, a message saying *I'm up from \$110* describes a backward-looking comparison to a reference price, while a message saying *this stock will go up* expresses a forward-looking belief about future returns. A dictionary-based sentiment measure, which is designed to capture ordinary optimism or pessimism, would treat both uses of *up* similarly, even though only the first reflects gain/loss framing relative to a past price anchor.³

To better recover this distinction, I use a large language model (LLM), GPT-4o, to supervise a classification pipeline that evaluates investor posts along three economically motivated dimensions: whether the message is about a holder's position; whether it is gain framed or loss framed; and whether the comparison is backward-looking or forward-looking. This structure allows the LLM classifier to distinguish messages that reflect reference-dependent thinking from messages that merely express optimism, pessimism, price expectations, or other general commentary. The resulting classifications align closely with my manual coding in a held-out validation sample and exhibit strong out-of-sample performance, indicating that the pipeline captures economically meaningful gain/loss framing rather than generic sentiment or keyword usage.

I then classify the full StockTwits corpus using a two-stage pipeline that balances semantic accuracy with feasibility at corpus scale. In the first stage, the LLM serves as the *teacher*, using detailed classification rubrics to label a representative training sample.⁴ In the second stage, a *student* model (DeBERTa-v3-base) is trained on these labels and applied to the full corpus, making large-scale inference feasible. This pipeline outputs soft probabilities for the object, frame, and anchor of every post.

Using these message-level probabilities, I construct two families of daily measures of reference-dependent sentiment over the sample period: a market-wide measure of the gain/loss composition of reference-dependent discourse, and stock-day measures that capture cross-sectional variation in this composition across individual stocks. This construction places the greatest weight on messages that jointly have a high probability of being about a holder's position, a backward-looking comparison, and a gain

³The *past prices* that anchor these comparisons are usually the investor's own cost basis. Common examples include a single purchase price (*I'm green on \$NVDA from \$110*), a blended average cost (*still underwater on my \$85 avg*), and a stated entry on an open lot (*up 18% from my entry*). What unifies these anchors is that each defines a reference point relative to which the current price is encoded as a gain or a loss in the sense of reference-dependent thinking.

⁴See Internet Appendix K for the detailed classification rubrics.

or loss frame. Among this reference-point-weighted set of messages, GainTalk measures how strongly the prevailing reference-dependent sentiment is tilted toward gains rather than losses. Section 2 provides the formal construction.

With these measures in place, the paper makes four empirical contributions.

First, I show that reference-dependent sentiment is associated with selling in the cross-section of individual stocks. Following Grinblatt and Han (2005), I use capital gain overhang to proxy for the average unrealized gain or loss of a stock's shareholder base, and retail sell flow to isolate selling (Barber et al. 2024). Higher GainTalk is associated with more retail selling, and this association is stronger in stocks whose shareholders sit on larger embedded gains: with firm and date fixed effects, a one-standard-deviation increase in capital gain overhang raises the GainTalk coefficient from 0.0152 to 0.0295, that is, from about 1.5 to about 3.0 percent higher abnormal retail sell flow per standard deviation of GainTalk (interaction $t = 19.26$). Thus, reference-dependent sentiment is associated with selling where shareholders have embedded gains to realize, the sell-side margin on which the disposition effect operates. Separating gain-framing and loss-framing intensity sharpens this pattern: gain-framing intensity is associated with more retail selling only in stocks with embedded gains ($+0.0126$, $t = 10.30$), and loss-framing intensity with less retail selling only in stocks with embedded losses (-0.0165 , $t = 8.88$).

Second, I show that aggregate turnover provides a market-level footprint of the same state-dependent trading. Higher GainTalk is associated with more turnover, and this association is stronger on days with higher market returns, which proxy for the prevailing gain/loss state: a one-standard-deviation increase in GainTalk is associated with abnormal turnover higher by about 4.1 percent on a day with a 1 percent market gain (interaction $t = 4.06$). The difference, 0.0120 per percentage point of return, is about 6.3 percent of the standard deviation of abnormal log market turnover (0.190).

Moreover, when gain-framed discourse is prevalent on days when aggregate returns are positive, turnover rises above what returns alone predict, consistent with a greater propensity to realize gains in up markets: a one-standard-deviation increase in gain-framing intensity on up-market days is associated with 0.91 percent higher abnormal turnover ($t = 2.55$). When loss-framed discourse is prevalent on days when aggregate returns are negative, turnover falls below what returns alone predict, consistent

with greater reluctance to realize losses in down markets: a one-standard-deviation increase in loss-framing intensity on down-market days is associated with 1.58 percent lower abnormal turnover ($t=3.56$).

Third, around earnings announcements I test whether reference-dependent sentiment predicts subsequent risk-adjusted returns. Using a double-sort strategy following [Frazzini \(2006\)](#), I sort stocks by the earnings surprise and by pre-event reference-dependent sentiment. A portfolio that buys good-news stocks with high pre-event gain-framed discourse and sells bad-news stocks with high pre-event loss-framed discourse earns a Fama-French five-factor (FF5) alpha of 1.031% per month over a three-month holding period ($t=3.64$), or about 12.4% over twelve months; the reverse framing portfolio, which buys good-news stocks with high pre-event loss-framed discourse and sells bad-news stocks with high pre-event gain-framed discourse, earns alphas indistinguishable from zero (-0.078% , $t=0.29$). Earnings announcements therefore provide an event setting in which reference-dependent sentiment appears in subsequent risk-adjusted returns, and the continuation rather than reversal of returns is difficult to reconcile with informed selling.

Fourth, I use a platform-level shock to framing legitimacy, the extent to which investors regard gain-loss evaluation as an appropriate basis for trading, to examine whether public reinforcement of gain-loss framing strengthens this association. In November 2023, StockTwits introduced a feature that prominently displayed the day's top 30 gainers and losers. On the gainer side, a one-standard-deviation increase in GainTalk is associated with about 17 percent more same-day abnormal retail selling after the rollout for stocks that reappear on the list on consecutive trading days, roughly four times the 4 percent for non-recurring stocks on the list. Before the rollout, the two groups show similar associations (about 4.4 and 3.5 percent). Among stocks repeatedly classified as losers, selling rises without a comparable change in this association, so the pattern is not simply a uniform increase in trading. The evidence is consistent with a channel in which repeated public classification reinforces gain-loss framing and strengthens its link to selling.

Together, these findings are consistent with reference-dependent preferences: investors evaluate positions relative to reference points ([Kahneman and Tversky 1979](#)), and when preferences are defined over realized gains and losses they exhibit disposition-

effect trading ([Barberis and Xiong 2012](#); [Ingersoll and Jin 2013](#)). Reference-dependent sentiment predicts a stronger link between the prevailing gain-loss state and trading activity.⁵

Finally, this paper develops a scalable methodology for measuring reference-dependent sentiment in naturally occurring investor-generated text. Reference-dependent thinking is difficult to identify in large-scale text data using standard textual methods because it depends on the economic context and semantic meaning of language rather than the presence of particular words. The same words can express a backward-looking gain/loss framing, a forward-looking return belief, or generic commentary, depending on the anchor and logic of the comparison.

An LLM-supervised classification pipeline implements that distinction at corpus scale. A teacher model labels whose position is being evaluated, whether the frame is a gain or a loss, and whether the anchor is in the past or the future; a student model applies those labels to the full corpus.

Related Literature These findings speak to several literatures. This paper contributes to the literature on the disposition effect. One strand concentrates on formal models of why investors sell winners more readily than losers, including preference-based and realization-utility explanations ([Shefrin and Statman 1985](#); [Barberis and Xiong 2009, 2012](#); [Ingersoll and Jin 2013](#)). Another strand tests these ideas empirically in brokerage and market data ([Odean 1998](#); [Grinblatt and Keloharju 2001](#); [Grinblatt and Han 2005](#); [Ben-David and Hirshleifer 2012](#); [Frydman et al. 2014](#); [An et al. 2024](#); [Andries et al. 2025](#)). [Grinblatt and Han \(2005\)](#) construct capital gain overhang as a stock-level proxy for investors’ embedded gain-loss positions and show that it predicts selling pressure and returns; [Ben-David and Hirshleifer \(2012\)](#) document related patterns in retail order flow. Relative to this literature, I study a distinct marker of when disposition-effect trading is stronger: reference-dependent thinking revealed in public investor discourse. In the cross-section, gain-framed discourse is associated with more retail selling where capital gain overhang is higher. At the market level, gain-framed discourse is associated with more turnover, and this association is stronger on days with higher market returns.

⁵Internet Appendix [A](#) provides a stylized model in which reference-dependent preferences interact with the prevalence of gain/loss framing and with the weight investors place on that framing. Proposition [1](#) derives four comparative statics that motivate and correspond directly to the empirical tests in Sections [4](#), [5](#), and [7](#).

Around earnings announcements, a disposition spread that sorts on the surprise and pre-event reference-dependent sentiment earns significant FF5 alphas, whereas the reverse spread does not. Public reinforcement of gain-loss framing through a platform leaderboard of daily gainers and losers strengthens this association for stocks repeatedly classified as gainers, complementing evidence that making a purchase price more salient strengthens disposition-effect trading ([Frydman and Rangel 2014](#); [Frydman and Wang 2020](#)). The findings suggest that disposition-effect trading is not only a function of investors' embedded gains and losses, but also of when reference-dependent thinking is expressed in the market.

A second contribution is to the growing literature that uses investor-generated social media content to measure beliefs, disagreement, sentiment, and attention ([Cookson and Niessner 2020, 2023](#); [Cookson et al. 2026](#); [Gu, Hirshleifer, Teoh, and Wu 2025](#)). [Cookson and Niessner \(2020\)](#) establish StockTwits as a setting in which investor disagreement can be measured and has predictive content, and [Cookson et al. \(2026\)](#) show that social media sentiment informs corporate decisions. [Gu, Hirshleifer, Teoh, and Wu \(2025\)](#) construct a visual sentiment index from GIFs shared on StockTwits and show that it predicts market returns and trading activity. Relative to measures of optimism, pessimism, disagreement, or attention, I study a different object: reference-dependent sentiment, the tendency of gain/loss framing of positions relative to past price anchors. Ordinary sentiment captures whether investors are bullish or bearish about future returns; attention captures how intensively a security is discussed. Reference-dependent sentiment instead captures how investors encode current prices relative to those anchors. In this sense, the paper expands the use of social-media text from measuring beliefs about future prices to measuring how public reflection on past gains and losses predicts disposition-effect trading.

Third, this paper contributes to the literature on social finance. [Hirshleifer \(2020\)](#) articulates social finance as the study of how social processes shape economic thinking and behavior, with social transmission bias as a central building block: ideas and signals are systematically transformed as they pass from person to person. [Han, Hirshleifer, and Walden \(2022\)](#) develop this agenda formally, modeling social transmission bias among overconfident investors and showing that it can generate excess trading and return momentum. Empirical work in this tradition shows that social interaction

materially alters investment decisions. [Hong et al. \(2004\)](#) find that social households are more likely to participate in the stock market. [Escobar and Pedraza \(2023\)](#) show that peers' successful experiences raise entry into the stock market and lead investors to purchase similar and extreme-return stocks. [Heimer \(2016\)](#) provides the closest precedent for disposition trading, showing that access to an investment-specific peer network nearly doubles a trader's susceptibility to the disposition effect, consistent with impression management. Relative to work that emphasizes peer incentives or the transmission of beliefs and trades, I focus on the content of investment-specific discourse. This paper shows that discourse can be used to measure a form of aggregate sentiment, or mental-model salience, that has not been measured in previous literature: reference-dependent salience.

Fourth, this paper contributes to the literature on cognitive constraints in financial decision-making. Prior work shows that limited attention shapes how investors process information and how markets react to news ([Barber and Odean 2008](#); [Hirshleifer et al. 2009](#)). A growing literature explores the role of noisy or random behavior, which has been associated with patterns of global insensitivity to variation in choice parameters ([Enke and Graeber 2023](#); [Enke et al. 2024](#)). In contrast to that perspective, one reading of the evidence here is that investor behavior is not uniformly noisy but selectively sensitive: market returns and embedded positions create objective gain and loss states, yet investors need not perceive those states accurately, and reference-dependent discourse is more strongly associated with trading when its framing matches the gain or loss state of the position. Under this interpretation, reference-dependent sentiment is not only a tilt in public framing, but also a proxy for the accuracy of investors' gain-loss perceptions, so that periods of clearer reference-dependent talk coincide with greater responsiveness along the disposition channel.

Finally, this paper contributes to the literature on textual analysis in finance. Early work relies on word lists and bag-of-words methods ([Tetlock 2007](#); [Loughran and McDonald 2011](#)), which are well suited for measuring tone but less suited for constructs that depend on semantic context. More recent work uses contextual language models to classify sentiment in financial text ([Huang et al. 2023](#)) and applies large language models off the shelf to interpret news and forecast returns ([Lopez-Lira and Tang 2023](#)). Relative to dictionary methods and off-the-shelf LLM applications, I contribute

a teacher-student design for recovering an economically defined construct at corpus scale: an LLM labels a training sample under a detailed rubric, and a compact student classifier is then applied to the full corpus of 185.1 million posts. Dictionary methods cannot isolate reference-dependent framing because the relevant words are shared with ordinary sentiment and price forecasts, while direct LLM classification of the full corpus is infeasible. The approach illustrates more broadly how context-dependent economic constructs can be measured in large investor-generated text corpora.

The remainder of the paper is organized as follows. Section 2 describes the corpus, classification pipeline, and construction of GainTalk. Section 3 reports summary statistics for the market-day and stock-day analysis samples. Section 4 examines whether reference-dependent sentiment is associated with disposition-effect trading in the cross-section of stocks. Section 5 tests whether the same association appears in the aggregate market. Section 6 studies reference-dependent sentiment around earnings announcements and the associated calendar-time spread returns. Section 7 presents an event study of the platform shock to framing legitimacy. Section 8 concludes. Technical details and additional robustness checks are in the [Internet Appendix](#).

2 Data and Measure Construction

2.1 The StockTwits Corpus

I construct the reference-dependent sentiment measures from the full StockTwits archive. The platform is an investment-focused social network whose users are predominantly active retail traders discussing individual securities in real time. Prior work shows that StockTwits activity is informative for trading volume, volatility, and returns ([Cookson and Niessner 2020, 2023](#); [Gu, Hirshleifer, Teoh, and Wu 2025](#)).

I begin with 603,929,533 raw messages posted between July 2009 and September 2024. For each message I observe a unique identifier, the posting timestamp, the message text, and any cashtags it references. Following the single-security convention of [Cookson and Niessner \(2020\)](#) and [Cookson and Niessner \(2023\)](#), I retain only messages carrying exactly one cashtag, so that the gain/loss framing of a message can be assigned unambiguously to a single stock; this removes the 287.3 million messages with no cashtag and the 26.5 million that reference more than one. I further require

the cashtag to map to a CRSP common stock (share code 10 or 11) listed on the NYSE, AMEX, or NASDAQ (exchange code 1, 2, or 3) with a valid trading-day record on the posting date, which drops a further 100 million messages referencing indexes, ETFs, cryptocurrencies, or other non-CRSP tickers (e.g., \$SPY, \$QQQ, \$BTC). I then discard the 0.85 million messages with empty text. These filters leave 189,273,517 messages, or 31.3% of the raw message count.⁶

Because posting volume and ticker coverage are sparse before 2013, the analysis corpus is January 2013 to September 2024. It comprises 185,096,724 messages and covers 2,253 unique tickers on the average trading day.

2.2 Classification Design

To identify how reference-dependent thinking surfaces on social media, I design a prompt that instructs the LLM (GPT-4o) to label StockTwits message along the following three economically motivated dimensions.

The first dimension, *object* (HOLDER / NONE), identifies whether the message is about a holder’s position, as in *I’m up 80% from my \$110 entry* or *still underwater on my average*. What this dimension screens out is discourse that is not about a holder’s position: questions, links, general market commentary, price chatter that names no holder, and messages about a company’s products, sales, or other business metrics.

The second dimension, *frame* (GAIN / LOSS / NONE), identifies whether the message is gain framed or loss framed, that is, whether a holder’s position stands above or below a reference price. The typical examples are investors reporting how they have done: *glad I held through the dip, up 30% now* is gain framed, while *still down from where I bought* is loss framed. A message about a holder’s position that does not place it on either side is NONE, as in *flat from where I bought*. This dimension supplies the signed gain-versus-loss distinction that the sentiment measures are built around.

The third dimension, *anchor* (PAST / FUTURE / UNSPECIFIED), identifies whether the comparison is backward-looking or forward-looking, and it is what separates reference-dependent thinking from beliefs about future returns. A PAST message looks back to a price already established (*still below my cost basis from last year*); a FUTURE message

⁶Internet Appendix Table B.1 reports the sample-construction funnel step by step; Internet Appendix Table B.2 reports the platform’s posting volume and discourse composition across sub-periods.

looks ahead to a level not yet reached (*will hit \$200 by year end*) and carries no backward-looking reference-dependent content of the kind associated with the disposition effect, however optimistic it sounds. Many messages invoke no price anchor at all (*anyone watching AAPL today?*), and these are UNSPECIFIED. When a message mentions both a past outcome and a future expectation, it is coded PAST if it emphasizes the position’s achieved status and FUTURE if it explicitly centers on what will happen next. Without this dimension, gain-framed reference-dependent messages cannot be distinguished from optimistic price predictions. Figure 1 applies this rubric to five typical investor posts from the platform.

Insert Figure 1 near here.

I implement these dimensions with the two-stage pipeline, because running the LLM over the full corpus of 185.1 million messages is prohibitively costly. GPT-4o serves as the teacher: with detailed coding rubrics built on the definitions above, it labels a representative sample of 100,000 messages along all three dimensions.⁷ These labels then train a DeBERTa-v3-base student model, which is scalable enough to run over every message in the corpus. For each message the student returns soft probabilities across the labels in each dimension rather than a single label. Section 2.3 details how these probabilities are combined into the reference-dependent sentiment measures.

Table 1 illustrates the classifier on representative held-out messages. The messages in Panel A satisfy all three conditions at once: each is about a holder’s position, compares it with a past price, and frames that comparison as a gain or a loss. Those in Panel B fail at least one condition, and the failures are informative. Four are forward-looking price predictions, two gain-framed and two loss-framed, and two are undirected chatter that is not about a holder’s position. Together the two panels show what the measures are designed to capture: an investment-related gain/loss framing relative to a past price anchor.

Insert Table 1 near here.

⁷The full system prompt, dimensional definitions, worked examples, and JSON output schema are reproduced in Internet Appendix K.

2.3 Reference-Dependent Sentiment Construction

The three conditions have to hold together for a message to enter the reference-dependent sentiment measures, so I build those measures from the student's probabilities in a way that requires all three at once. GainMessage_m is the probability that message m is a gain message, meaning that it is about a holder's position, frames that position as a gain, and does so against a past price. It is the product of the three probabilities that carry these conditions,

$$\text{GainMessage}_m = P_m(\text{holder}) \cdot P_m(\text{gain}) \cdot P_m(\text{past}),$$

and LossMessage_m is defined identically, with $P_m(\text{loss})$ in place of $P_m(\text{gain})$. Internet Appendix D provides a self-contained summary of all variable definitions.

The product of the three probabilities, rather than their average, is what imposes the definition on the measure. The probability is high only when all three signals are present together, and a message missing any one of them is scaled near zero by the factor it lacks. Undirected chatter is not about a holder's position (*\$GME is this room dead or what*) and loses the holder factor; a bullish forecast carries no backward-looking anchor (*\$AAPL headed to 300*) and loses the past factor. Either way the message contributes almost nothing, however enthusiastically it talks about gains.

The market-day reference-dependent sentiment measures average these probabilities across the set $\mathcal{M}_t$ of N_t messages posted on day t . GainIntensity_t and LossIntensity_t are then the percentage of the day's messages that are gain and loss messages,

$$\begin{aligned} \text{GainIntensity}_t &= \frac{100}{N_t} \sum_{m \in \mathcal{M}_t} \text{GainMessage}_m, \\ \text{LossIntensity}_t &= \frac{100}{N_t} \sum_{m \in \mathcal{M}_t} \text{LossMessage}_m. \end{aligned} \tag{1}$$

My main variable is the balance between the two sides,

$$\text{GainTalk}_t = \frac{\text{GainIntensity}_t}{\text{GainIntensity}_t + \text{LossIntensity}_t}, \tag{2}$$

the share of the day's reference-dependent sentiment that is gain framed ($\text{GainTalk}_t \in [0, 1]$, reported in percent in the tables that follow). The distinction between GainTalk and the two intensities matters for what each one picks up. GainTalk picks up the direction of the conversation: whether reference-dependent discourse leans toward gains or toward losses, regardless of how much of it there is. It is therefore comparable

across days and stocks with very different posting volumes, and it is not mechanically higher on high-attention days when gain and loss messages both become more common. GainIntensity and LossIntensity pick up how much of the day’s conversation is about being up or being down relative to past prices. Because each is measured relative to all messages, it captures how widespread gain framing or loss framing is in the market.

For the stock-level tests (Section 4), I build the same measure one stock at a time. For each stock i on date t , I sum these probabilities over every message that mentions the stock and take the same gain-to-total ratio,

$$\text{GainTalk}_{i,t} = \frac{\sum_{m \in (i,t)} \text{GainMessage}_m}{\sum_{m \in (i,t)} \text{GainMessage}_m + \sum_{m \in (i,t)} \text{LossMessage}_m}. \quad (3)$$

I keep only stock-days with at least five classified messages, to avoid noisy estimates from thinly covered cells. When GainTalk enters the regressions below, at either the market-day or the stock-day level, I standardize it to mean zero and unit variance using the full-sample mean and standard deviation unless noted otherwise.

In form, GainTalk parallels the StockTwits net-sentiment construction in [Cookson et al. \(2026\)](#): that paper classifies posts as bullish or bearish with a Maximum Entropy classifier and defines stock-day forward sentiment as⁸

$$\text{Forward Sentiment}_{i,t} = \frac{N_{i,t}^{\text{bull}} - N_{i,t}^{\text{bear}}}{N_{i,t}^{\text{bull}} + N_{i,t}^{\text{bear}}}, \quad (4)$$

where $N_{i,t}^{\text{bull}}$ ($N_{i,t}^{\text{bear}}$) is the number of posts for stock i on day t classified as bullish (bearish). GainTalk replaces those poles with GainIntensity and LossIntensity and reports the share $G/(G+L)$ rather than the net $(G-L)/(G+L)$. The two formats are affine transforms of each other, $(G-L)/(G+L) = 2G/(G+L) - 1$, so after within-sample standardization they are identical. What differs is the content: GainTalk balances past-anchored talk about gains against losses, whereas forward sentiment balances forward-looking bullish against bearish labels. Because forward sentiment captures where posters expect the price to go rather than how they frame where it already stands, I use it as a control in the tests below.

⁸Text preprocessing of unlabeled messages follows [Cookson et al. \(2026\)](#).

The definition is selective by design. For most messages the most likely label is NONE on object, NONE on frame, and UNSPECIFIED on anchor. Messages whose most likely labels jointly identify a holder’s position, a definite frame, and a definite anchor account for 15.53 million posts, or 8.39% of the analysis sample; 8.84 million, or 4.78%, have a most likely gain or loss frame and a past anchor. Internet Appendix Table E reports this most-likely-label distribution. GainIntensity and LossIntensity average the per-message probability products over all posts. Table 2 (Panel A) summarizes the resulting daily series over April 2013 to September 2024: GainTalk averages 46.76% with a standard deviation of 6.31 percentage points, and GainIntensity and LossIntensity average 2.113% and 2.399%, respectively.

2.4 Validation

In the previous section, GainTalk is defined as the proportion of the day’s reference-dependent sentiment that is gain framed. The natural question is whether the measure is capturing what the construction intends. I check three things: whether the student model recovers the teacher’s labels on messages it has never seen, whether the messages it classifies as reference-dependent sentiment are distinguishable from messages that express ordinary optimism or pessimism, and whether the daily series moves with markets in the way the mechanism implies.

The first check is out-of-sample accuracy. The LLM labels a corpus of about 100,000 messages. I set aside 10% of them (9,995 messages) for evaluation and train the DeBERTa-v3-base student model on the rest. I then compare the student model’s labels with the teacher’s on those 9,995 messages. The student model recovers the teacher’s object label 88.0% of the time, the frame label 86.9% of the time, and the anchor label 88.9% of the time, and it gets all three right at once for 75.2% of messages. One caveat is that these reference labels come from GPT-4o rather than from human annotation, so agreement with the teacher would overstate performance if the teacher were itself noisy. A random sample of 1,000 messages was therefore manually reviewed on all three heads. The teacher and the human reviewer assigned the same object, frame, and anchor labels simultaneously in 79.2% of them, and the student matched the human on all three heads in 76.8% of them. Full performance metrics and the human-audit table are in Internet Appendix C.

Accuracy alone does not establish that the classifier is recovering reference-dependent thinking. To check that, I rank unigrams and bigrams by prevalence within each of four categories, using a random sample of 2 million messages from the classified corpus: *Gain-framed* (gain frame, a holder’s position, past anchor), *Loss-framed* (loss frame, a holder’s position, past anchor), and the future-anchored counterparts (gain or loss frame, future anchor). As Internet Appendix Table F shows, the most salient gain-framed phrases, including *in the green*, *currently up*, *am up*, *finally green*, and *in profit*, express that the writer is evaluating a held position as a gain relative to a past entry price. The top loss-framed phrases, including *am down*, *currently down*, *underwater*, *bag holder*, and *at loss*, express that the writer is in a loss state relative to past entry prices. Future-anchored phrases, by contrast, are directional outlook language (*green soon*, *explode soon*, *higher soon*, *moon tomorrow*, *dump coming*, *fall hard*, *red soon*, *crash soon*). The gain-framed and loss-framed lists are position-relative and past-anchored; the future-anchored lists are outlook language about where the price is headed. What the classifier is picking out, then, is backward-looking evaluation of a holder’s position against a past price, not generic market optimism or pessimism.

Insert Figure 2 near here.

The final check asks whether the reference-dependent sentiment tracks changes in investors’ gain/loss composition over time. Figure 2 plots the daily GainTalk z-score together with trailing CRSP value-weighted market returns at the 3-, 21-, 63-, and 252-trading-day horizons. GainTalk is defined in equation (2) and is standardized to zero mean and unit variance over the full sample. Pearson correlations between the GainTalk z-score and the four trailing-return series are 0.40, 0.28, 0.21, and 0.25, respectively ($N = 2,737$ StockTwits trading days). The co-movement is strongest at the shortest horizon and declines as the return window lengthens, so GainTalk tracks recent market gains and losses more closely than longer-run cumulative returns.

3 Descriptive Statistics

3.1 Market variables

The market variables are constructed from the CRSP daily stock file. Daily market return is the CRSP value-weighted market return (percentage points). Daily turnover rate is aggregate dollar trading volume divided by aggregate market capitalization, expressed as a percentage. It is the market activity series underlying abnormal log market turnover, the outcome in the daily market-level regressions of Section 5. Table 2 reports summary statistics for the reference-dependent discourse and sentiment measures. Internet Appendix Table G reports trading outcomes and the lagged return and volatility controls used in the daily regressions.

Insert Table 2 near here.

The discourse series are moderately persistent. Daily GainTalk has a first-order autocorrelation of 0.70, decaying to roughly half its level within two trading days, and GainIntensity is more persistent (autocorrelation 0.85) than LossIntensity (0.54). Internet Appendix Table I.6 shows that inference in the market-level tests is robust to Newey-West lags of up to 63 days and to a moving-block bootstrap.

3.2 Stock variables

The cross-sectional tests of Section 4 use a stock-day panel. I aggregate messages to the stock-day level, retain stock-days with at least five classified posts, and merge each covered stock-day with its CRSP trading record and quote-midpoint retail sell flow (Barber et al. 2024). The estimation sample is April 2015 to September 2024 and contains 1,891,685 stock-days. Table 2 (Panel B) reports summary statistics; Internet Appendix Table G (Panel B) reports trading outcomes and controls.

The covered stock-days account for 58.9% of aggregate CRSP common-stock market capitalization on the average trading day. Coverage is tilted toward larger firms, as expected for a retail-discussion platform: retail attention concentrates on familiar, liquid names with broad ownership and media coverage, so large stocks generate more messages and appear more often in the sample.

GainTalk is dispersed at both levels of the design. At the market day, the standard deviation is 6.31 percentage points (Table 2, Panel A), so a one-standard-deviation move shifts the gain share of reference-dependent discourse by about 6.3 pp. At the stock day, the standard deviation rises to 34.74 percentage points (Table 2, Panel B), suggesting that GainTalk is substantially more dispersed across stocks than across market days. Part of the stock-day dispersion reflects the small number of reference-dependent messages behind many stock-days. Such sampling noise attenuates the stock-level estimates toward zero rather than generating them: on stock-days with at least one full reference-dependent message, the GainTalk $\times$ CGO interaction is larger than in the full panel (Internet Appendix Table H.3).

4 Reference-Dependent Sentiment and Stock-Level Trading Activity

In this section, I study whether reference-dependent sentiment contains information about trading in the cross-section of individual stocks: whether stock-level GainTalk is associated with retail sell flow, and whether that association is stronger where capital gain overhang is higher. Two measurement choices matter for that cross-sectional test. First, the gain/loss state should be specific to each stock's cost-basis position: what matters is whether a given stock's existing shareholder base has an unrealized gain or loss. Second, trading should be signed. Disposition-effect trading is asymmetric on the sell side (realizing winners; holding losers), so a direct test should speak to sell flow rather than to total volume alone. This section therefore uses stock-specific capital gain overhang as a proxy for the prevailing gain/loss state and stock-specific retail sell flow as a proxy for selling intensity.

4.1 Capital Gain Overhang and Retail Sell Flow

Following Grinblatt and Han (2005), I measure each stock's embedded gain or loss with *capital gain overhang* (CGO), the unrealized return of the current price relative to a turnover-weighted reference purchase price. For each CRSP common stock i , the

reference price evolves as

$$\text{RP}_{i,t} = V_{i,t-1} \tilde{P}_{i,t-1} + (1 - V_{i,t-1}) \text{RP}_{i,t-1}, \quad (5)$$

where $\tilde{P}_{i,t}$ is the split-adjusted price, and $V_{i,t} = \text{vol}_{i,t} / \text{shrout}_{i,t}$ is the daily share turnover rate. I compute the recursion on daily CRSP data beginning in January 2008, initializing each stock's reference price at its first observed split-adjusted price.⁹ Stock-level CGO is then

$$\text{CGO}_{i,t} = \frac{\tilde{P}_{i,t} - \text{RP}_{i,t}}{\tilde{P}_{i,t}}. \quad (6)$$

$\text{CGO}_{i,t} > 0$ (< 0) indicates that the average invested dollar in stock i is held at an unrealized gain (loss). Following [Frazzini \(2006\)](#), capital gain overhang is the natural stock-level state for disposition-effect trading: CGO summarizes the gain or loss domain faced by investors in the stock more directly than a same-day return.

In the stock-level test, the outcome is retail sell flow. Using quote-midpoint retail sell volume ([Barber et al. 2024](#)), I measure retail sell shares divided by shares outstanding and express that ratio in basis points,

$$\text{SellFlow}_{it} = 10,000 \times \frac{\text{MRSVol}_{it}}{\text{SHROUT}_{it}},$$

where MRSVol_{it} is quote-midpoint retail sell volume, in shares, and SHROUT_{it} is shares outstanding. Abnormal log retail sell flow is the deviation of $\log(1 + \text{SellFlow}_{it})$ from its mean over the 121 trading days in $[t-140, t-20]$,

$$\text{AbnLogSell}_{it} = \log(1 + \text{SellFlow}_{it}) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(1 + \text{SellFlow}_{is}),$$

with the window borrowed directly from [Cookson and Niessner \(2020\)](#). Because this window uses the 140 trading days before t , and the construction requires at least 60 valid baseline days, the stock-day estimation sample is April 2015 to September 2024. On a stock-day with sell flow of s basis points, a difference of β in AbnLogSell corresponds to a difference of about $\beta(1+s)$ basis points in SellFlow . Relative to total trading volume, this signed retail series isolates the sell-side activity through which the disposition effect operates: realizing embedded gains raises sell flow, while reluctance to

⁹The weights in (5) already sum to one. By the April 2015 start of the stock-day panel, names present from 2008 have seven years of daily updating. For stocks that list after January 2008, the first observed price is close to the purchase price of the initial shareholder base.

realize embedded losses lowers it. Stock-day GainTalk is the gain share of reference-dependent discourse among posts that mention the stock, constructed as in Section 2 (equation (3)).

4.2 GainTalk and Retail Sell Flow

I ask whether the association between GainTalk and retail sell flow is stronger where capital gain overhang is higher. I estimate

$$\begin{aligned} \text{AbnLogSell}_{it} = & \alpha + \beta_1 \text{GainTalk}_{it} + \beta_2 \text{CGO}_{it} + \delta (\text{GainTalk}_{it} \times \text{CGO}_{it}) \\ & + X'_{it}\gamma + \text{FE} + \varepsilon_{it}, \end{aligned} \quad (7)$$

where GainTalk and CGO are standardized using the full-sample mean and standard deviation. The coefficient of interest is δ : a positive value means that the association between GainTalk and retail selling is stronger where capital gain overhang is higher. The controls X_{it} include lagged abnormal log retail sell flow, cumulative stock returns over $t-5$ to $t-1$ and $t-21$ to $t-6$, Volatility ($t-5$ to $t-1$), log social-media volume, forward sentiment (Cookson et al. 2026), RavenPack sentiment, and RavenPack log news count, together with the same-day stock return. These controls absorb the trading predictors emphasized in Cookson and Niessner (2020) as well as widely used proxies for social-media attention, social-media sentiment, and traditional-news sentiment and intensity. Columns (1) and (2) omit the interaction, and columns (3) and (4) include it. Columns (1) and (3) include no fixed effects, and columns (2) and (4) include firm and date fixed effects. Columns (5) and (6) replace GainTalk with before-market-open (BMO) GainTalk and overnight GainTalk and include firm and date fixed effects. Following the convention of Cookson and Niessner (2023), BMO GainTalk uses posts from 06:00 to 09:30 ET on day t , and overnight GainTalk uses posts from the 16:00 ET close on day $t-1$ to the 09:30 ET open on day t , so both are fixed before day- t trading begins.

Insert Table 3 near here.

Table 3 reports the estimates. Columns (1) and (2) omit the interaction. With firm and date fixed effects (column (2)), a one-standard-deviation rise in GainTalk is associated with about 1.5 percent higher abnormal retail sell flow ($t = 21.98$), about 1.8

percent of the standard deviation of abnormal log retail sell flow (0.852; Internet Appendix Table G).

One finding worth noting is that the forward sentiment coefficients are negative in every column, which is the opposite sign from GainTalk and helps clarify what the two measures pick up. When investors form a bullish belief about a stock they hold, a common response is to wait for that belief to materialize, so stronger forward-looking bullishness can coincide with less selling. However, when the belief turns bearish, the natural response is to unwind in order to avoid a perceived upcoming loss, which coincides with more selling. Reference-dependent sentiment operates in the opposite direction. Gain-framed sentiment is associated with more selling, consistent with a greater propensity to realize gains, whereas loss-framed sentiment is associated with less selling, consistent with greater reluctance to realize losses. Therefore, the two concepts have distinct theoretical underpinnings and empirical manifestations.

Taken together, the first two columns show that after controlling for forward sentiment (Cookson et al. 2026), social-media attention, and traditional news sentiment and intensity, the GainTalk coefficient remains positive and significant, suggesting that reference-dependent sentiment captures a meaningful dimension beyond existing sentiment and attention measures.

The interaction δ is positive and significant at the one percent level in columns (3)–(6). With firm and date fixed effects (column (4)), $\delta = +0.0143$ ($t = 19.26$). In column (4), the association between GainTalk and retail selling is stronger where capital gain overhang is higher: a one-standard-deviation increase in CGO raises the GainTalk coefficient from 0.0152 to 0.0295. In magnitude, a one-standard-deviation increase in GainTalk is associated with abnormal retail sell flow higher by about 1.5 percent at mean overhang and by about 3.0 percent where overhang is one standard deviation above its mean. The interaction of 0.0143 is about 1.7 percent of that standard deviation. Before-market-open GainTalk $\times$ CGO is +0.0167 ($t = 16.40$) in column (5), and overnight GainTalk $\times$ CGO is +0.0121 ($t = 15.30$) in column (6).

To separate selling pressure from buying pressure, Internet Appendix Table H.1 replaces the outcome with abnormal net retail flow, sell flow minus buy flow relative to its trailing mean. With firm and date fixed effects, GainTalk is associated with higher net selling (+0.0355, $t = 4.19$), and the GainTalk $\times$ CGO interaction remains positive

(+0.0289, $t=4.05$).

The before-market-open and overnight estimates lead to the same conclusion. Both interactions are positive and significant, and the before-market-open interaction (+0.0167) is larger than the same-day estimate in column (4). Because both measures are fixed before day- t trading begins, the association is unlikely to reflect discourse that responds to same-day trading.

One interpretation is that capital gain overhang measures an objective embedded gain or loss, but investors need not perceive it accurately. Gain-framed discourse may be elevated precisely when investors recognize that their positions sit in the gain domain. GainTalk is then more strongly associated with selling in stocks where gains are in fact embedded, coinciding with a stronger disposition-effect pattern. Under this interpretation, reference-dependent discourse proxies for the accuracy of investors' gain-loss perceptions.

Another plausible interpretation is that capital gain overhang places each stock's shareholder base in a gain or loss domain, while reference-dependent discourse captures whether investors socially frame that position. Investors who see others discussing being up, being back to green, or locking in gains on a stock may also be evaluating their own positions in it through reference points. That social cue can coincide with a stronger disposition-effect pattern and with more selling where gains are embedded. This interpretation is grounded in the social transmission of reference-dependent thinking.

The two interpretations are not mutually exclusive: investors may perceive their gain-loss positions more accurately and also take cues from how others frame theirs. The influencer evidence in Internet Appendix Table I.8 further points to a social channel. At the market level, GainTalk of influencers, whose framing reaches more investors, is more strongly associated with abnormal turnover than GainTalk of other posters, and overnight influencer GainTalk predicts other posters' GainTalk after the open more strongly than the reverse. Whether the mechanism runs through the social transmission of reference-dependent thinking or through more accurate gain-loss perception, the common implication is that disposition-style trading is not constant but varies systematically with reference-dependent sentiment measured in observable public discourse.

4.3 Gain- and Loss-Framing Sentiment by Capital Gain Overhang

The interaction in Section 4.2 pools the two sides of reference-dependent discourse into one coefficient, and GainTalk itself is a ratio, so a high value can reflect either more talk about gains or less talk about losses. A sign-split exercise separates those sides. Reference-dependent sentiment predicts something more specific than a single slope on CGO: talking about gains should be associated with more selling where the shareholder base sits on an unrealized gain, and talking about losses with less selling where it sits on an unrealized loss, with each side largely inactive in the other's state.

I therefore replace stock-day GainTalk with the stock-day analogues of GainIntensity_{it} and LossIntensity_{it} in equation (1), computed over the messages that mention stock *i* on date *t* and standardized over the full sample. I estimate

$$\begin{aligned} \text{AbnLogSell}_{it} = & \alpha + \lambda_+ (\text{GainIntensity}_{it} \times D_{it}) + \phi_+ (\text{GainIntensity}_{it} \times (1 - D_{it})) \\ & + \phi_- (\text{LossIntensity}_{it} \times D_{it}) + \lambda_- (\text{LossIntensity}_{it} \times (1 - D_{it})) \quad (8) \\ & + \eta D_{it} + \rho_+ \text{CGO}_{it}^+ + \rho_- \text{CGO}_{it}^- + X'_{it}\gamma + \text{FE} + \varepsilon_{it}, \end{aligned}$$

where λ_+ and λ_- are the matched-state slopes of gain and loss intensity, ϕ_+ and ϕ_- are the cross-state slopes, $D_{it} = 1[\text{CGO}_{it} > 0]$ equals one when capital gain overhang is positive, and $\text{CGO}_{it}^+ = \max(\text{CGO}_{it}, 0)$ and $\text{CGO}_{it}^- = \min(\text{CGO}_{it}, 0)$ are a linear spline in raw CGO with a knot at zero. The indicator and the spline use raw CGO, not the standardized CGO in equation (7), and η , ρ_+ , and ρ_- absorb the level and the CGO shape of each state. The controls X_{it} follow Table 3. Columns (1) and (2) use same-day abnormal log retail sell flow, and columns (3) and (4) use the next day's outcome. Within each outcome the fixed effects are none, then firm and date. Reference-dependent sentiment predicts $\lambda_+ > 0$ and $\lambda_- < 0$, with the off-diagonal slopes ϕ_+ and ϕ_- near zero.

Insert Table 4 near here.

Table 4 reports the estimates. I focus on columns (2) and (4), which add firm and date fixed effects: column (2) uses same-day abnormal log retail sell flow and column (4) uses the outcome on $t+1$.

For same-day selling, the spline in column (2) controls for the sign and the size of capital gain overhang: $\max(\text{CGO}_{it}, 0)$ is +0.3665 ($t = 18.38$) and $\min(\text{CGO}_{it}, 0)$ is -0.1338 ($t = 20.92$), so abnormal selling rises as embedded gains grow and also rises

as embedded losses deepen. The intensity interactions then measure how selling co-varies with gain-framing and loss-framing intensity inside each state. The significant terms line up with matched states, as the mechanism predicts. Gain-framing intensity matters only in the gain state: $\text{GainIntensity}_{it} \times 1[\text{CGO}_{it} > 0]$ is $+0.0126$ ($t = 10.30$), while the same intensity in the loss state is $+0.0014$ ($t = 0.90$) and insignificant. Loss-framing intensity matters only in the loss state: $\text{LossIntensity}_{it} \times 1[\text{CGO}_{it} \leq 0]$ is -0.0165 ($t = 8.88$), while its interaction with the gain state is -0.0014 ($t = 1.55$). In magnitude, one standard deviation of gain-framing intensity in a gain stock is associated with abnormal log retail sell flow higher by about 1.5 percent of its standard deviation, and one standard deviation of loss-framing intensity in a loss stock with abnormal log retail sell flow lower by about 1.9 percent of its standard deviation.

For selling on $t+1$, column (4) shows the same matched-state pattern. $\text{GainIntensity}_{it} \times 1[\text{CGO}_{it} > 0]$ is $+0.0064$ ($t = 7.25$) and $\text{LossIntensity}_{it} \times 1[\text{CGO}_{it} \leq 0]$ is -0.0025 ($t = 2.99$); the cross-state interactions remain small (-0.0001 , $t = 0.09$; and $(-0.0011, t = 1.12)$). Relative to column (2), the gain-state coefficient is about half as large and the loss-state coefficient is smaller still, so the association is not confined to the day of the discourse but does decay. Columns (1) and (3), which omit fixed effects, deliver the same matched-state signs for the two outcomes ($+0.0151$ and -0.0111 with $t = 11.31$ and $t = 9.35$; $+0.0086$ and -0.0070 with $t = 8.40$ and $t = 7.32$).

Overall, the sign-split separates what the linear $\text{GainTalk} \times \text{CGO}$ interaction averages together. Gain framing is associated with more retail selling where capital gain overhang is positive, and loss framing with less retail selling where it is not. Gain framing is economically small in the loss state, and loss framing is economically small in the gain state. The pattern is therefore not an artifact of a ratio that confounds more gain talk with less loss talk, nor of a single coefficient that averages across gain and loss stocks.

5 Reference-Dependent Sentiment and Aggregate Trading Activity

The tests so far are in the cross-section of stocks. In this section, I study whether reference-dependent sentiment also contains information about market-wide trading

activity. I begin with the association between GainTalk and the level of abnormal turnover, then interact GainTalk with the market return, and finally separate gain-framed and loss-framed intensity by the sign of the return. The sequence moves from a level relation to a test of state dependence that matches the predictions of reference-dependent preferences, with the market return as a proxy for the prevailing gain/loss state and abnormal turnover as a proxy for trading intensity.

5.1 GainTalk and Abnormal Market Turnover

The dependent variable is abnormal log market turnover. Let TO_t denote aggregate dollar volume divided by market capitalization. I define

$$\text{AbnLogTO}_t = \log(TO_t) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(TO_s),$$

so that day- t log turnover is measured relative to its own historical moving average over the 121 trading days from $t-140$ to $t-20$, the same [Cookson and Niessner \(2020\)](#) window as for abnormal log retail sell flow in Section 4. The trailing baseline removes slow-moving trends in market activity and isolates abnormal trading intensity on day t . Because this window uses the 140 trading days before t , the daily estimation sample is April 2013 to September 2024 ($N = 2,737$ trading days).¹⁰ While aggregate turnover mixes buys and sells, it remains the natural market-level outcome: realizing a winner is a trade, holding a loser is not. A stronger disposition-effect pattern therefore shows up as greater volume. Section 4 uses retail sell flow, where the selling margin is measured more directly.

The key independent variable is daily market-level GainTalk from equation (2). I estimate

$$\text{AbnLogTO}_t = \alpha + \beta_1 \text{GainTalk}_t + X_t' \gamma + \varepsilon_t, \quad (9)$$

where X_t contains the market-level counterparts of the controls in equation (7): the same-day market return, lagged abnormal log turnover, cumulative market returns over $t-5$ to $t-1$ and $t-21$ to $t-6$, Volatility ($t-5$ to $t-1$), log social-media volume, forward sentiment, RavenPack sentiment, and RavenPack log news count.

¹⁰The daily series used to form those lags begins in September 2012. Counting 140 trading days forward from that date yields the first complete baseline window in April 2013.

I estimate the specification on the CRSP universe and on S&P 500 constituents, with year-quarter and day-of-week fixed effects, and report t -statistics based on Newey-West standard errors with five lags. The coefficient of interest is β_1 , the association between GainTalk and abnormal turnover after controlling for the variables in X_t .

Insert Table 5 near here.

Table 5 reports the estimates. Columns (1)–(2) estimate equation (9). In the CRSP universe with year-quarter and day-of-week fixed effects (column (1)), a one-standard-deviation rise in GainTalk is associated with about 3.1 percent higher abnormal turnover ($t=6.71$). Column (2) delivers a similar estimate for S&P 500 constituents, also significant at the one percent level. Given that the average daily market turnover is about 75 basis points in my sample, that is about 2.3 basis points of additional turnover. Relative to the standard deviation of abnormal log turnover (0.190; Table G), the column (1) estimate of 0.0312 is about one-sixth of a typical daily move in the outcome.

As in the stock-level tests, forward sentiment enters with the opposite sign from GainTalk in every column.

The natural next question is whether the association between GainTalk and turnover is stronger in gain states. I take the same-day market return as a proxy for the gain/loss state: a higher return is, on average, a day on which more positions sit in the gain domain. Columns (3)–(4) therefore add the interaction between GainTalk and the market return,

$$\text{AbnLogTO}_t = \alpha + \beta_1 \text{GainTalk}_t + \beta_2 R_{m,t} + \delta (\text{GainTalk}_t \times R_{m,t}) + X_t' \gamma + \varepsilon_t, \quad (10)$$

with the same controls and with year-quarter and day-of-week fixed effects in both columns. As with the GainTalk $\times$ CGO interaction in equation (7), a positive δ means that the association between GainTalk and abnormal turnover is stronger on days with higher market returns.

Columns (3) and (4) report estimates of equation (10). The interaction coefficient δ is positive and significant at the one percent level in both columns: +0.0120 ($t=4.06$) for the CRSP universe and +0.0118 ($t=3.72$) for S&P 500 constituents. The association between GainTalk and turnover is therefore stronger on up-market days. In column (3), each percentage point of market return adds about 1.2 percent to the abnormal turnover

associated with a one-standard-deviation increase in GainTalk, so that on a day with a 1 percent market gain, a one-standard-deviation increase in GainTalk is associated with abnormal turnover higher by about 4.1 percent. Given the average daily market turnover of about 75 basis points, the interaction adds about 0.9 basis points of turnover on such a day, for a total of about 3.1 basis points. Relative to the standard deviation of abnormal log turnover (0.190; Table G), the interaction of 0.0120 is about 6 percent of a typical daily move in the outcome, and the combined estimate of 0.0414 is about one-fifth. The GainTalk level coefficient remains positive and significant, so the interaction does not displace the level association documented in columns (1)–(2).

The two interpretations in Section 4.2, social transmission of reference-dependent thinking and more accurate gain-loss perception, carry over with the market return in place of capital gain overhang. Either way, the association between GainTalk and turnover depends on the prevailing gain/loss state.

Columns (5) and (6) replace abnormal log turnover with abnormal log aggregate retail sell flow and keep the GainTalk $\times$ market-return interaction. RetailSell is quote-midpoint retail sell dollar volume divided by aggregate market capitalization, and AbnLogSell uses the same $[t-140, t-20]$ window as AbnLogTO. Column (5) uses the CRSP universe and column (6) uses S&P 500 constituents. The specification otherwise matches columns (3)–(4), including year-quarter and day-of-week fixed effects, with the lagged outcome equal to lagged AbnLogSell. The interaction is +0.0191 ($t=6.02$) in the CRSP universe and +0.0197 ($t=5.64$) among S&P 500 constituents. The GainTalk level coefficient remains positive: +0.0344 ($t=8.34$) and +0.0237 ($t=5.31$).

5.2 Gain- and Loss-Framing Sentiment by Market State

I next repeat the sign-split exercise of Section 4.3 at the market level, with the market return in place of capital gain overhang and abnormal turnover in place of retail sell flow. I replace the GainTalk ratio with GainIntensity $_t$ and LossIntensity $_t$ from equation (1), each standardized over the full sample, and estimate

$$\begin{aligned} \text{AbnLogTO}_t = & \alpha + \lambda_+ (\text{GainIntensity}_t \times D_t) + \phi_+ (\text{GainIntensity}_t \times (1 - D_t)) \\ & + \phi_- (\text{LossIntensity}_t \times D_t) + \lambda_- (\text{LossIntensity}_t \times (1 - D_t)) \\ & + \eta D_t + \rho_+ R_t^+ + \rho_- R_t^- + X_t' \gamma + \varepsilon_t, \end{aligned} \quad (11)$$

where the coefficients are as in equation (8), $D_t = 1[R_{m,t} > 0]$ equals one when the market return is positive, and $R_t^+ = \max(R_{m,t}, 0)$ and $R_t^- = \min(R_{m,t}, 0)$ are a linear spline in the market return with a knot at zero. The controls X_t are the same as in equation (9). Within each turnover universe the fixed effects vary from none to year-quarter to year-quarter and day-of-week. The predictions are as in Section 4.3: $\lambda_+ > 0$ and $\lambda_- < 0$, with ϕ_+ and ϕ_- near zero.

Insert Table 6 near here.

Table 6 reports the estimates. I focus on column (3), the CRSP universe with year-quarter and day-of-week fixed effects. The linear spline in the market return controls for the sign and magnitude of the return. The GainIntensity and LossIntensity interactions then measure how abnormal turnover covaries with gain-framing and loss-framing intensity within up and down markets.

The significant interactions line up with matched states, as the mechanism predicts. Gain-framing intensity is significant only in gain states: $\text{GainIntensity}_t \times 1[R_{m,t} > 0]$ is $+0.0091$ ($t=2.55$), while the same measure interacted with the loss state is $+0.0069$ ($t=1.64$) and insignificant. Loss-framing intensity is significant in loss states: $\text{LossIntensity}_t \times 1[R_{m,t} \leq 0]$ is -0.0158 ($t=3.56$), while its interaction with the gain state is -0.0065 ($t=1.63$).

In magnitude, one standard deviation of gain-framing intensity in an up market is associated with abnormal turnover higher by about five percent of its standard deviation, and one standard deviation of loss-framing intensity in a down market with abnormal turnover lower by about eight percent of its standard deviation. The remaining columns deliver the same matched-state pattern.

Overall, the market-level sign-split separates what the pooled GainTalk interaction averages together. Separating the two framing intensities shows that the association with abnormal turnover is asymmetric and state dependent: gain framing matters when the market return is positive, loss framing when it is negative. Gain framing is insignificant in loss states, and the loss-framing slope is smaller in gain states than in loss states. Moreover, this structure helps distinguish reference-dependent thinking from a general attention story. If social discourse simply captured broad attention, one might

expect higher trading regardless of the market state. A reference-dependent mechanism instead predicts a direction-specific pattern: gain framing should be associated with stronger gain realization, while loss framing should be associated with a stronger tendency to hold losers. The matched-state estimates line up with that prediction.

5.3 Overnight GainTalk: Pre-Determined Robustness

Table 5, columns (3)–(4), shows that the association between GainTalk and turnover is stronger on days with higher market returns. That interaction is estimated within day t , and because GainTalk is built in part from posts written during the trading session, it may partly respond to same-day trading rather than precede it. As in Section 4.2, I therefore replace same-day GainTalk with BMO and overnight GainTalk, which are fixed before day- t trading begins. Table 7 substitutes these measures for contemporaneous GainTalk and interacts them with either the full-day market return or the overnight return on the S&P 500 ETF (SPY); columns (6)–(7) further horse-race overnight GainTalk against overnight social-media and news volume.

Insert Table 7 near here.

The interaction remains positive and significant once GainTalk is measured before the open. In column (1), BMO GainTalk interacted with the full-day return yields $+0.0087$ ($t = 2.79$). The association is not confined to same-day turnover: column (2) shows that the association between BMO GainTalk and next-day abnormal turnover is also stronger when the return is higher ($+0.0065$, $t = 1.78$). When the return state is aligned with the pre-open window, column (3) shows that BMO GainTalk $\times$ overnight return equals $+0.0287$ ($t = 5.78$). Overnight GainTalk $\times$ overnight return equals $+0.0259$ ($t = 5.17$) in column (4), and overnight GainTalk $\times$ the full-day return equals $+0.0068$ ($t = 2.11$) in column (5). Attention placebos do not absorb the GainTalk interaction. Overnight message volume interacted with the return is insignificant, and overnight news volume interacted with the return is significant only at the ten percent level, while overnight GainTalk $\times$ return remains significant in both columns ($+0.0081$, $t = 2.49$; $+0.0074$, $t = 2.27$).

Overall, the overnight evidence indicates that GainTalk measured before the open is

associated with abnormal turnover later in the day, and that this association is stronger on days with higher market returns.

5.4 Placebo Test: Reference-Dependent Sentiment versus Competing Channels

The previous tests show that GainTalk is associated with higher abnormal turnover and that this association is stronger on days with higher market returns. A natural concern is that these patterns simply pick up other sentiment, news, or attention measures rather than reference-dependent sentiment.

Forward-looking social-media sentiment and social-media attention, together with traditional news tone and news intensity, all co-move with market conditions and with trading activity. Earlier specifications already include several of these variables as level controls, but that leaves open whether their interactions with the market return can account for the GainTalk $\times$ return coefficient that Section 5.1 emphasizes.

Table 8 therefore puts these alternatives into the same interaction design. The outcome remains same-day abnormal log market turnover on the CRSP universe with year-quarter and day-of-week fixed effects. Column (1) interacts forward sentiment alone with the market return. Columns (2)–(5) introduce GainTalk and then cumulatively add forward sentiment, RavenPack news sentiment, RavenPack news count, and social-media volume, each as a main effect and interacted with the return. The object of interest is whether $\hat{\delta}^{\text{GT}}$ survives once these alternatives are allowed to interact with the market return alongside GainTalk.

Insert Table 8 near here.

The GainTalk interaction remains positive and significant. In the most stringent specification (column (5)), $\hat{\delta}^{\text{GT}} = +0.0134$ ($t = 3.40$), close in magnitude to the preferred estimate in Table 5, column (3). The alternative measures interacted with the return do not reproduce a consistent pattern: forward sentiment $\times$ return is indistinguishable from zero in the most stringent specification, RavenPack sentiment $\times$ return is insignificant, and social-media volume is associated with a higher level of abnormal turnover, but that association does not vary with the return. News count carries

a smaller interaction that is significant only at the ten percent level in column (4) and insignificant in column (5), so it does not displace GainTalk.

These horse races clarify the interpretation of the earlier results. The stronger association between GainTalk and turnover on up-market days is not an artifact of bullish forward-looking beliefs, news tone, news intensity, or attention. What matters is the reference-dependent composition of discourse: how strongly the day's conversation is tilted toward gains rather than losses.

6 Earnings News, GainTalk, and Disposition Spreads

The evidence so far concerns trading activity. Sections 4 and 5 show that reference-dependent sentiment speaks to which stocks attract retail sell flow and to when market turnover is high. Those tests speak to quantities, and quantities need not appear in prices: if liquidity provision absorbs the associated order flow, volume can vary without returns moving. Whether reference-dependent sentiment carries information about asset prices is therefore a separate question, which this section addresses.

Following [Frazzini \(2006\)](#), at each month-start I sort stocks independently into quintiles on earnings news and on pre-announcement GainTalk. The news variable is the market-model cumulative abnormal return around the most recent earnings announcement, $CAR[-2, +1]$, a return-based surprise that requires no assumption about the market's earnings expectation. The market model is estimated over the 200 trading days from -210 to -11 before the announcement. Pre-announcement GainTalk is stock-day GainTalk averaged over the 60 trading days ending the day before that announcement, so it summarizes how the stock's investor base framed its position going into the event. Portfolios are equally weighted, held for $K = 1, 2, 3, 6$, and 12 months with overlapping vintages, and priced in calendar time on the Fama-French five factors.

Unlike [Frazzini \(2006\)](#), where capital gain overhang infers the embedded gain from past prices and turnover, GainTalk captures how investors frame that position in discourse. The same state-dependent pattern then implies a sign-specific prediction. When unexpected good news arrives and pushes the stock further into the gain domain, stocks with high pre-announcement GainTalk should exhibit stronger positive post-announcement drift. Conversely, when unexpected bad news arrives and pushes the

stock further into the loss domain, stocks with low pre-announcement GainTalk (high LossTalk) should exhibit stronger negative post-announcement drift. In both cases, the disposition-effect trading documented in Sections 4 and 5 works against the news: gain-framed holders sell into good news, and loss-framed holders hold on through bad news, so prices adjust slowly. A portfolio long good-news/high-GainTalk stocks and short bad-news/low-GainTalk stocks should therefore earn positive abnormal returns. The reverse sort should not.

Insert Table 9 near here.

Table 9 reports monthly five-factor alphas. Pre-announcement GainTalk separates stocks with very different post-announcement returns. The high-GainTalk spread, which holds good-news stocks in the top GainTalk quintile and shorts bad-news stocks in the bottom GainTalk quintile for three months, earns a five-factor alpha of 1.031% per month ($t = 3.64$). The low-GainTalk spread, which reverses the GainTalk sort, earns -0.078% ($t = 0.29$), not significantly different from zero. The results are consistent with slow adjustment after news when pre-announcement GainTalk and the sign of the news agree.

The magnitudes are economically large and decay with the holding period. A monthly alpha of 1.031% corresponds to about 12.4 percent annualized. The high-GainTalk spread falls monotonically from $+1.201\%$ at one month ($t = 3.64$) to $+0.713\%$ at twelve months ($t = 3.02$). The high-GainTalk long-short alpha exceeds the low-GainTalk long-short alpha by $+1.191\%$ at one month ($t = 2.74$) and by $+0.792\%$ at twelve months ($t = 2.28$). That decay is consistent with an adjustment that completes gradually rather than with a permanent premium, and the five-factor adjustment means that the high-GainTalk spread is not accounted for by standard factor exposures.¹¹ The result is robust to value weighting (Internet Appendix Table J.2).

One alternative reading is that gain-framed discourse ahead of good news reflects the absence of a fundamental catalyst, so that selling after the announcement is informed rather than reference-dependent. If sellers instead knew that the good news

¹¹The spreads are similar when pre-announcement GainTalk is replaced by its monthly cross-sectional residual with respect to capital gain overhang, prior returns, size, and turnover (Internet Appendix Table J.1).

overstated fundamental value, their information would imply weak or negative subsequent returns for good-news, high-GainTalk stocks: no further drift if prices incorporated that information immediately, and a partial reversal of the announcement return if they incorporated it gradually. Table 9 shows the opposite: at the three-month holding period, the good-news, high-GainTalk portfolio earns a five-factor alpha of +0.431% per month ($t = 3.26$), and the bad-news, low-GainTalk portfolio earns -0.600% ($t = 2.20$). Returns continue in the direction of the news rather than reverse, which is consistent with slow adjustment after news and difficult to reconcile with informed selling.

Overall, reference-dependent sentiment is not confined to trading quantities. The same gain framing whose association with retail selling is stronger where capital gain overhang is high also identifies the stocks whose prices adjust slowly to earnings news, and the drift appears where framing and the sign of the news agree. Reference-dependent sentiment therefore carries information about prices as well as about trading activity.

7 An Event Study of a Plausible Platform Shock to Framing Legitimacy

Whether a position sits at a paper gain or a paper loss does not by itself determine how strongly investors trade on it. What also matters is whether investors actively evaluate their positions in gain-loss terms. GainTalk captures the gain-versus-loss framework of this evaluation in public discourse. Sections 4 and 5 show that higher GainTalk is associated with trading patterns consistent with greater gain realization, particularly for stocks with larger embedded gains and on days with higher market returns.

But thinking about gains and losses does not necessarily mean relying on them when making trading decisions. An investor's purchase price is a sunk cost, and whether a position is currently at a gain or loss does not, by itself, determine whether selling is optimal. Two investors may evaluate their positions in exactly the same gain-framed terms, yet differ in whether they regard that evaluation as an appropriate basis for deciding to sell. I refer to this perceived appropriateness as *framing legitimacy*. When framing legitimacy is higher, investors may place greater weight on gain-loss evaluations in their trading decisions. The prediction is that the association between GainTalk

and retail selling becomes stronger.

To examine this idea, I use a change in how StockTwits presents stock performance to its users. On November 11, 2023, StockTwits introduced leaderboards prominently displaying the day's top 30 gainers and losers. Unlike changes in individual stock performance, the timing of this platform-wide introduction is less likely to be determined by any particular stock's price path or news.

The leaderboard does more than report price movements. By explicitly organizing stocks into gainers and losers, the platform presents gain-loss comparisons as a prominent way to evaluate stock performance. Although these comparisons are based on recent price changes rather than investors' personal purchase prices, they share the broader practice of evaluating an outcome relative to a past price. For investors who already think about their positions in these terms, the platform's use of this framework may reinforce the perception that gain-loss evaluation is a meaningful and appropriate basis for judging an investment. This provides a setting to examine whether greater public reinforcement of gain-loss framing is associated with a stronger relationship between GainTalk and retail selling.

Since historical leaderboard membership is not directly available, I reconstruct the daily lists using the ranking rule implied by the platform's display. I apply the same rule to the pre-introduction period to identify stocks that would have appeared on the leaderboard had the feature existed.

Because the leaderboard was introduced platform-wide, the rollout alone does not provide a cross-sectional control group. I therefore exploit differences in how strongly the platform reinforces gain-loss framing across stocks. I define a stock as *recurring* when it appears on the same side of the reconstructed leaderboard on consecutive trading days. Recurrence is defined both before and after the rollout, but what changes is whether the platform publicly displays the same classification on consecutive days.

The key distinction is between a single classification and repeated reinforcement of the same evaluation. When a stock appears on the leaderboard once, the platform evaluates its performance in terms of a gain or loss relative to an earlier price. When the same stock appears again in the same category, the platform repeats that evaluation. The stock is not simply experiencing another price movement; it is being publicly characterized as a gainer or loser for a second consecutive day. For investors who notice

these repeated appearances, gain-loss evaluation may therefore become a more established and seemingly appropriate way of judging that stock, rather than just one of many possible ways to describe its performance.

This repeated reinforcement may matter for trading decisions. Although the leaderboard does not reveal investors' personal purchase prices, it repeatedly presents backward-looking gain-loss comparisons as a meaningful way to evaluate stocks. Investors who already think about their positions in these terms may consequently place greater weight on that evaluation when deciding whether to sell. Under this framing-legitimacy mechanism, GainTalk should become more strongly associated with retail selling for recurring stocks after the leaderboard introduction. Non-recurring stocks provide the comparison group, since their performance receives less repeated reinforcement through the interface.

The framing-legitimacy mechanism may have different implications for the two sides of the leaderboard. On the gainer side, repeated classification reinforces gain-framed evaluation. Investors who place greater weight on this framework may become more willing to realize gains, so GainTalk should become more strongly associated with retail selling. On the loser side, however, two trading motives may operate in opposite directions. Greater attention to stocks experiencing large price declines may encourage selling, while greater reliance on loss-framed evaluation may discourage investors from realizing losses. As a result, the GainTalk–selling relationship need not strengthen in the same way on the loser side. The framing-legitimacy mechanism therefore predicts a clearer increase in the GainTalk–selling relationship for recurring gainers than for recurring losers.

Recurrence is not randomly assigned. Stocks may repeatedly appear among the largest movers because of persistent price trends, news, or investor attention. These factors may also affect trading, regardless of the leaderboard. Rather than relying on differences in selling levels between recurring and non-recurring stocks, I examine whether the relationship between GainTalk and selling changes differently for these two groups after the leaderboard introduction. Specifically, the $\text{GainTalk} \times \text{Recur} \times \text{Post}$ coefficient measures the post-introduction change in the recurring-minus-non-recurring difference in the GainTalk–selling slope.

This design yields three empirical predictions. First, before the introduction, when

repeated classification was not explicitly displayed through the leaderboard, the GainTalk–selling slopes should differ relatively little between recurring and non-recurring stocks. Second, after the introduction, repeated classification should reinforce gain-loss framing more strongly for recurring stocks. On the gainer side, greater reliance on gain-framed evaluation should strengthen the relationship between GainTalk and retail selling, leading to a positive $\text{GainTalk} \times \text{Recur} \times \text{Post}$ coefficient. On the loser side, the prediction is less clear. Greater attention to declining stocks may encourage selling, while greater reliance on loss-framed evaluation may encourage holding. Third, if a single leaderboard appearance were sufficient to strengthen the role of gain-loss evaluation in trading decisions, the GainTalk–selling relationship should also increase for non-recurring stocks, which is reflected in $\text{GainTalk} \times \text{Post}$. By contrast, the repeated-reinforcement mechanism predicts an additional increase for recurring stocks, captured by the triple interaction.

The estimation sample consists of reconstructed leaderboard stock-days from May 11, 2023, to May 11, 2024, covering six months on either side of the introduction. Post equals one beginning November 11, 2023. GainTalk is standardized separately for the gainer-side and loser-side estimation samples. Columns (1) and (3) examine contemporaneous abnormal log retail sell flow, following Table 3, while columns (2) and (4) use the following day’s outcome. All specifications include firm and date fixed effects, together with the controls described in Section 4. Reported t -statistics are based on standard errors double-clustered by firm and date.

Insert Table 10 near here.

Table 10 reports the results. The coefficient on GainTalk is positive and significant on the gainer side, consistent with the positive relationship between gain-framed discourse and retail selling documented earlier. Before the leaderboard introduction, $\text{GainTalk} \times \text{Recur}$ is small and statistically insignificant across all four columns, suggesting little evidence that the GainTalk–selling relationship differed between recurring and non-recurring stocks. After the introduction, however, a clear difference emerges between the two sides of the leaderboard. On the gainer side, the triple interaction is positive and significant for both contemporaneous selling ($+0.1210$, $t = 2.64$) and next-day selling ($+0.0910$, $t = 2.59$). On the loser side, the corresponding estimates

are close to zero and statistically insignificant ($+0.0118, t = 0.38$; $+0.0076, t = 0.26$). In contrast, $\text{GainTalk} \times \text{Post}$ is small and insignificant in all four specifications, suggesting no comparable change in the GainTalk–selling relationship for non-recurring stocks. Interestingly, $\text{Recur} \times \text{Post}$ is negative on the gainer side but positive and significant on the loser side. Thus, while repeated classification is associated with higher selling levels among losers, it strengthens the GainTalk–selling relationship only among gainers. This distinction suggests that the results are not simply driven by a uniform increase in trading activity following repeated leaderboard exposure. Instead, they are consistent with a framing-legitimacy mechanism in which repeated public reinforcement of gain framing strengthens its relationship with gain-realization decisions.

Insert Figure 3 near here.

Figure 3 examines how this relationship evolves around the leaderboard introduction. The event-time coefficients, θ_k , measure changes in the recurring-minus-non-recurring difference in the GainTalk–selling slope relative to the omitted bin -1 . Before the introduction, the estimates show no systematic pattern of differential changes between recurring and non-recurring stocks. After the introduction, however, the two sides of the leaderboard exhibit different patterns. On the gainer side, the GainTalk–selling slope increases sharply for recurring stocks, with the largest increase occurring in bin 0, the first 28 days after the introduction. The estimates remain positive in subsequent bins, although their magnitudes are smaller. On the loser side, the corresponding estimates remain close to zero throughout the post-introduction period, despite a positive change in the selling-level difference between recurring and non-recurring stocks. The event-time evidence therefore reinforces the distinction between changes in selling activity and changes in the relationship between gain-loss framing and selling. Because recurrence is defined using consecutive trading days, the repeated-reinforcement mechanism does not require the effect to build gradually over event time. The absence of systematic pre-introduction differences also makes an explanation based solely on persistent characteristics of recurring stocks less compelling, although time-varying shocks that differentially affect these stocks cannot be fully excluded.

Taken together, the results reveal a clear difference between the two sides of the leaderboard. On the gainer side, the GainTalk–selling relationship becomes significantly stronger for recurring stocks after the introduction, while no comparable difference is detectable before the introduction. On the loser side, repeated classification is associated with higher retail selling, but does not strengthen the relationship between GainTalk and selling. This contrast suggests that the results are not simply driven by greater attention or trading activity following repeated leaderboard exposure. Instead, they support the framing-legitimacy mechanism. By repeatedly classifying stocks as gainers or losers, the platform reinforces gain-loss framing as an appropriate framework for evaluating investments. When this reinforcement emphasizes gains, investors may place greater weight on gain-framed evaluations when deciding whether to realize gains. On the loser side, greater attention may encourage selling, while reliance on loss-framed evaluation may encourage holding. The findings therefore highlight a distinction between increased trading activity and greater reliance on gain-loss framing in trading decisions.

8 Conclusion

This paper studies when investors’ unrealized gains and losses become more consequential for trading. Using 185.1 million StockTwits posts and an LLM-supervised classification pipeline, I construct a measure of reference-dependent sentiment from investors’ backward-looking gain-loss evaluations expressed in public discourse.

I find that gain-framed discourse is associated with greater retail selling in stocks with embedded gains, while loss-framed discourse is associated with less selling in stocks with embedded losses. Similar patterns appear in aggregate market turnover, and reference-dependent sentiment predicts returns following earnings announcements.

Exploiting the introduction of StockTwits leaderboards as a platform-level shock to framing legitimacy, I find that GainTalk becomes significantly more strongly associated with retail selling for stocks repeatedly classified as top gainers, while repeated classification as a top loser raises selling without strengthening this association. This evidence suggests that the perceived legitimacy of gain-loss framing matters for how strongly investors rely on such evaluations when making trading decisions.

Overall, these findings suggest that reference-dependent sentiment identifies when the disposition effect shows up more strongly in trading and in prices around earnings news.

References

- Li An, Joseph Engelberg, Matthew Henriksson, Baolian Wang, and Jared Williams. The portfolio-driven disposition effect. *Journal of Finance*, 79(5):3459–3495, 2024. doi: 10.1111/jofi.13378.
- Marianne Andries, Maxime Bonelli, and David Sraer. Financial advisors and investors’ bias. *Review of Financial Studies*, 2025. doi: 10.1093/rfs/hhaf090.
- Brad M. Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21(2):785–818, 2008.
- Brad M. Barber, Xing Huang, Philippe Jorion, Terrance Odean, and Christopher Schwarz. A (sub)penny for your thoughts: Tracking retail investor activity in TAQ. *Journal of Finance*, 79(4):2403–2427, 2024. doi: 10.1111/jofi.13334.
- Nicholas Barberis and Wei Xiong. What drives the disposition effect? an analysis of a long-standing preference-based explanation. *Journal of Finance*, 64(2):751–784, 2009. doi: 10.1111/j.1540-6261.2009.01448.x.
- Nicholas Barberis and Wei Xiong. Realization utility. *Journal of Financial Economics*, 104(2):251–271, 2012. doi: 10.1016/j.jfineco.2011.10.005.
- Itzhak Ben-David and David Hirshleifer. Are investors really reluctant to realize their losses? Trading responses to past returns and the disposition effect. *Review of Financial Studies*, 25(8):2485–2532, 2012.
- James J. Choi and Adriana Z. Robertson. What matters to individual investors? Evidence from the horse’s mouth. *Journal of Finance*, 75(4):1965–2020, 2020.

- J. Anthony Cookson and Marina Niessner. Why don't we agree? Evidence from a social network of investors. *Journal of Finance*, 75(1):173–228, 2020. doi: 10.1111/jofi.12852. URL <https://doi.org/10.1111/jofi.12852>.
- J. Anthony Cookson and Marina Niessner. Does disagreement facilitate informed trading? *Journal of Financial and Quantitative Analysis*, 58(7):2879–2911, 2023. doi: 10.1017/S0022109022001053.
- J. Anthony Cookson, Runjing Lu, William Mullins, and Marina Niessner. The social signal. *Journal of Financial Economics*, 158:103870, 2024. doi: 10.1016/j.jfineco.2024.103870.
- J. Anthony Cookson, Marina Niessner, and Christoph Schiller. Can social media inform corporate decisions? Evidence from merger withdrawals. *Journal of Finance*, 81(1): 91–142, 2026. doi: 10.1111/jofi.13508. URL <https://onlinelibrary.wiley.com/doi/10.1111/jofi.13508>.
- Benjamin Enke and Thomas Graeber. Cognitive uncertainty. *Quarterly Journal of Economics*, 138(4):2021–2067, 2023.
- Benjamin Enke, Thomas Graeber, Ryan Oprea, and Jeffrey Yang. Behavioral attenuation. Technical report, National Bureau of Economic Research, 2024.
- Laura Escobar and Alvaro Pedraza. Active trading and (poor) performance: The social transmission channel. *Journal of Financial Economics*, 150(1):139–165, 2023. doi: 10.1016/j.jfineco.2023.103706.
- Andrea Frazzini. The disposition effect and underreaction to news. *Journal of Finance*, 61(4):2017–2046, 2006. doi: 10.1111/j.1540-6261.2006.00896.x.
- Cary Frydman and Antonio Rangel. Debiasing the disposition effect by reducing the saliency of information about a stock's purchase price. *Journal of Economic Behavior & Organization*, 107:541–552, 2014. doi: 10.1016/j.jebo.2014.01.017.
- Cary Frydman and Baolian Wang. The impact of salience on investor behavior: Evidence from a natural experiment. *Journal of Finance*, 75(1):229–276, 2020. doi: 10.1111/jofi.12851.

- Cary Frydman, Nicholas Barberis, Colin Camerer, Peter Bossaerts, and Antonio Rangel. Using neural data to test a theory of investor behavior: An application to realization utility. *Journal of Finance*, 69(2):907–946, 2014. doi: 10.1111/jofi.12126.
- Mark Grinblatt and Bing Han. Prospect theory, mental accounting, and momentum. *Journal of Financial Economics*, 78(2):311–339, 2005. doi: 10.1016/j.jfineco.2004.10.006.
- Mark Grinblatt and Matti Keloharju. What makes investors trade? *Journal of Finance*, 56(2):589–616, 2001. doi: 10.1111/0022-1082.00338.
- Ming Gu, David Hirshleifer, Siew Hong Teoh, and Shijia Wu. GIFfluence: A visual approach to investor sentiment and the stock market. Working paper, University of Southern California, 2025.
- Bing Han, David Hirshleifer, and Johan Walden. Social transmission bias and investor behavior. *Journal of Financial and Quantitative Analysis*, 57(1):390–412, 2022.
- Rawley Z. Heimer. Peer pressure: Social interaction and the disposition effect. *Review of Financial Studies*, 29(11):3177–3209, 2016. doi: 10.1093/rfs/hhw063.
- David Hirshleifer. Presidential address: Social transmission bias in economics and finance. *Journal of Finance*, 75(4):1779–1831, 2020. doi: 10.1111/jofi.12906.
- David Hirshleifer, Sonya Seongyeon Lim, and Siew Hong Teoh. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64(5):2289–2325, 2009.
- Harrison Hong, Jeffrey D. Kubik, and Jeremy C. Stein. Social interaction and stock-market participation. *Journal of Finance*, 59(1):137–163, 2004.
- Allen H. Huang, Hui Wang, and Yi Yang. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023. doi: 10.1111/1911-3846.12832.
- Jonathan E. Ingersoll and Lawrence J. Jin. Realization utility with reference-dependent preferences. *Review of Financial Studies*, 26(3):723–767, 2013. doi: 10.1093/rfs/hhs116.
- Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

- Hongqi Liu, Cameron Peng, Wei A. Xiong, and Wei Xiong. Taming the bias zoo. *Journal of Financial Economics*, 143(2):716–741, 2022.
- Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? return predictability and large language models. Working paper, 2023. URL <https://ssrn.com/abstract=4412788>.
- Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65, 2011.
- Terrance Odean. Are investors reluctant to realize their losses? *Journal of Finance*, 53(5):1775–1798, 1998.
- Thomas Renault. Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, 84:25–40, 2017. doi: 10.1016/j.jbankfin.2017.07.002.
- Hersh Shefrin and Meir Statman. The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance*, 40(3):777–790, 1985.
- Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168, 2007.

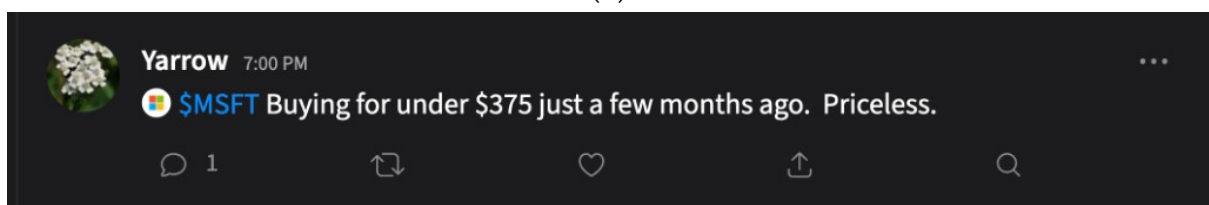
Tables and Figures

Figures

Figure 1: **Examples of Reference-Dependent Discourse.**

The figure applies the GPT-4o teacher-stage classification rubric (Internet Appendix K) to five investor posts. Each post is labeled on the three dimensions used to construct reference-dependent sentiment: object (whether the message is about a holder’s position), frame (gain versus loss), and anchor (whether the comparison is backward-looking). Panels (a), (b), (d), and (e) satisfy the joint condition that enters the measures: a holder’s position, a signed frame, and a past anchor. Panel (c) is labeled None on object, with a loss frame and a past anchor, and therefore does not enter the measures.

(a)



LLM classification Holder; Gain; Past.

Short reason Buying under \$375 a few months ago is a past entry; “priceless” frames the current price as a gain versus that cost basis.

(b)

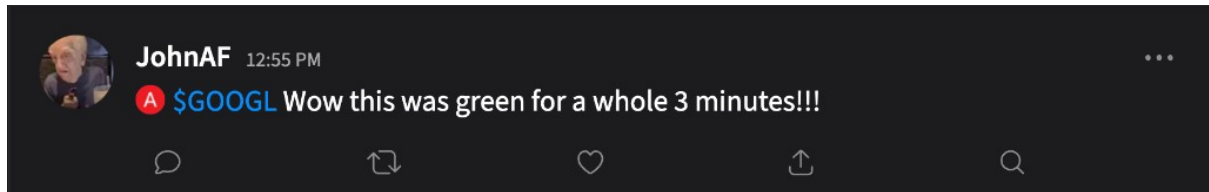


LLM classification Holder; Gain; Past.

Short reason The 1991 purchase is the cost basis; “do the math” encodes the current price as a large gain.

Figure 1: Examples of Reference-Dependent Discourse. (continued)

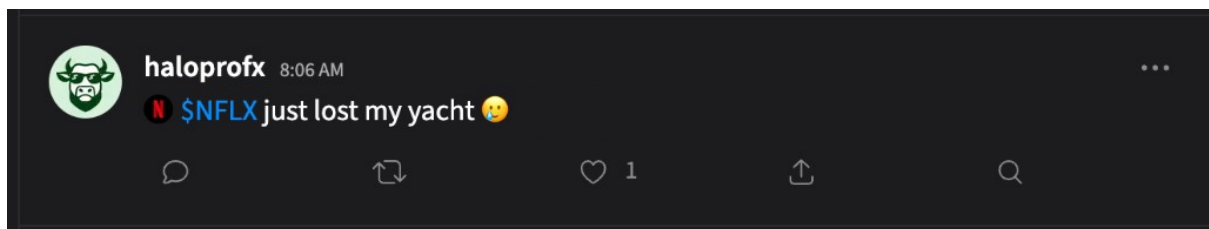
(c)



LLM classification None; Loss; Past.

Short reason “Green for a whole 3 minutes” sarcastically reports that the gain did not last, a loss relative to a just-preceding intra-day state.

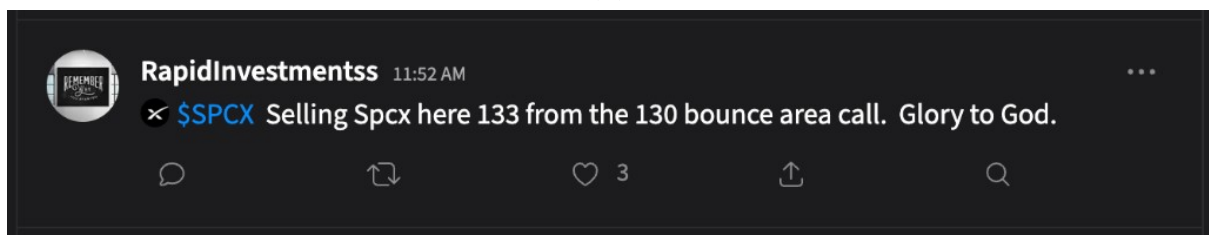
(d)



LLM classification Holder; Loss; Past.

Short reason “Lost my yacht” reports a personal loss as an already-incurred outcome.

(e)

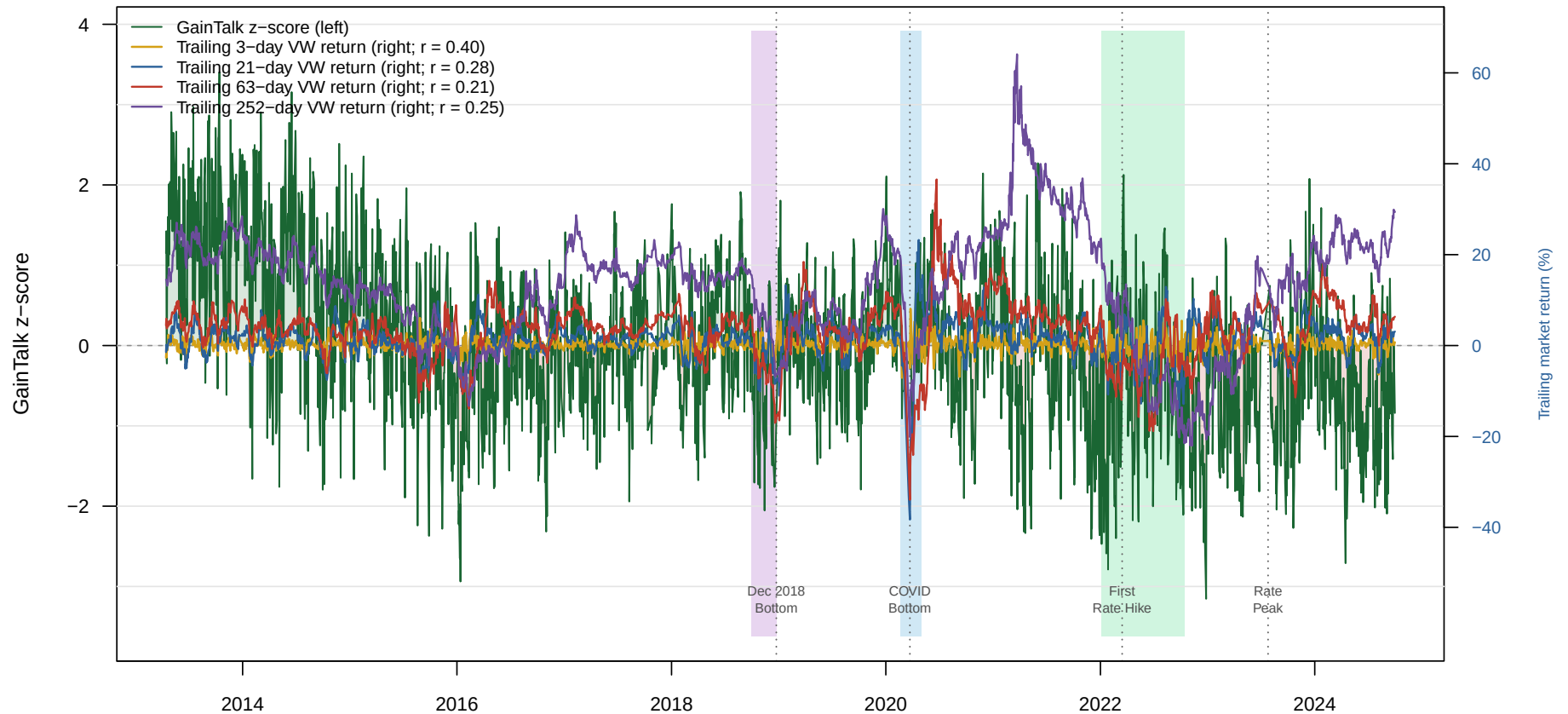


LLM classification Holder; Gain; Past.

Short reason Selling at 133 from the 130 bounce is a realized gain versus a past reference.

Figure 2: GainTalk and Trailing Market Returns.

This figure plots the daily GainTalk z-score (left axis) and trailing CRSP value-weighted market returns at 3, 21, 63, and 252 trading days (right axis). GainTalk is defined in equation (2) and is standardized to zero mean and unit variance over the full sample. Trailing k -day market return is the sum of daily CRSP value-weighted returns (percent) from day $t - k + 1$ through day t . Dashed vertical lines mark the December 2018 market bottom, the COVID bottom (March 2020), the first rate hike (March 2022), and the rate peak (July 2023). Shaded bands mark the 2018 Q4 sell-off, the COVID-19 crash, and the 2022 rate-shock episode. Pearson correlations between the GainTalk z-score and the 3-, 21-, 63-, and 252-day trailing returns are 0.40, 0.28, 0.21, and 0.25 ($N = 2,737$ StockTwits trading days, April 2013 to September 2024).



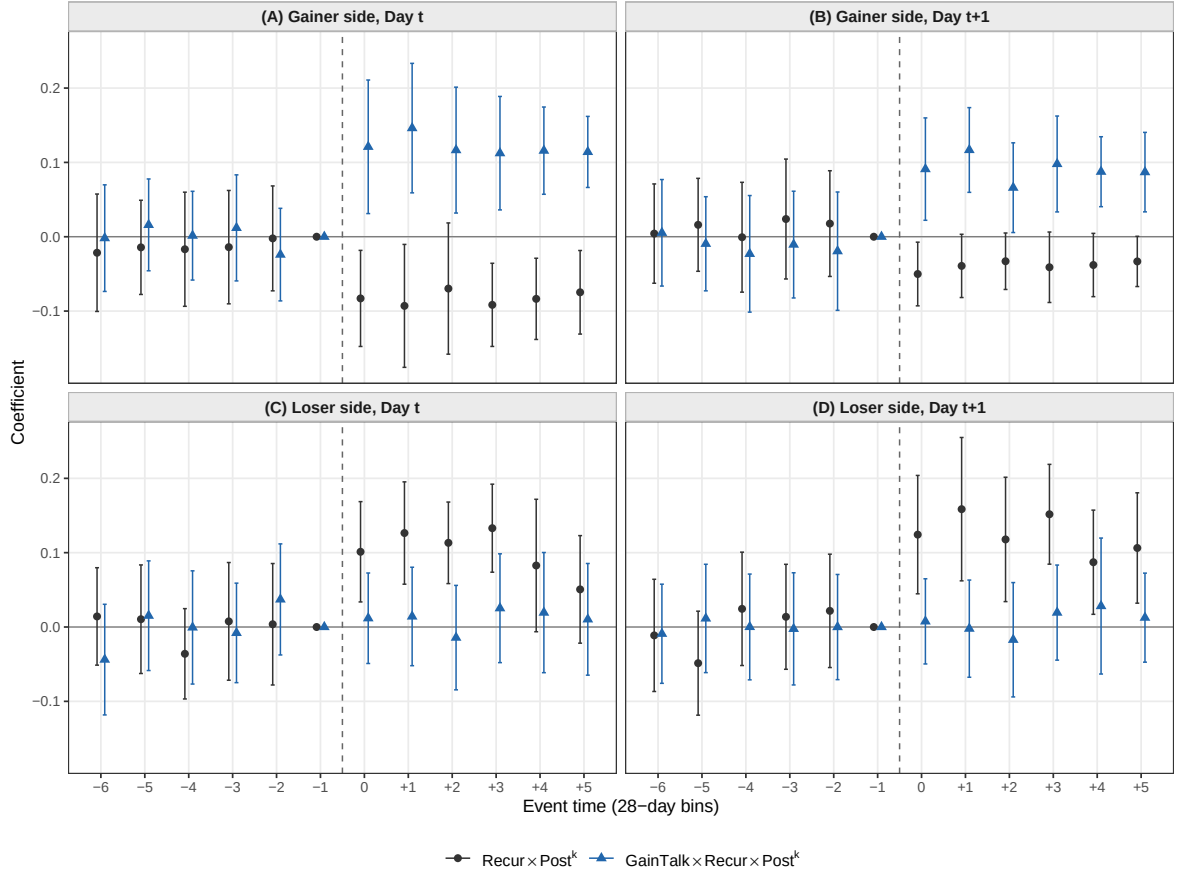


Figure 3: Event-Time Pattern of Leaderboard Recurrence, GainTalk, and Abnormal Retail Selling

The figure reports event-time estimates corresponding to Table 10. It plots, relative to omitted bin -1 ,

$$\begin{aligned}
 Y_{it} = & \beta_1 \text{GainTalk}_{it} + \beta_2 \text{Recur}_{it} + \beta_3 \text{GainTalk}_{it} \times \text{Recur}_{it} \\
 & + \sum_{k \neq -1} \gamma_k \text{Recur}_{it} \times \text{Post}_t^k + \sum_{k \neq -1} \delta_k \text{GainTalk}_{it} \times \text{Post}_t^k + \sum_{k \neq -1} \theta_k \text{GainTalk}_{it} \times \text{Recur}_{it} \times \text{Post}_t^k \\
 & + X'_{it} \lambda + \alpha_i + \tau_t + \varepsilon_{it}.
 \end{aligned}$$

Gray circles are γ_k ($\text{Recur} \times \text{Post}^k$); blue triangles are θ_k ($\text{GainTalk} \times \text{Recur} \times \text{Post}^k$). Panel (A) matches col. (1) of that table (gainer side, day t); panel (B) col. (2) (gainer side, day $t+1$); panel (C) col. (3) (loser side, day t); panel (D) col. (4) (loser side, day $t+1$). Post_t^k equals one when date t falls in a contiguous 28-calendar-day event-time bin k . Bin -1 (2023-10-14 to 2023-11-10) is omitted. Bin 0 is 2023-11-11 to 2023-12-08. The leftmost bin pools $k \leq -6$; the rightmost bin pools $k \geq +5$. Recur equals one when the stock was on the same side of the leaderboard on the previous trading day. GainTalk is estimation-sample standardized within side, as in that table. Sample, controls, and clustering match Table 10 (Post from 2023-11-11; six months on either side). Whiskers are 95% confidence intervals. Panels (A)–(B) use $N = 3,782$ gainer-side stock-days; panels (C)–(D) use $N = 3,164$ loser-side stock-days.

Table 1: Classifier Illustration: Reference-Dependent Sentiment versus Ordinary-Sentiment Posts

The table presents representative StockTwits messages from the held-out test set together with their object, frame, and anchor labels. Panel A shows messages that satisfy all three conditions for entering the reference-dependent sentiment measures: a holder's position, a gain or loss frame, and a past anchor. Panel B shows messages that fail at least one condition. Short reason is a one-line reading of why the message enters the measures or fails a required head. The full evaluation table appears in Internet Appendix C.

Panel A: Reference-Dependent Thinking Posts

	Object	Frame	Anchor	Short reason
<i>\$APPY Sold balance for +27%</i>	Holder	Gain	Past	Realized gain on a sale
<i>\$PLTR Finally green on my average.</i>	Holder	Gain	Past	Paper gain vs. average cost
<i>\$AAPL UP 18 bucks from where I bought it.</i>	Holder	Gain	Past	Paper gain vs. purchase price
<i>\$PCLN bought 100 shares at \$1336. Could be in some pain now.</i>	Holder	Loss	Past	Paper loss vs. cost basis
<i>\$GMTX %14 down. I wanted to make a profit lol</i>	Holder	Loss	Past	Paper loss on an open position
<i>\$MUX 24% below where I bought it 7 days ago.</i>	Holder	Loss	Past	Paper loss vs. a recent purchase

Panel B: Non-Reference-Dependent Thinking Posts

	Object	Frame	Anchor	Short reason
<i>\$TNXP nice little rise forecasted Monday</i>	None	Gain	Future	Forecast; not a holder's position
<i>\$TSLA little chilli Dipper before the next leg up</i>	None	Gain	Future	Directional view; not a holder's position
<i>\$AAPL will trade down to 510 today.. easy.</i>	None	Loss	Future	Forecast; not a holder's position
<i>\$GPRO Is gonna dump down to 40</i>	None	Loss	Future	Forecast; not a holder's position
<i>\$GME is this room dead or what</i>	None	None	Unspec.	Not a holder's position
<i>\$AMC anyone else have off tomorrow to watch the show?</i>	None	None	Unspec.	Not a holder's position

Table 2: Summary Statistics: Reference-Dependent Discourse and Sentiment Variables

The table reports summary statistics for the reference-dependent discourse and sentiment measures. Panel A uses the daily market-level sample (April 2013 to September 2024). Panel B uses the stock-day panel used in the cross-sectional tests of Section 4 (April 2015 to September 2024), formed by merging CRSP with retail sell flow identified from off-exchange trades using the quote-midpoint method (Barber et al. 2024), requiring at least five classified StockTwits posts per stock-day. Panel C reports pairwise Pearson correlations among GainTalk, Forward Sentiment, and RavenPack Sentiment in the same two samples; the lower triangle is reported. Variable definitions are in Table D.

Variable	<i>N</i>	Mean	SD	P10	P25	Median	P75	P90
<i>Panel A: Market-Day</i>								
GainTalk (%)	2,737	46.76	6.31	38.53	42.42	46.67	50.86	54.75
GainIntensity (%)	2,737	2.113	0.440	1.612	1.798	2.051	2.361	2.692
LossIntensity (%)	2,737	2.399	0.429	1.873	2.102	2.374	2.657	2.947
Forward Sentiment	2,737	0.404	0.126	0.243	0.327	0.405	0.477	0.573
RavenPack Sentiment	2,737	0.087	0.056	0.012	0.049	0.088	0.125	0.159
RavenPack Log News Count	2,737	7.632	0.418	7.157	7.356	7.575	7.913	8.227
Log social-media volume	2,737	10.681	0.837	9.476	10.075	10.792	11.229	11.755
<i>Panel B: Stock-Day</i>								
GainTalk (%)	1,891,685	50.55	34.74	3.37	17.25	47.76	86.82	97.94
GainIntensity (%)	1,891,685	1.400	3.437	0.000	0.000	0.004	1.101	4.447
LossIntensity (%)	1,891,685	1.455	3.397	0.000	0.000	0.002	1.302	4.845
Forward Sentiment	1,891,685	0.435	0.415	0.000	0.111	0.500	0.750	1.000
RavenPack Sentiment	1,891,685	0.012	0.247	−0.240	0.000	0.000	0.000	0.368
RavenPack Log News Count	1,891,685	0.444	0.696	0.000	0.000	0.000	0.693	1.386
Log social-media volume	1,891,685	2.831	1.075	1.792	2.079	2.485	3.219	4.290
<i>Panel C: Pearson Correlations</i>								
Variable	GainTalk	Forward Sentiment			RavenPack Sentiment			
Market-Day								
GainTalk	1.000							
Forward Sentiment	−0.018	1.000						
RavenPack Sentiment	0.014	0.052			1.000			
Stock-Day								
GainTalk	1.000							
Forward Sentiment	0.217	1.000						
RavenPack Sentiment	0.109	0.186			1.000			

Table 3: GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow

The table reports stock-day OLS estimates of same-day abnormal log retail sell flow on full-sample standardized GainTalk, full-sample standardized CGO, and their interaction. Abnormal log retail sell flow is $\log(1 + \text{SellFlow}_{it})$ relative to its mean over $[t-140, t-20]$, where $\text{SellFlow}_{it} = 10,000 \times \text{MRSVol}_{it} / \text{SHROUT}_{it}$ is quote-midpoint retail sell volume over shares outstanding, in basis points (Barber et al. 2024). $\text{CGO}_{i,t} = (\tilde{P}_{i,t} - \text{RP}_{i,t}) / \tilde{P}_{i,t}$ (Grinblatt and Han 2005). GainTalk and CGO are standardized using the full-sample mean and standard deviation. Cols. (5) and (6) replace GainTalk with BMO GainTalk (posts from 6:00am to 9:30am ET on t) and overnight GainTalk (posts from 4:00pm ET on $t-1$ to 9:30am ET on t), respectively, on stock-days with non-missing BMO and overnight GainTalk. Controls comprise same-day stock return, lagged abnormal log retail sell flow, cumulative returns ($t-5$ to $t-1$ and $t-21$ to $t-6$), Volatility ($t-5$ to $t-1$), log social-media volume, forward sentiment (Cookson et al. 2026), RavenPack sentiment, and RavenPack log news count. t -statistics in parentheses; standard errors are double-clustered by firm and date. Sample: April 2015 to September 2024; cols. (1)–(4): $N = 1,891,685$ stock-days. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0146*** (16.41)	+0.0151*** (21.98)	+0.0147*** (15.64)	+0.0152*** (20.69)		
BMO GainTalk					+0.0017** (2.16)	
Overnight GainTalk						+0.0006 (0.89)
GainTalk $\times$ CGO			+0.0135*** (15.02)	+0.0143*** (19.26)		
BMO GT $\times$ CGO					+0.0167*** (16.40)	
Overnight GT $\times$ CGO						+0.0121*** (15.30)
CGO	−0.0611*** (26.54)	−0.0616*** (23.30)	−0.0591*** (25.72)	−0.0592*** (22.50)	−0.0588*** (22.35)	−0.0585*** (22.20)
Forward sentiment	−0.0091*** (7.23)	−0.0053*** (8.20)	−0.0091*** (7.26)	−0.0052*** (8.15)	−0.0054*** (8.05)	−0.0051*** (7.95)
Log social-media volume	+0.0877*** (14.91)	+0.2559*** (43.08)	+0.0874*** (14.89)	+0.2554*** (43.07)	+0.2546*** (42.80)	+0.2551*** (42.95)
RavenPack sentiment	+0.0338*** (8.67)	+0.0221*** (8.49)	+0.0323*** (8.32)	+0.0206*** (7.95)	+0.0209*** (8.05)	+0.0204*** (7.88)
RavenPack log news count	+0.0921*** (21.67)	+0.0266*** (9.49)	+0.0922*** (21.71)	+0.0267*** (9.53)	+0.0271*** (9.62)	+0.0265*** (9.48)
Same-day return	+0.0115*** (27.78)	+0.0095*** (26.00)	+0.0117*** (27.89)	+0.0096*** (26.10)	+0.0095*** (25.85)	+0.0096*** (26.05)
Cum. return $t-5$ to $t-1$	+0.1178*** (8.12)	+0.1039*** (9.72)	+0.1189*** (8.20)	+0.1039*** (9.75)	+0.0998*** (9.35)	+0.1015*** (9.55)
Cum. return $t-21$ to $t-6$	+0.1306*** (10.52)	+0.1098*** (9.10)	+0.1301*** (10.52)	+0.1090*** (9.11)	+0.1055*** (8.85)	+0.1078*** (9.02)
Volatility ($t-5$ to $t-1$)	+0.3481*** (7.24)	+0.0872*** (2.84)	+0.3491*** (7.27)	+0.0898*** (2.92)	+0.0905*** (2.94)	+0.0889*** (2.88)
Lagged AbnLogSell	+0.6623*** (209.87)	+0.5799*** (165.37)	+0.6623*** (209.91)	+0.5798*** (165.39)	+0.5789*** (164.85)	+0.5794*** (165.10)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,891,685	1,238,017	1,697,519
Adj. R^2	0.505	0.568	0.505	0.569	0.565	0.568

Table 4: Stock-Level Sign-Split Test: Reference-Dependent Discourse and Abnormal Log Retail Sell Flow

The table reports stock-day OLS estimates of abnormal log retail sell flow on full-sample standardized GainIntensity_{it} and LossIntensity_{it}, each interacted with a gain-state or loss-state indicator, and a linear spline in CGO with a knot at 0. Cols. (1) and (2) use same-day abnormal log retail sell flow. Cols. (3) and (4) use next-day abnormal log retail sell flow, measured on $t+1$. GainIntensity_{it} and LossIntensity_{it} are equation (1) computed over the messages that mention stock i on date t . The gain state is $1[\text{CGO}_{it} > 0]$ and the loss state is $1[\text{CGO}_{it} \leq 0]$. The spline terms are $\max(\text{CGO}_{it}, 0)$ and $\min(\text{CGO}_{it}, 0)$. CGO in the state indicators and the spline is raw CGO. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Same-day AbnLogSell		Next-day AbnLogSell	
	(1)	(2)	(3)	(4)
GainIntensity _{it} × 1[CGO _{it} > 0]	+0.0151*** (11.31)	+0.0126*** (10.30)	+0.0086*** (8.40)	+0.0064*** (7.25)
GainIntensity _{it} × 1[CGO _{it} ≤ 0]	+0.0058* (1.75)	+0.0014 (0.90)	−0.0018* (1.72)	−0.0001 (0.09)
LossIntensity _{it} × 1[CGO _{it} > 0]	−0.0020 (1.61)	−0.0014 (1.55)	−0.0013 (1.18)	−0.0011 (1.12)
LossIntensity _{it} × 1[CGO _{it} ≤ 0]	−0.0111*** (9.35)	−0.0165*** (8.88)	−0.0070*** (7.32)	−0.0025*** (2.99)
1[CGO _{it} > 0]	−0.0074** (2.47)	−0.0085*** (3.36)	−0.0072** (2.17)	−0.0040 (1.38)
max(CGO _{it} , 0)	+0.2522*** (12.97)	+0.3665*** (18.38)	+0.2636*** (12.99)	+0.3539*** (17.58)
min(CGO _{it} , 0)	−0.1732*** (24.53)	−0.1338*** (20.92)	−0.1271*** (19.92)	−0.1318*** (20.53)
Controls	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Date FE	No	Yes	No	Yes
Observations	1,891,685	1,891,685	1,886,172	1,886,172
Adj. R ²	0.509	0.573	0.351	0.409

Table 5: GainTalk, Market Turnover, and Aggregate Retail Sell Flow

The table reports daily OLS estimates of abnormal log trading activity on standardized GainTalk and the same-day market return. Cols. (1)–(4) use abnormal log market turnover; cols. (5)–(6) use abnormal log aggregate retail sell flow. Cols. (1), (3), and (5) use the CRSP universe; cols. (2), (4), and (6) use S&P 500 constituents. Cols. (3)–(6) add GainTalk $\times$ market return. All columns include year-quarter and day-of-week fixed effects. $\text{AbnLogTO} = \log(\text{TO}_t) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(\text{TO}_s)$ (window specification following [Cookson and Niessner 2020](#)), where TO is aggregate dollar volume divided by market capitalization. AbnLogSell applies the same window to RetailSell, where $\text{RetailSell}_t = 100 \times \sum_i (P_{it} \times \text{MRSVol}_{it}) / \sum_i \text{MktCap}_{it}$ and MRSVol_{it} is quote-midpoint retail sell volume ([Barber et al. 2024](#)). GainTalk is the standardized gain-frame share of reference-dependent posts. The lagged dependent variable is lagged AbnLogTO in cols. (1)–(4) and lagged AbnLogSell in cols. (5)–(6). Other controls comprise forward market sentiment ([Cookson et al. 2026](#)), log social-media volume, RavenPack sentiment, RavenPack log news count, cumulative market returns ($t-5$ to $t-1$ and $t-21$ to $t-6$), and Volatility ($t-5$ to $t-1$). t -statistics in parentheses; Newey-West standard errors with 5 lags. Cols. (1)–(4) cover April 2013 to September 2024 ($N=2,737$). Cols. (5)–(6) cover August 2015 to September 2024 ($N=2,151$). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	AbnLogTO				AbnLogSell	
	CRSP	S&P 500	CRSP	S&P 500	CRSP	S&P 500
	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0312*** (6.71)	+0.0273*** (5.59)	+0.0294*** (6.27)	+0.0256*** (5.19)	+0.0344*** (8.34)	+0.0237*** (5.31)
GainTalk $\times$ Market return			+0.0120*** (4.06)	+0.0118*** (3.72)	+0.0191*** (6.02)	+0.0197*** (5.64)
Forward market sentiment	−0.0928*** (8.66)	−0.0996*** (8.87)	−0.0904*** (8.47)	−0.0972*** (8.68)	−0.0844*** (9.43)	−0.0934*** (9.12)
Log social-media volume	+0.2289*** (10.84)	+0.1978*** (8.70)	+0.2274*** (10.83)	+0.1963*** (8.68)	+0.2164*** (10.52)	+0.1563*** (6.54)
RavenPack sentiment	+0.0211*** (6.11)	+0.0234*** (6.24)	+0.0208*** (6.01)	+0.0231*** (6.14)	+0.0112*** (3.41)	+0.0141*** (3.57)
RavenPack log news count	+0.0465*** (5.15)	+0.0454*** (4.81)	+0.0466*** (5.18)	+0.0455*** (4.84)	+0.0384*** (5.99)	+0.0423*** (5.44)
Market return	−0.0159*** (4.23)	−0.0173*** (4.27)	−0.0116*** (2.91)	−0.0131*** (3.05)	+0.0097** (2.52)	+0.0104** (2.50)
Cum. return $t-5$ to $t-1$	−0.0066*** (3.65)	−0.0069*** (3.53)	−0.0063*** (3.48)	−0.0066*** (3.37)	−0.0022 (1.40)	−0.0024 (1.29)
Cum. return $t-21$ to $t-6$	−0.0015 (1.34)	−0.0023** (2.08)	−0.0014 (1.30)	−0.0023** (2.03)	+0.0001 (0.16)	−0.0005 (0.55)
Volatility ($t-5$ to $t-1$)	+0.0367** (2.57)	+0.0344** (2.19)	+0.0363** (2.52)	+0.0341** (2.15)	+0.0227 (1.64)	+0.0189 (1.07)
Lagged dependent variable	+0.3602*** (16.41)	+0.3886*** (17.43)	+0.3585*** (16.32)	+0.3867*** (17.35)	+0.5322*** (25.82)	+0.5325*** (24.08)
Year-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,737	2,737	2,737	2,737	2,151	2,151
Adj. R^2	0.563	0.579	0.566	0.581	0.777	0.755

Table 6: Sign-Split Test: Reference-Dependent Discourse and Abnormal Log Market Turnover

The table reports daily OLS estimates of same-day abnormal log market turnover on standardized GainIntensity_t and LossIntensity_t , each interacted with a gain-state or loss-state indicator, a linear spline in the market return with a knot at zero, and controls. Cols. (1)–(3) use the CRSP universe; cols. (4)–(6) use S&P 500 constituents. Within each universe, fixed effects vary from none to year-quarter to year-quarter and day-of-week. $\text{AbnLogTO}_t = \log(\text{TO}_t) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(\text{TO}_s)$ (window specification following [Cookson and Niessner 2020](#)), where TO is aggregate dollar volume divided by market capitalization. $R_t^+ = \max(R_{m,t}, 0)$ and $R_t^- = \min(R_{m,t}, 0)$. Everything else follows Table 5. t -statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	CRSP universe			S&P 500		
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{GainIntensity}_t \times \mathbf{1}[R_{m,t} > 0]$	+0.0095** (2.00)	+0.0107*** (2.98)	+0.0091** (2.55)	+0.0082* (1.65)	+0.0089** (2.30)	+0.0072* (1.86)
$\text{GainIntensity}_t \times \mathbf{1}[R_{m,t} \leq 0]$	+0.0058 (1.10)	+0.0080* (1.92)	+0.0069 (1.64)	+0.0044 (0.78)	+0.0071 (1.60)	+0.0058 (1.31)
$\text{LossIntensity}_t \times \mathbf{1}[R_{m,t} > 0]$	−0.0021 (0.43)	−0.0054 (1.36)	−0.0065 (1.63)	−0.0033 (0.62)	−0.0041 (0.96)	−0.0055 (1.28)
$\text{LossIntensity}_t \times \mathbf{1}[R_{m,t} \leq 0]$	−0.0124** (2.24)	−0.0153*** (3.46)	−0.0158*** (3.56)	−0.0134** (2.31)	−0.0160*** (3.36)	−0.0168*** (3.52)
$\mathbf{1}[R_{m,t} > 0]$	−0.0351 (1.61)	−0.0375** (2.01)	−0.0350* (1.89)	−0.0349 (1.51)	−0.0372* (1.89)	−0.0348* (1.80)
$R_t^+ = \max(R_{m,t}, 0)$	+0.0396*** (3.94)	+0.0212*** (2.65)	+0.0187** (2.36)	+0.0449*** (4.24)	+0.0241*** (2.80)	+0.0215** (2.53)
$R_t^- = \min(R_{m,t}, 0)$	−0.0683*** (6.21)	−0.0455*** (5.63)	−0.0463*** (5.74)	−0.0771*** (6.59)	−0.0508*** (5.79)	−0.0519*** (5.92)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year-quarter FE	No	Yes	Yes	No	Yes	Yes
Day-of-week FE	No	No	Yes	No	No	Yes
Observations	2,737	2,737	2,737	2,737	2,737	2,737
Adj. R^2	0.264	0.429	0.634	0.283	0.441	0.646

Table 7: Overnight GainTalk and Abnormal Log Market Turnover

This table reports daily OLS estimates of abnormal log market turnover on overnight and before-market-open GainTalk, interacted with the contemporaneous full-day or overnight market return. The overnight window runs from 16:00 ET on day $t-1$ to 09:30 ET on day t ; the before-market-open window is 06:00 to 09:30 ET. Cols. (1)–(3) use BMO GainTalk; col. (2) uses next-day turnover. Cols. (4)–(5) use overnight GainTalk; cols. (6)–(7) add overnight social-media and news volume as attention placebos. All columns use the CRSP universe and year-quarter and day-of-week fixed effects. Everything else follows Table 5. t -statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	GainTalk window and return state					Attention placebos	
	(1) Same-day	(2) Next-day	(3) Same-day	(4) Same-day	(5) Same-day	(6) Same-day	(7) Same-day
Overnight GainTalk				+0.0095*** (3.01)	+0.0078** (2.48)	+0.0095*** (2.96)	+0.0076** (2.31)
BMO GainTalk	+0.0025 (0.83)	+0.0021 (0.66)	+0.0038 (1.15)				
Overnight vol.						+0.0895*** (6.46)	
Overnight news vol.							+0.0008 (0.13)
Market return	−0.0084** (2.20)	−0.0135*** (3.22)			−0.0076* (1.95)	−0.0104*** (2.58)	−0.0080** (2.02)
Overnight return			−0.0088 (1.39)	−0.0079 (1.26)			
BMO GT × Return	+0.0087*** (2.79)	+0.0065* (1.78)					
BMO GT × Overnight return			+0.0287*** (5.78)				
Overnight GT × Overnight return				+0.0259*** (5.17)			
Overnight GT × Return					+0.0068** (2.11)	+0.0081** (2.49)	+0.0074** (2.27)
Overnight vol. × Return						+0.0032 (0.94)	
Overnight news vol. × Return							+0.0056* (1.87)
Dep. var. timing	day t	day $t+1$	day t	day t	day t	day t	day t
Return-state window	full day	full day	overnight	overnight	full day	full day	full day
Controls				Yes			
Year-quarter FE				Yes			
Day-of-week FE				Yes			
Observations	2,737	2,736	2,737	2,737	2,737	2,737	2,737
Adj. R^2	0.563	0.451	0.569	0.568	0.563	0.549	0.552

Table 8: Reference-Dependent Sentiment versus Competing Channels

This table reports daily OLS estimates of same-day abnormal log market turnover on standardized GainTalk and competing sentiment, news, and attention measures, each interacted with the market return. Col. (1) enters forward market sentiment (Cookson et al. 2026) alone; cols. (2)–(5) add GainTalk and cumulatively horse-race the alternatives listed in the column headers. All columns use the CRSP universe and year-quarter and day-of-week fixed effects. Everything else follows Table 5, excluding any measure that enters as a main effect. *t*-statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	(1) Forward Sent. (solo)	(2) Horse Race + Forward Sent.	(3) Horse Race + RavenPack sent.	(4) Horse Race + RavenPack count.	(5) Horse Race + Social vol.
GainTalk		+0.0124** (2.55)	+0.0138*** (2.88)	+0.0220*** (4.77)	+0.0280*** (6.02)
forward sent.	−0.0402*** (3.92)	−0.0531*** (4.54)	−0.0537*** (4.56)	−0.0634*** (5.81)	−0.0970*** (8.38)
RavenPack sent.			+0.0205*** (5.41)	+0.0195*** (5.25)	+0.0208*** (6.16)
RavenPack count.				+0.0340*** (9.71)	+0.0439*** (6.24)
Social vol.					+0.2153*** (11.20)
Market return ($\hat{\beta}_2$)	−0.0162*** (4.08)	−0.0139*** (3.22)	−0.0147*** (3.42)	−0.0167*** (3.98)	−0.0126*** (2.95)
GainTalk × Return		+0.0127*** (4.17)	+0.0121*** (4.00)	+0.0123*** (4.19)	+0.0134*** (3.40)
forward sent. × Return	+0.0085** (2.33)	+0.0073** (2.01)	+0.0074** (2.10)	+0.0067* (1.92)	+0.0010 (0.13)
RavenPack sent. × Return			−0.0028 (0.96)	−0.0030 (1.04)	−0.0030 (1.03)
RavenPack count. × Return				+0.0053* (1.74)	+0.0047 (1.58)
Social vol. × Return					+0.0057 (0.64)
Controls			Yes		
Year-quarter FE			Yes		
Day-of-week FE			Yes		
Observations	2,737	2,737	2,737	2,737	2,737
Adj. R^2	0.508	0.512	0.520	0.544	0.573

Table 9: High- and Low-GainTalk Spread Alphas

The table reports equal-weighted calendar-time Fama-French five-factor alphas for spread portfolios formed on earnings news and pre-announcement GainTalk, following [Frazzini \(2006\)](#). At each month-start, stocks are independently sorted into quintiles by the market-model cumulative abnormal return around their most recent earnings announcement, $CAR[-2, +1]$, and by GainTalk averaged over the 60 trading days ending the day before that announcement. The high-GainTalk spread is long good-news/high-GainTalk stocks and short bad-news/low-GainTalk stocks; the low-GainTalk spread is long good-news/low-GainTalk stocks and short bad-news/high-GainTalk stocks. Rows report overlapping rolling periods of $K = 1, 2, 3, 6$, and 12 months. Columns report alphas for each leg, for each long-short spread, and for the difference between the two spreads. Alphas are in monthly percent. t -statistics in parentheses; Newey-West standard errors with $K-1$ lags. Sample: 2013-04 to 2024-09 ($N = 138$ months). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Rolling period	High-GainTalk spread			Low-GainTalk spread			High – low	N months
	Bad news	Good news	L/S	Bad news	Good news	L/S		
+1	−0.725** (2.41)	+0.476*** (3.66)	+1.201*** (3.64)	−0.065 (0.38)	−0.055 (0.26)	+0.010 (0.04)	+1.191*** (2.74)	138
+2	−0.650** (2.47)	+0.458*** (3.55)	+1.107*** (3.95)	+0.003 (0.02)	−0.135 (0.62)	−0.139 (0.53)	+1.246*** (2.85)	138
+3	−0.600** (2.20)	+0.431*** (3.26)	+1.031*** (3.64)	+0.017 (0.11)	−0.061 (0.27)	−0.078 (0.29)	+1.109** (2.40)	138
+6	−0.485* (1.92)	+0.411*** (2.83)	+0.896*** (3.42)	+0.175 (1.13)	+0.046 (0.24)	−0.129 (0.54)	+1.025** (2.45)	138
+12	−0.324 (1.31)	+0.389*** (4.00)	+0.713*** (3.02)	+0.161 (1.25)	+0.082 (0.45)	−0.079 (0.45)	+0.792** (2.28)	138

Table 10: Leaderboard Recurrence, GainTalk, and Abnormal Retail Selling

The table reports stock-day OLS estimates of abnormal log retail sell flow on leaderboard stock-days. The leaderboard is the inferred StockTwits daily top 30 gainers and losers, reconstructed from the platform's display conventions. Cols. (1)–(2) use the gainer side and cols. (3)–(4) the loser side; cols. (1) and (3) use day t and cols. (2) and (4) use day $t+1$. *Recur* equals one when the stock was on the same side of the leaderboard on the previous trading day. *Post* equals one from 2023-11-11. The sample is the six months on either side of that date (2023-05-11 to 2024-05-11). *GainTalk* is the gain-framed share of reference-dependent messages, standardized within each side's estimation sample. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Gainer side		Loser side	
	(1) Day t	(2) Day $t+1$	(3) Day t	(4) Day $t+1$
GainTalk	+0.0348* (1.89)	+0.0213** (2.14)	+0.0042 (0.52)	−0.0038 (0.44)
Recur	+0.0186 (0.61)	−0.0287 (0.84)	−0.0112 (0.29)	+0.0268 (0.64)
GainTalk × Recur	+0.0092 (0.28)	−0.0068 (0.17)	−0.0111 (0.32)	+0.0117 (0.31)
GainTalk × Post	+0.0058 (0.37)	+0.0111 (0.68)	−0.0103 (0.61)	+0.0049 (0.27)
Recur × Post	−0.0826** (2.12)	−0.0391 (1.58)	+0.1012*** (2.94)	+0.1243*** (3.06)
GainTalk × Recur × Post	+0.1210*** (2.64)	+0.0910*** (2.59)	+0.0118 (0.38)	+0.0076 (0.26)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	3,782	3,782	3,164	3,164
Adj. R^2	0.578	0.546	0.636	0.612

Internet Appendix

Contents

A	A Stylized Model of Reference-Dependent Sentiment, Return and Turnover	57
A.1	Environment	57
A.2	Selling decisions	58
A.3	Aggregation and the central proposition	59
A.4	Empirical predictions and mapping to the data	61
B	Corpus Coverage	64
C	Classifier Validation	65
D	Variable Definitions	67
D.1	Per-Post Soft Scores	67
D.2	Daily Market-Level Discourse Measures	68
D.3	Stock-Level Discourse Measures	68
E	Classifier Label Distribution	70
F	Textual Validation	71
G	Trading Activity and Controls	73
H	Robustness and Extensions for Stock-Level GainTalk and Abnormal Retail Sell Flow	74
I	Robustness and Extensions for Market-Level GainTalk and Abnormal Market Turnover	85
J	Earnings News, GainTalk, and Disposition Spreads: Additional Results	93
K	LLM Teacher-Stage Annotation Protocol	97

A A Stylized Model of Reference-Dependent Sentiment, Return and Turnover

The empirical sections document four regularities. GainTalk is associated with more selling. This association is stronger in gain states, and in the cross-section of stocks it rises with embedded gains and, more weakly, with deeper embedded losses. Gain-framing intensity matters where gains are prevalent and loss-framing intensity where losses are prevalent. After the leaderboard introduction, the GainTalk–selling slope steepens significantly for stocks repeatedly classified as gainers; for stocks repeatedly classified as losers, the slope estimate is small and insignificant, while the selling level rises. This section motivates these regularities with a deliberately spare model of reference-dependent realization ([Kahneman and Tversky 1979](#); [Shefrin and Statman 1985](#); [Barberis and Xiong 2012](#)). The model separates the forces the empirical argument relies on: the disposition asymmetry in baseline selling, frame congruence, an asymmetry in how strongly holders rely on gain versus loss evaluations, and the weight holders place on gain-loss evaluation, which Section 7 calls framing legitimacy. Attention shifts selling levels and, for stocks displayed as losers, prompts holders who frame their positions as losses to sell. The model is written as if framing moved selling. It can also be read as a reduced-form description of a common reference-dependent mindset that drives both the prevailing framing and selling propensities, which is the reading maintained in the empirical sections.

A.1 Environment

Consider stock i on date t ; I drop the subscripts when no confusion arises. A continuum of holders $j \in [0, 1]$ each holds one share bought at cost basis c_j , and the current price is p . The cross-sectional distribution of cost bases has CDF F , so the mass of holders at a gain (*winners*) and at a loss (*losers*) is

$$W \equiv F(p), \quad L \equiv 1 - W.$$

The stock-level tests use capital gain overhang to proxy for the prevailing gain/loss state of a stock’s holders, which W summarizes in the model. In a representative-market

analogue, the market return locally moves the winner mass in the same direction,

$$W_t = W_0 + \tau R_t + o(R_t), \quad \tau > 0, \quad (12)$$

so the market return proxies for the prevailing gain/loss state.

Reference-dependent sentiment. Discourse about the stock contains gain-framed and loss-framed reference-dependent messages with intensities $\gamma \geq 0$ and $\lambda \geq 0$, the model counterparts of GainIntensity and LossIntensity in equation (1). GainTalk is the gain share

$$\theta \equiv \frac{\gamma}{\gamma + \lambda} \in [0, 1],$$

as in equations (2) and (3). Because GainTalk measures the direction of reference-dependent discourse rather than its volume, the share-based comparative statics are conditional on the total amount of such discourse: they hold $\gamma + \lambda > 0$ fixed and normalize it to one, so that $\gamma = \theta$ and $\lambda = 1 - \theta$.

A.2 Selling decisions

Each holder decides on date t whether to sell her share. The *baseline* selling probability follows the disposition pattern,

$$\pi_0^W > \pi_0^L > 0, \quad (13)$$

a reduced-form summary of the [Kahneman and Tversky \(1979\)](#) value function applied to the realization choice in [Shefrin and Statman \(1985\)](#) and [Odean \(1998\)](#). Three assumptions govern how reference-dependent sentiment relates to selling.

A1 (Frame congruence). Gain framing is associated with the selling propensity of holders at a gain, and loss framing with that of holders at a loss. A gain frame is essentially uninformative about a loser's decision to hold, and a loss frame about a winner's decision to sell.

A2 (State-pool relevance). The association scales with the share of holders in the matching state: the prevailing frame is more relevant to a holder when it describes the position of more of the stock's holders. This assumption admits the two readings in Section 4.2. Under the perception reading, a frame that describes most

holders' positions is more likely to reflect an accurate perception of the gain/loss state. Under the social reading, holders who share a state reinforce one another's framing ([Han et al. 2022](#)).

A3 (Gain-side reliance). Holders rely more on a gain evaluation when deciding whether to sell a winner than on a loss evaluation when deciding whether to hold a loser: $\eta_W > \eta_L > 0$. This is in the spirit of realization utility ([Barberis and Xiong 2012](#)), in which realizing a gain delivers a burst of utility that makes the gain evaluation directly relevant to the sale.

Two further objects enter the selling decision. *Framing legitimacy* $\omega_{it} \geq 0$ is the weight holders place on gain-loss evaluation when they trade. *Attention* shifts the selling probabilities of winners and losers by a_{it}^W and a_{it}^L , which may take either sign. For a stock that the platform displays as a loser, attention can also prompt loss-framed holders to sell, at rate $\alpha_{it} \geq 0$; otherwise $\alpha_{it} = 0$. A1–A3 together yield

$$\pi^W = \pi_0^W + a^W + \omega \eta_W \gamma W, \quad (14)$$

$$\pi^L = \pi_0^L + a^L - (\omega \eta_L - \alpha) \lambda L. \quad (15)$$

Parameters are restricted to the region in which π^W and π^L lie in $(0, 1)$, and the comparative statics below are stated within that region. Gain framing raises winners' selling, and loss framing lowers losers' selling, each in proportion to the size of the matching pool and to framing legitimacy. The distinction among the new objects is that a^W and a^L shift selling regardless of how holders frame their positions, α raises selling by loss-framed holders of displayed losers, and framing legitimacy changes how much the frame matters.

A.3 Aggregation and the central proposition

Aggregating (14)–(15) across holders and substituting $\gamma = \theta$ and $\lambda = 1 - \theta$ yields date- t selling

$$S = W \pi^W + L \pi^L = \underbrace{W \pi_0^W + L \pi_0^L}_{S^0} + \underbrace{W a^W + L a^L}_A + \omega [\eta_W \theta W^2 - \eta_L (1 - \theta) L^2] + \alpha (1 - \theta) L^2. \quad (16)$$

The first component is the standard disposition prediction: selling rises with the winner pool because winners are more eager to sell than losers. The second is the attention

component, which does not depend on framing. The third is the sentiment component. The last is attention-induced selling by loss-framed losers. At the stock level, S is the model counterpart of retail sell flow. In the representative-market analogue, W_t replaces W , and aggregate turnover inherits the same form because its selling side rises when winners realize and falls when losers hold.

Proposition 1 (Comparative statics). *Maintain (12) and (16), let $\omega > 0$, and, in (P1)–(P3), let $\alpha = 0$. Then*

(P1) **Level association.**

$$\frac{\partial S}{\partial \theta} = \omega (\eta_W W^2 + \eta_L L^2) > 0. \quad (17)$$

(P2) **State dependence.**

$$\frac{\partial^2 S}{\partial \theta \partial W} = 2\omega (\eta_W W - \eta_L L), \quad (18)$$

which is increasing in W , negative for $W < W^*$, and positive for $W > W^*$, where $W^* \equiv \eta_L/(\eta_W + \eta_L) < 1/2$. The GainTalk–selling slope is therefore U-shaped in the gain/loss state, with its minimum at W^* . At mirror-image states $w > 1/2$ and $1 - w$, the slope changes faster at w , so the gain side is steeper. Holding ω , η_W , and η_L fixed, the cross-partial averaged over states with $E[W] = 1/2$ equals $\omega(\eta_W - \eta_L) > 0$. In the representative-market analogue, locally, $\partial^2 S / \partial \theta \partial R_t \approx 2\omega \tau (\eta_W W_t - \eta_L L_t)$, with the same average sign.

(P3) **Matched states.** *Before the normalization $\gamma + \lambda = 1$,*

$$\frac{\partial S}{\partial \gamma} = \omega \eta_W W^2, \quad \frac{\partial S}{\partial \lambda} = -\omega \eta_L L^2. \quad (19)$$

Gain-framing intensity raises selling by more when a larger share of holders is at a gain, and loss-framing intensity lowers selling by more when a larger share is at a loss.

(P4) **Framing legitimacy and attention.** *Let D_{it} indicate that the platform displays a repeated classification of the stock, and suppose $\omega_{it} = \omega_0 + \psi D_{it}$ with $\psi > 0$, $A_{it} = A_{it}^0 + \kappa D_{it}$ with κ of either sign, where A_{it}^0 collects attention differences that do not depend on D_{it} , and $\alpha_{it} = \chi D_{it}$ with $\chi \geq 0$ on the loser list. Then the change in the GainTalk–selling slope associated with the display is*

$$\Delta \frac{\partial S}{\partial \theta} = \begin{cases} \psi (\eta_W W^2 + \eta_L L^2) > 0 & \text{on the gainer list,} \\ \psi (\eta_W W^2 + \eta_L L^2) - \chi L^2 & \text{on the loser list,} \end{cases} \quad (20)$$

while κ shifts the level of selling and leaves the slope unchanged. On the loser list, attention-induced selling by loss-framed holders offsets greater reliance on loss-framed evaluation, so the slope response is smaller than on the gainer list and is zero or negative when χ is large enough. The display also shifts the level of selling at a given θ through ω , by $\psi[\eta_W\theta W^2 - \eta_L(1 - \theta)L^2]$. If the gainer list places a share $\bar{W} > 1/2$ of holders at a gain and the loser list places the same share at a loss, the slope response on the gainer side exceeds the response on the loser side by $\psi(\eta_W - \eta_L)(2\bar{W} - 1) + \chi\bar{W}^2 > 0$.

Sketch of proof. (P1) and (P2) follow from differentiating (16) with respect to θ and then W , using $L = 1 - W$. The cross-partial in (18) is linear in W with slope $2\omega(\eta_W + \eta_L) > 0$ and root W^* ; taking expectations with $E[W] = 1/2$ gives $\omega(\eta_W - \eta_L)$, and the market-level expression follows from (12). (P3) follows from differentiating the sentiment component before substituting $\gamma = \theta$ and $\lambda = 1 - \theta$. For (P4), θ enters S only through the sentiment component and the α term, so the slope depends on D_{it} only through ω_{it} and α_{it} , and κ drops out of $\partial S/\partial\theta$; the α term contributes $-\alpha L^2$ to the slope. Evaluating the slope change at $W = \bar{W}$ on the gainer list and at $W = 1 - \bar{W}$ on the loser list and differencing gives $\psi(\eta_W - \eta_L)[\bar{W}^2 - (1 - \bar{W})^2] + \chi\bar{W}^2 = \psi(\eta_W - \eta_L)(2\bar{W} - 1) + \chi\bar{W}^2$. ■

The restriction $\eta_W > \eta_L$ delivers the steeper gain side in (P2) and, together with $\chi \geq 0$, the larger gainer-side response in (P4). An additive attention shift ($\psi = \chi = 0$, $\kappa \neq 0$) moves selling levels without changing the GainTalk–selling slope on either side of the list.

A.4 Empirical predictions and mapping to the data

Proposition 1 delivers four predictions that correspond to the empirical sections of the paper. The comparative statics are stated for raw selling, whereas the empirical outcomes are abnormal log retail sell flow and abnormal log turnover, so the mapping is qualitative: the model speaks to signs and relative magnitudes rather than to coefficient values.

Prediction 1 (Level association). GainTalk is positively associated with selling. In the stock-level tests of Section 4, a one-standard-deviation increase in GainTalk is associated with about 1.5 percent higher abnormal retail sell flow ($t = 21.98$; Table 3, column (2)). At the market level, the corresponding estimate for abnormal

turnover is $+0.0312$ ($t = 6.71$; Table 5, column (1)).

Prediction 2 (State dependence). The association is stronger in gain states on average and, qualitatively, U-shaped in the gain/loss state, with a steeper gain side. In the cross-section, $\text{GainTalk} \times \text{CGO}$ is $+0.0143$ ($t = 19.26$; Table 3, column (4)). Allowing the slope to differ between positive and negative CGO, $\text{GainTalk} \times \text{CGO}^+$ is $+0.1207$ ($t = 13.56$) and $\text{GainTalk} \times \text{CGO}^-$ is -0.0279 ($t = 9.11$) (Table H.10, column (3)): the GainTalk slope rises with embedded gains and also, more weakly, with deeper embedded losses. At the market level, $\text{GainTalk} \times R_{m,t}$ is $+0.0120$ ($t = 4.06$; Table 5, column (3)).

Prediction 3 (Matched states). Gain-framing intensity is associated with more selling where gains are more prevalent, and loss-framing intensity with less selling where losses are more prevalent. At the stock level, $\text{GainIntensity} \times \mathbf{1}[\text{CGO} > 0]$ is $+0.0126$ ($t = 10.30$) and $\text{LossIntensity} \times \mathbf{1}[\text{CGO} \leq 0]$ is -0.0165 ($t = 8.88$), while the cross-state slopes are small and insignificant (Table 4). At the market level, the matched-state slopes are $+0.0091$ ($t = 2.55$) in up markets and -0.0158 ($t = 3.56$) in down markets (Table 6).

Prediction 4 (Framing legitimacy). Repeated public classification steepens the GainTalk-selling slope on the gainer side. On the loser side, attention-induced selling by loss-framed holders offsets greater reliance on loss-framed evaluation, so the slope response is smaller and need not be positive. The display also shifts selling levels. In Section 7, D_{it} corresponds to $\text{Recur} \times \text{Post}$. Provided that recurring and non-recurring stocks do not experience differential changes in their gain/loss state around the introduction, the triple interaction corresponds to (20). On the gainer side, $\text{GainTalk} \times \text{Recur} \times \text{Post}$ is $+0.1210$ ($t = 2.64$) and $+0.0910$ ($t = 2.59$). On the loser side, the estimates are $+0.0118$ ($t = 0.38$) and $+0.0076$ ($t = 0.26$) and statistically insignificant, which is consistent with the two forces offsetting on the loser side. $\text{GainTalk} \times \text{Post}$ is insignificant, as the model predicts for non-recurring stocks. $\text{Recur} \times \text{Post}$, which captures the selling-level response, is $+0.1012$ ($t = 2.94$) and $+0.1243$ ($t = 3.06$) on the loser side and -0.0826 ($t = 2.12$) in column (1) on the gainer side (Table 10).

Distinguishing the channel. The interactions in Proposition 1 arise from the match between the frame and the gain/loss state. Measures that do not distinguish gains from losses therefore do not generate them. Consistent with this, interacting forward sentiment, RavenPack news sentiment, RavenPack news count, and social-media volume with the market return leaves the GainTalk $\times$ return coefficient at +0.0134 ($t = 3.40$; Table 8, column (5)). In the model, attention moves selling levels through A and, on the loser list, prompts loss-framed holders to sell through α . The leaderboard evidence is consistent with the framing-legitimacy interpretation: recurring gainers show the slope response predicted by an increase in ω , and recurring losers show a selling-level response with an insignificant slope response, as when α offsets the increase in ω . Attention that raises the weight holders place on gain-loss evaluation would enter the model through ω rather than A .

What the model omits. The model is silent on three issues. First, it does not model the endogenous generation of gain- and loss-framed posts: θ is taken as an observable summary of prevailing framing. Second, it does not solve for prices. The earnings-announcement evidence in Section 6 follows the logic of Grinblatt and Han (2005) and Frazzini (2006): when selling by winners works against good news and holding by losers works against bad news, prices adjust slowly. Parts (P2) and (P3) of Proposition 1 supply the trading side of that argument. Third, it treats ω as a reduced form for the public reinforcement of gain-loss framing described in Section 7 rather than deriving why repeated classification raises it.

B Corpus Coverage

Table B.1: **Sample Construction Funnel.**

The table reports the construction of the analysis corpus from the raw StockTwits archive, which runs from July 2009 through September 2024. The sample retains messages with exactly one cashtag that maps to a CRSP common stock (share codes 10–11) listed on the NYSE, AMEX, or NASDAQ (exchange codes 1–3) with a valid trading-day record on the posting date, and discards empty-text messages. These filters leave 189,273,517 messages (31.3% of the raw count). Posting volume and ticker coverage are sparse in the early years, and GainIntensity is elevated relative to later sub-periods (Table B.2), so messages posted before 2013 are dropped. The final row is the 2013 to 2024 analysis corpus: 185,096,724 messages, covering 2,253 unique tickers on the average trading day.

Step	Posts
Raw messages	603.9M
No cashtag	−287.3M
Multi-cashtag	−26.5M
Non-CRSP ticker	−100.0M
Empty text	−0.85M
Posted before 2013	−4.2M
Final	185.1M

Table B.2: Corpus Coverage by Sub-Period.

The table reports weekly StockTwits corpus statistics by sub-period, from July 2009 through September 2024. Weeks are Wednesday-ending weeks with at least one post in that window. Posts per Week is the mean weekly post count. GainIntensity and LossIntensity are the un-weighted mean, across weeks, of equation 1 applied to that week’s posts. GainTalk is the mean share of discourse-relevant posts that are gain-framed. After 2013, GainTalk is relatively stable (49.7%, 46.8%, and 44.0%) as weekly posting volume rises from 89,838 to 507,850. In 2009 to 2012, GainIntensity averages 3.7%, about 1.6 times the 2.3% in 2013 to 2016, while LossIntensity is 2.5%, already close to later sub-periods. Weekly volume in that window is only 22,745 posts: the early platform was a smaller community whose posts were more concentrated on individual positions, so a larger share of messages are classified as gain-framed. The analysis corpus therefore drops messages posted before 2013 (Table B.1).

Sub-Period	Weeks	Posts/Week	GainIntensity	LossIntensity	GainTalk
2009–2012	183	22,745	3.7%	2.5%	59.0%
2013–2016	209	89,838	2.3%	2.3%	49.7%
2017–2020	209	319,597	2.2%	2.4%	46.8%
2021–2024	196	507,850	1.8%	2.3%	44.0%
Full sample	797	237,482	2.4%	2.4%	49.7%

C Classifier Validation

Table C.1: Student Classifier Test-Set Performance.

Performance of the DeBERTa-v3-base student model against GPT-4o teacher labels on the held-out test set ($n = 9,995$), which is 10% of the labeled corpus held out by message-group (zero leakage). Panel A reports accuracy, macro-F1, and per-class F1 for the primary heads. Object is HOLDER / NONE; Frame is GAIN / LOSS / NONE; Anchor is PAST / FUTURE / UNSPECIFIED. Panel B reports joint accuracy across the primary heads. Frame-head calibration uses post-hoc temperature scaling ($T = 1.76$; ECE = 0.025 with 15 bins). Mean macro-F1 across the primary heads is $(0.877 + 0.865 + 0.848)/3 = 0.863$.

Panel A: Primary Classification Heads

Head	Classes	Accuracy	Macro-F1	Per-class F1		
				Class 1	Class 2	Class 3
Object	HOLDER / NONE	88.0%	0.877	0.875	0.879	
Frame	GAIN / LOSS / NONE	86.9%	0.865	0.867	0.846	0.882
Anchor	PAST / FUTURE / UNSPEC.	88.9%	0.848	0.863	0.759	0.923

Panel B: Joint Performance

Head	Classes	Accuracy	Macro-F1	Per-class F1		
				Class 1	Class 2	Class 3
Joint (Object $\cap$ Frame $\cap$ Anchor)	All correct	75.2%				

Table C.2: **Manual Joint Label Agreement.**

Pairwise joint agreement on a random sample of $N = 1,000$ StockTwits messages that were manually reviewed on all three primary heads. A message counts as an agreement only when the two sources match simultaneously on object (OBJECT / NONE), frame (GAIN / LOSS / NONE), and anchor (PAST / FUTURE / UNSPECIFIED). There is no external ground truth in this audit: the table reports consistency across labeling sources, not accuracy against a correct answer.

Pair	N	Agree	Joint agreement
Teacher and human	1,000	792	79.2%
Student and human	1,000	768	76.8%

D Variable Definitions

Table D reports the main variables used in the paper. The remainder of this section restates the LLM-derived constructions that build GainTalk, GainIntensity, and LossIntensity.

D.1 Per-Post Soft Scores

The student classifier assigns each post m a probability distribution over labels in three independent heads:

- *Object*: $P_m(\text{holder})$, $P_m(\text{none})$
- *Anchor*: $P_m(\text{past})$, $P_m(\text{future})$, $P_m(\text{unspecified})$
- *Frame*: $P_m(\text{gain})$, $P_m(\text{loss})$, $P_m(\text{none})$

$P_m(\text{holder})$ is the probability that the post is about a holder’s position, either the poster’s own or another holder’s. The per-post gain-message and loss-message probabilities are

$$\text{GainMessage}_m = P_m(\text{holder}) \cdot P_m(\text{gain}) \cdot P_m(\text{past}), \quad (21)$$

$$\text{LossMessage}_m = P_m(\text{holder}) \cdot P_m(\text{loss}) \cdot P_m(\text{past}). \quad (22)$$

A probability is high only when the post is simultaneously about a holder’s position, anchors to a past price, and expresses a directional gain (or loss); any dimension receiving near-zero probability drives the product to zero.

D.2 Daily Market-Level Discourse Measures

Let $\mathcal{M}_t$ be the set of all posts on day t and $N_t = |\mathcal{M}_t|$. Averaging the per-post probabilities yields the daily *level* measures (expressed as a percentage of all posts):

$$\text{GainIntensity}_t = \frac{100}{N_t} \sum_{m \in \mathcal{M}_t} \text{GainMessage}_m, \quad (23)$$

$$\text{LossIntensity}_t = \frac{100}{N_t} \sum_{m \in \mathcal{M}_t} \text{LossMessage}_m. \quad (24)$$

Their sum is the total daily reference-dependent intensity:

$$\text{Total RP}_t = \text{GainIntensity}_t + \text{LossIntensity}_t.$$

The *directional ratio* GainTalk isolates gain/loss tilt independently of total RP volume:

$$\text{GainTalk}_t = \frac{\text{GainIntensity}_t}{\text{GainIntensity}_t + \text{LossIntensity}_t} \in [0, 1], \quad (25)$$

with LossTalk defined symmetrically as $\text{LossTalk}_t = 1 - \text{GainTalk}_t$. Days with zero total RP (i.e. $\text{GainIntensity}_t = \text{LossIntensity}_t = 0$) are excluded from the sample.

D.3 Stock-Level Discourse Measures

For each stock i and date t , the stock-level measure is the same gain-to-total ratio as in equation (2), summing GainMessage_m and LossMessage_m over every message that mentions stock i :

$$\text{GainTalk}_{i,t} = \frac{\sum_{m \in (i,t)} \text{GainMessage}_m}{\sum_{m \in (i,t)} \text{GainMessage}_m + \sum_{m \in (i,t)} \text{LossMessage}_m}, \quad (26)$$

which is the construction in equation (3). Stock-days with fewer than five classified messages are excluded. When GainTalk enters the stock-day regressions in Section 4, it is standardized to mean zero and unit variance using the full-sample mean and standard deviation, unless a table notes a within-date standardization.

Table D: Variable Definitions

The table reports definitions of the main variables. Discourse measures are constructed from the DeBERTa-v3-base student classifier, as detailed in Section 2 and below. When a discourse measure or Forward Sentiment enters a regression, it is standardized to mean zero and unit variance using the full-sample mean and standard deviation unless a table notes a within-date standardization.

Variable	Definition
<i>A. Discourse measures</i>	
GainMessage	Probability that a post is about a holder’s position, frames that position as a gain, and anchors to a past price: $P_m(\text{holder}) \cdot P_m(\text{gain}) \cdot P_m(\text{past})$.
LossMessage	Same product with $P_m(\text{loss})$ in place of $P_m(\text{gain})$.
GainIntensity	Mean of GainMessage as a share of all posts, defined as in equation (1). Reported in percent in the summary-statistics tables.
LossIntensity	Mean of LossMessage as a share of all posts, defined analogously. Reported in percent in the summary-statistics tables.
Total RP	Sum of GainIntensity and LossIntensity.
GainTalk	Share of reference-dependent discourse that is gain-framed, $\text{GainIntensity}/(\text{GainIntensity} + \text{LossIntensity})$, defined as in equations (2) and (3). Reported in percent ($\times 100$) in the summary-statistics tables. Stock-days require at least five classified messages.
LossTalk	$1 - \text{GainTalk}$.
Overnight GainTalk	GainTalk from messages posted from 16:00 ET on day $t-1$ to 09:30 ET on day t .
Before-market-open GainTalk	GainTalk from messages posted from 06:00 ET to 09:30 ET.
<i>B. Market trading</i>	
Turnover (TO)	Aggregate dollar volume divided by market capitalization.
AbnLogTO	$\log(\text{TO}_t) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(\text{TO}_s)$ (window specification following Cookson and Niessner 2020), requiring a complete 121-day window.
Market return	Daily CRSP value-weighted market return, in percentage points.
R^+ / R^-	$\max(R_{m,t}, 0)$ and $\min(R_{m,t}, 0)$.
<i>C. Stock-day trading and overhang</i>	
Retail Sell Flow	Quote-midpoint retail sell volume divided by shares outstanding, in basis points (Barber et al. 2024).
AbnLogSell	$\log(1 + \text{SellFlow}_{it})$ minus its mean over available trading days in $[t-140, t-20]$ (window specification following Cookson and Niessner 2020), requiring at least 60 valid baseline days. SellFlow_{it} is Retail Sell Flow, in basis points.
Stock return	Daily CRSP stock return, in percentage points.
Log market cap	Natural logarithm of market capitalization.
Reference price (RP)	Turnover-weighted reference purchase price, defined as in equation (5) (Grinblatt and Han 2005).
Capital gain overhang (CGO)	$(\bar{P}_{i,t} - \text{RP}_{i,t})/\bar{P}_{i,t}$, defined as in equation (6).
<i>D. Sentiment, news, and attention</i>	
Forward Sentiment	$(N^{\text{bull}} - N^{\text{bear}})/(N^{\text{bull}} + N^{\text{bear}})$, defined as in equation (4) (Cookson et al. 2026). At the stock-day level, missing values are set to zero before standardization, matching the estimation sample.
RavenPack Sentiment	News-count-weighted mean of firm-day RavenPack tone (market-day); firm-day news-tone mean (stock-day). Stock-days with no news are set to zero.
RavenPack log news count	$\log(1 + N^{\text{news}})$, where N^{news} is the market-day RavenPack news count (market-day sample) or the firm-day count (stock-day sample). Stock-days with no news are set to zero.
Social-media volume	Number of classified StockTwits posts.
Log social-media volume	$\log(1 + N)$, where N is the number of classified StockTwits posts.
<i>E. Return and volatility controls</i>	
Cum. return $t-5$ to $t-1$	Cumulative return over the prior five trading days, in percentage points.
Cum. return $t-21$ to $t-6$	Cumulative return from $t-21$ to $t-6$, in percentage points.
Volatility $t-5$ to $t-1$	Market-day: mean EGARCH(1,1) conditional volatility of the CRSP value-weighted return over $t-5$ to $t-1$, in percentage points. Stock-day: standard deviation of daily CRSP returns over the same window, in percentage points.
Lagged AbnLogTO / AbnLogSell	One-day lag of the corresponding abnormal log outcome.
<i>F. Earnings news and event study</i>	
CAR $[-2, +1]$	Market-model cumulative abnormal return over $[-2, +1]$ trading days around the most recent earnings announcement.
Pre-announcement GainTalk	Stock-day GainTalk averaged over the 60 trading days ending the day before the announcement.
High-GainTalk spread	Long good-news/high-GainTalk stocks and short bad-news/low-GainTalk stocks (Frazzini 2006).
Low-GainTalk spread	Long good-news/low-GainTalk stocks and short bad-news/high-GainTalk stocks.
Recur	Equals one when the stock was on the same side of the leaderboard on the previous trading day.
Post	Equals one from 2023-11-11 in the leaderboard event-study sample.

E Classifier Label Distribution

Table E reports the distribution of the student classifier’s most likely labels in the 2013–2024 analysis sample. Panel A gives the marginal share of each label in the three primary heads. Panel B lists the four cells in which the most likely labels identify a holder’s position together with a definite frame and anchor; those cells account for 8.39% of the analysis sample, and the two past-anchored gain and loss cells account for 4.78%.

Table E: Classifier Label Distribution.

Panel A reports the share of all 185.1 million classified posts in the 2013–2024 analysis sample whose most likely label on each head equals the listed class. Object is HOLDER / NONE; Frame is GAIN / LOSS / NONE; Anchor is PAST / FUTURE / UNSPECIFIED. Panel B reports the four cells in which the most likely labels identify a holder’s position together with a definite frame and a definite anchor. Those cells account for 8.39% of the analysis sample; the two past-anchored gain and loss cells account for 4.78%. GainIntensity and LossIntensity average the per-message probability products over all posts.

Panel A: Marginal Label Distributions ($N = 185,096,724$ posts)

Head	Label	Share
Frame	Gain	24.5%
	Loss	13.2%
	None	62.3%
Anchor	Past	7.3%
	Future	10.9%
	Unspecified	81.8%
Object	Holder	10.5%
	None	89.5%

Panel B: Definite Frame-Anchor Combinations

#	Frame	Anchor	Count	% total	% of 4
1	Loss	Past	4,881,445	2.64	31.43
2	Gain	Future	4,366,305	2.36	28.11
3	Gain	Past	3,963,374	2.14	25.52
4	Loss	Future	2,319,953	1.25	14.94
Total			15,531,077	8.39	100.00

F Textual Validation

Table F: **Most Salient Phrases by Classifier Category.**

This table lists the 30 most salient words and phrases for each of four message categories, constructed from a random sample of 2 million messages from the classified corpus. Phrases are ranked by prevalence: the share of messages in the focal category that contain the phrase, among unigrams and bigrams that are more common in that category than in the complementary category. Panel A reports the reference-dependent thinking categories *Gain-framed* and *Loss-framed* (gain/loss frame, a holder’s position, past temporal anchor), defined by the DeBERTa-v3 classifier’s predicted probabilities. Panel B reports the future-anchored counterparts (gain or loss frame, future temporal anchor), defined by the same DeBERTa-v3 classifier. Text is cleaned by decoding HTML entities, removing cashtags, @mentions, URLs, and duplicates before extracting phrases.

Panel A: Reference-Dependent Thinking (Backward-Looking, Past-Anchored)

Category	Most salient words and phrases (ranked by prevalence)
<i>Gain-framed</i>	<i>huge gain, nice gain, good gains, so glad, my average, in the green, up over, holding from, in profit, long from, way up, sitting on, am up, paying off, above my, my entry, cost basis, initial investment, from entry, made back, loving this, almost sold, green again, held overnight, original position, currently up, finally green, held from, unrealized, glad held</i>
<i>Loss-framed</i>	<i>big loss, huge loss, break even, lost everything, keep losing, total loss, biggest loss, massive loss, am losing, at loss, paper loss, lost thousands, lost lot, bag holder, bag holding, am down, still down, underwater, recoup, down big, only red, being down, currently down, am stuck, down 10k, stuck holding, down 20k, far down, am bag, being bag</i>

Table F: **Most Salient Phrases by Classifier Category.** (continued)

Panel B: Future-Anchored Outlook (Gain/Loss Frame, Future Anchor)

Category	Most salient words and phrases (ranked by prevalence)
<i>Gain-framed</i> (future)	<i>green soon, explode soon, higher soon, moon tomorrow, highs soon, fly tomorrow, fly past, plus tomorrow, gain tomorrow, dollars soon, tomorrow easy, 10x from, rip soon, rocket soon, double soon, highs next, 2+ soon, 2+ tomorrow, 20+ tomorrow, rocket tomorrow, share next, 1+ soon, green next, plus soon, rocket past, easy tomorrow, gonna spike, highs tomorrow, tomorrow guaranteed, soon load</i>
<i>Loss-framed</i> (future)	<i>dump coming, fall hard, red soon, dump tomorrow, dump hard, headed lower, crash coming, down coming, bad apples, downsides, dump incoming, bankruptcy soon, facing significant, significant downsides, drop incoming, heading below, downhill from, dump soon, downside coming, drop fast, down incoming, dump time, next dump, bankruptcy imminent, bloodbath tomorrow, fall soon, headed below, crash soon, under soon, below soon</i>

G Trading Activity and Controls

Table G reports summary statistics for trading outcomes, firm characteristics, and the lagged return and volatility controls in the market-day and stock-day samples of Section 3.

Table G: Summary Statistics: Trading Activity and Controls

The table reports summary statistics for trading outcomes, firm characteristics, and the lagged return and volatility controls. Panel A uses the daily market-level sample (April 2013 to September 2024). Panel B uses the stock-day panel used in the cross-sectional tests of Section 4 (April 2015 to September 2024), formed by merging CRSP with retail sell flow identified from off-exchange trades using the quote-midpoint method (Barber et al. 2024), requiring at least five classified StockTwits posts per stock-day. The abnormal sell-flow baseline uses the 140 trading days before t , requiring at least 60 valid baseline days. Variable definitions are in Table D.

Variable	N	Mean	SD	P10	P25	Median	P75	P90
<i>Panel A: Market-Day</i>								
Abnormal Log Turnover (AbnLogTO)	2,737	-0.004	0.190	-0.225	-0.132	-0.019	0.097	0.236
Turnover (%)	2,737	0.750	0.150	0.598	0.653	0.724	0.818	0.926
Market Return (%)	2,737	0.064	0.879	-1.006	-0.371	0.073	0.567	1.129
Lagged Abnormal Log Turnover (AbnLogTO)	2,737	-0.004	0.190	-0.225	-0.132	-0.019	0.097	0.236
Cum. Return $t-5$ to $t-1$ (%)	2,737	0.321	1.942	-2.055	-0.678	0.462	1.463	2.524
Cum. Return $t-21$ to $t-6$ (%)	2,737	1.013	3.318	-2.951	-0.801	1.253	2.941	4.743
Volatility $t-5$ to $t-1$ (%)	2,737	0.923	0.402	0.544	0.643	0.816	1.117	1.458
<i>Panel B: Stock-Day</i>								
Abn. Log Retail Sell Flow	1,891,685	0.190	0.852	-0.587	-0.259	0.033	0.471	1.157
Turnover (%)	1,891,685	2.362	6.690	0.311	0.538	1.024	2.067	4.169
Daily Return (%)	1,891,685	0.223	6.449	-4.523	-1.822	0.015	1.896	4.792
Log Market Cap	1,891,685	7.481	2.537	4.154	5.636	7.477	9.332	10.883
Market Cap (\$M)	1,891,685	25,655.6	114,537.1	63.7	280.3	1,766.8	11,290.6	53,272.1
Capital Gain Overhang	1,891,685	-0.106	0.526	-0.456	-0.174	0.000	0.120	0.210
Lagged Abn. Log Retail Sell Flow	1,891,685	0.136	0.833	-0.622	-0.288	-0.003	0.399	1.044
Cum. Return $t-5$ to $t-1$ (%)	1,891,685	0.521	13.452	-10.045	-3.980	0.238	4.236	10.319
Cum. Return $t-21$ to $t-6$ (%)	1,891,685	1.486	24.034	-17.241	-6.724	0.809	7.964	18.589
Volatility $t-5$ to $t-1$ (%)	1,891,685	3.544	4.809	0.905	1.472	2.523	4.241	6.756

H Robustness and Extensions for Stock-Level GainTalk and Abnormal Retail Sell Flow

Table H.1: **GainTalk, Capital Gain Overhang, and Abnormal Net Retail Flow**

The table reports the specifications in Table 3, with the dependent variable replaced by abnormal net retail flow. Abnormal net retail flow is $\text{SellFlow}_{it} - \text{BuyFlow}_{it}$ relative to its mean over $[t-140, t-20]$, where $\text{SellFlow}_{it} = 10,000 \times \text{MRSVol}_{it} / \text{SHROUT}_{it}$ and $\text{BuyFlow}_{it} = 10,000 \times \text{MRBVol}_{it} / \text{SHROUT}_{it}$ are quote-midpoint retail sell and buy volume over shares outstanding, in basis points (Barber et al. 2024). The lagged dependent variable is the lag of this net flow. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0624*** (6.68)	+0.0355*** (4.19)	+0.0625*** (6.68)	+0.0355*** (4.19)		
BMO GainTalk					+0.0978*** (8.66)	
Overnight GainTalk						+0.0587*** (7.50)
GainTalk $\times$ CGO			+0.0273*** (3.81)	+0.0289*** (4.05)		
BMO GT $\times$ CGO					+0.0402*** (2.99)	
Overnight GT $\times$ CGO						+0.0246*** (2.99)
CGO	-0.1170*** (8.84)	-0.3173*** (11.58)	-0.1156*** (8.87)	-0.3162*** (11.64)	-0.4154*** (11.80)	-0.3831*** (12.87)
Forward sentiment	-0.0861*** (9.09)	-0.0297*** (4.36)	-0.0861*** (9.09)	-0.0297*** (4.36)	+0.0001 (0.01)	-0.0068 (1.03)
Log social-media volume	-0.0017 (0.07)	+0.2847*** (7.83)	-0.0019 (0.08)	+0.2845*** (7.82)	+0.3348*** (7.64)	+0.3076*** (7.92)
RavenPack sentiment	-0.0073 (0.36)	+0.0559*** (2.73)	-0.0083 (0.41)	+0.0552*** (2.70)	+0.1422*** (5.14)	+0.0720*** (3.21)
RavenPack log news count	+0.1579*** (8.36)	-0.0839*** (5.33)	+0.1580*** (8.36)	-0.0839*** (5.32)	-0.0963*** (4.70)	-0.0900*** (5.16)
Same-day return	+0.0941*** (17.13)	+0.1015*** (16.82)	+0.0942*** (17.09)	+0.1016*** (16.79)	+0.1050*** (16.38)	+0.1067*** (17.89)
Cum. return $t-5$ to $t-1$	+1.8508*** (12.95)	+1.3811*** (8.40)	+1.8515*** (12.95)	+1.3811*** (8.40)	+1.5415*** (7.77)	+1.3257*** (7.62)
Cum. return $t-21$ to $t-6$	+0.1014 (1.23)	+0.2115** (1.97)	+0.1011 (1.23)	+0.2111** (1.97)	+0.8721*** (6.36)	+0.8041*** (6.86)
Volatility ($t-5$ to $t-1$)	-6.7771*** (15.28)	-3.3599*** (7.28)	-6.7764*** (15.28)	-3.3591*** (7.28)	-5.9747*** (7.48)	-5.5552*** (7.75)
Lagged AbnNet	+0.2399*** (31.98)	+0.1987*** (31.84)	+0.2399*** (31.98)	+0.1987*** (31.84)	+0.1986*** (29.95)	+0.1977*** (31.50)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,891,685	1,238,017	1,697,519
Adj. R^2	0.063	0.093	0.063	0.093	0.097	0.096

Table H.2: GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow: S&P 500 Firms

The table reports the specifications in Table 3, with the sample restricted to stock-days on which the firm is a point-in-time S&P 500 constituent. GainTalk and CGO are standardized within this sample. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0013 (1.61)	−0.0001 (0.15)	+0.0016* (1.81)	+0.0001 (0.16)		
BMO GainTalk					−0.0009 (1.24)	
Overnight GainTalk						−0.0029*** (5.24)
GainTalk × CGO			+0.0151*** (16.23)	+0.0113*** (17.22)		
BMO GT × CGO					+0.0074*** (9.20)	
Overnight GT × CGO						+0.0061*** (10.60)
CGO	−0.0320*** (12.91)	−0.0224*** (8.15)	−0.0305*** (12.38)	−0.0214*** (7.76)	−0.0220*** (7.88)	−0.0217*** (7.84)
Forward sentiment	−0.0037** (2.20)	−0.0040*** (4.04)	−0.0036** (2.19)	−0.0038*** (3.93)	−0.0038*** (3.81)	−0.0033*** (3.34)
Log social-media volume	+0.0336*** (5.54)	+0.1580*** (29.30)	+0.0324*** (5.41)	+0.1565*** (29.12)	+0.1566*** (29.22)	+0.1570*** (29.13)
RavenPack sentiment	+0.0250*** (5.15)	+0.0195*** (7.10)	+0.0253*** (5.23)	+0.0197*** (7.20)	+0.0199*** (7.27)	+0.0198*** (7.23)
RavenPack log news count	+0.0640*** (12.36)	+0.0463*** (22.35)	+0.0640*** (12.42)	+0.0463*** (22.43)	+0.0461*** (22.33)	+0.0463*** (22.35)
Same-day return	+0.0006 (0.60)	+0.0011 (1.44)	+0.0010 (0.95)	+0.0013* (1.82)	+0.0013* (1.73)	+0.0012 (1.60)
Cum. return $t-5$ to $t-1$	−0.1624*** (4.29)	−0.1394*** (4.78)	−0.1506*** (4.01)	−0.1319*** (4.52)	−0.1338*** (4.61)	−0.1257*** (4.34)
Cum. return $t-21$ to $t-6$	−0.0914*** (4.30)	−0.0987*** (4.88)	−0.0905*** (4.27)	−0.0982*** (4.85)	−0.0977*** (4.82)	−0.0974*** (4.82)
Volatility ($t-5$ to $t-1$)	+0.4873*** (3.19)	−0.1220 (1.08)	+0.4918*** (3.24)	−0.1007 (0.90)	−0.1118 (1.00)	−0.1101 (0.98)
Lagged AbnLogSell	+0.5907*** (74.23)	+0.4948*** (63.29)	+0.5887*** (74.19)	+0.4933*** (63.23)	+0.4943*** (63.17)	+0.4939*** (63.09)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	464,084	464,084	464,084	464,084	303,977	416,283
Adj. R^2	0.451	0.547	0.453	0.548	0.548	0.548

Table H.3: GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow: At Least One Reference-Dependent Message

The table reports the specifications in Table 3, with the sample restricted to stock-days on which the sum of same-day GainMessage and LossMessage is at least one. GainTalk and CGO are standardized within this sample. Everything else follows Table 3. *t*-statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0445*** (16.18)	+0.0408*** (21.01)	+0.0440*** (15.59)	+0.0400*** (20.07)		
BMO GainTalk					+0.0105*** (7.15)	
Overnight GainTalk						+0.0099*** (7.15)
GainTalk $\times$ CGO			+0.0140*** (5.78)	+0.0174*** (9.84)		
BMO GT $\times$ CGO					+0.0243*** (11.97)	
Overnight GT $\times$ CGO						+0.0177*** (10.15)
CGO	-0.1466*** (20.79)	-0.1166*** (19.65)	-0.1431*** (20.39)	-0.1122*** (18.89)	-0.1116*** (18.75)	-0.1110*** (18.62)
Forward sentiment	-0.0031 (0.67)	-0.0211*** (10.12)	-0.0030 (0.64)	-0.0208*** (10.01)	-0.0210*** (9.95)	-0.0205*** (9.88)
Log social-media volume	+0.0664*** (6.77)	+0.3267*** (28.59)	+0.0669*** (6.82)	+0.3270*** (28.62)	+0.3262*** (28.48)	+0.3268*** (28.55)
RavenPack sentiment	+0.0727*** (5.39)	+0.0427*** (6.20)	+0.0715*** (5.30)	+0.0416*** (6.04)	+0.0420*** (6.10)	+0.0413*** (5.98)
RavenPack log news count	+0.1734*** (15.11)	+0.0896*** (13.90)	+0.1733*** (15.11)	+0.0896*** (13.89)	+0.0902*** (13.98)	+0.0892*** (13.85)
Same-day return	+0.0171*** (24.66)	+0.0121*** (20.42)	+0.0171*** (24.63)	+0.0121*** (20.40)	+0.0120*** (20.28)	+0.0121*** (20.35)
Cum. return $t-5$ to $t-1$	+0.2478*** (11.36)	+0.2012*** (10.82)	+0.2433*** (11.22)	+0.1950*** (10.53)	+0.1910*** (10.35)	+0.1935*** (10.48)
Cum. return $t-21$ to $t-6$	+0.1563*** (6.11)	+0.1082*** (5.71)	+0.1550*** (6.11)	+0.1065*** (5.72)	+0.1040*** (5.65)	+0.1058*** (5.78)
Volatility ($t-5$ to $t-1$)	-0.1341** (2.12)	-0.4486*** (9.59)	-0.1234** (1.96)	-0.4350*** (9.33)	-0.4280*** (9.18)	-0.4320*** (9.25)
Lagged AbnLogSell	+0.7364*** (160.30)	+0.5625*** (80.01)	+0.7365*** (160.49)	+0.5625*** (80.08)	+0.5618*** (79.85)	+0.5622*** (79.92)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	941,797	941,797	941,797	941,797	788,218	834,485
Adj. R^2	0.587	0.700	0.587	0.700	0.717	0.704

Table H.4: GainTalk, Capital Gain Overhang, and Next-Day Abnormal Log Retail Sell Flow

The table reports the specifications in Table 3, with the dependent variable replaced by next-day abnormal log retail sell flow, measured on $t+1$. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0068*** (8.22)	+0.0074*** (11.25)	+0.0069*** (8.02)	+0.0075*** (10.91)		
BMO GainTalk					+0.0045*** (7.13)	
Overnight GainTalk						+0.0037*** (5.71)
GainTalk $\times$ CGO			+0.0093*** (9.62)	+0.0103*** (12.69)		
BMO GT $\times$ CGO					+0.0094*** (11.51)	
Overnight GT $\times$ CGO						+0.0070*** (9.03)
CGO	-0.0763*** (27.99)	-0.0755*** (25.40)	-0.0749*** (27.55)	-0.0737*** (24.84)	-0.0732*** (24.65)	-0.0729*** (24.52)
Forward sentiment	-0.0071*** (6.26)	-0.0054*** (7.89)	-0.0071*** (6.28)	-0.0054*** (7.86)	-0.0055*** (7.92)	-0.0053*** (7.78)
Log social-media volume	+0.0508*** (11.39)	+0.1863*** (38.91)	+0.0505*** (11.36)	+0.1860*** (38.89)	+0.1854*** (38.75)	+0.1857*** (38.82)
RavenPack sentiment	+0.0316*** (8.10)	+0.0215*** (7.99)	+0.0306*** (7.85)	+0.0204*** (7.60)	+0.0207*** (7.68)	+0.0202*** (7.55)
RavenPack log news count	+0.1009*** (22.90)	+0.0523*** (18.24)	+0.1010*** (22.93)	+0.0523*** (18.27)	+0.0528*** (18.35)	+0.0524*** (18.30)
Same-day return	+0.0089*** (24.90)	+0.0071*** (23.98)	+0.0089*** (25.09)	+0.0071*** (24.17)	+0.0070*** (24.05)	+0.0071*** (24.12)
Cum. return $t-5$ to $t-1$	+0.1393*** (8.14)	+0.1368*** (10.14)	+0.1400*** (8.19)	+0.1368*** (10.16)	+0.1345*** (10.05)	+0.1358*** (10.10)
Cum. return $t-21$ to $t-6$	+0.1689*** (10.59)	+0.1459*** (9.27)	+0.1685*** (10.59)	+0.1453*** (9.27)	+0.1438*** (9.15)	+0.1448*** (9.22)
Volatility ($t-5$ to $t-1$)	+0.3489*** (7.05)	+0.0085 (0.24)	+0.3498*** (7.07)	+0.0105 (0.30)	+0.0112 (0.32)	+0.0098 (0.28)
Lagged AbnLogSell	+0.5484*** (135.08)	+0.4705*** (108.49)	+0.5484*** (135.07)	+0.4704*** (108.47)	+0.4700*** (108.35)	+0.4702*** (108.40)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,886,172	1,886,172	1,886,172	1,886,172	1,234,409	1,692,571
Adj. R^2	0.349	0.407	0.350	0.407	0.409	0.409

Table H.5: GainTalk, Capital Gain Overhang, and Δ Abnormal Log Retail Sell Flow

The table reports the specifications in Table 3, with the dependent variable replaced by its one-day change, $\Delta \text{AbnLogSell}_{it} = \text{AbnLogSell}_{it} - \text{AbnLogSell}_{i,t-1}$. The lagged dependent variable is the lag of this change. Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0077*** (11.23)	+0.0082*** (14.79)	+0.0078*** (10.75)	+0.0083*** (14.20)		
BMO GainTalk					-0.0032*** (5.88)	
Overnight GainTalk						-0.0075*** (14.40)
GainTalk $\times$ CGO			+0.0132*** (18.32)	+0.0130*** (20.93)		
BMO GT $\times$ CGO					+0.0118*** (14.74)	
Overnight GT $\times$ CGO						+0.0056*** (8.73)
CGO	-0.0573*** (34.02)	-0.0497*** (32.23)	-0.0553*** (33.08)	-0.0475*** (31.10)	-0.0472*** (30.95)	-0.0469*** (30.82)
Forward sentiment	-0.0066*** (6.85)	-0.0083*** (14.38)	-0.0067*** (6.89)	-0.0083*** (14.34)	-0.0084*** (14.42)	-0.0082*** (14.28)
Log social-media volume	+0.0612*** (15.16)	+0.1554*** (48.92)	+0.0609*** (15.14)	+0.1550*** (48.95)	+0.1544*** (48.82)	+0.1547*** (48.88)
RavenPack sentiment	+0.0434*** (12.52)	+0.0358*** (14.67)	+0.0420*** (12.13)	+0.0345*** (14.16)	+0.0348*** (14.22)	+0.0342*** (14.08)
RavenPack log news count	+0.0849*** (23.13)	+0.0580*** (28.91)	+0.0851*** (23.18)	+0.0581*** (29.00)	+0.0585*** (29.08)	+0.0578*** (28.95)
Same-day return	+0.0100*** (27.50)	+0.0087*** (27.74)	+0.0101*** (27.65)	+0.0088*** (27.87)	+0.0087*** (27.80)	+0.0088*** (27.85)
Cum. return $t-5$ to $t-1$	+0.2651*** (20.17)	+0.3026*** (24.35)	+0.2661*** (20.34)	+0.3027*** (24.42)	+0.2985*** (24.15)	+0.3005*** (24.28)
Cum. return $t-21$ to $t-6$	+0.0386*** (7.72)	+0.0023 (0.69)	+0.0381*** (7.69)	+0.0015 (0.45)	+0.0020 (0.48)	+0.0018 (0.52)
Volatility ($t-5$ to $t-1$)	-1.7288*** (25.12)	-2.3561*** (25.05)	-1.7278*** (25.17)	-2.3543*** (25.06)	-2.3480*** (24.95)	-2.3510*** (25.00)
Lagged $\Delta \text{AbnLogSell}$	-0.2193*** (107.94)	-0.2389*** (134.81)	-0.2194*** (108.03)	-0.2390*** (134.91)	-0.2388*** (134.85)	-0.2389*** (134.88)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,886,172	1,886,172	1,886,172	1,886,172	1,234,409	1,692,571
Adj. R^2	0.098	0.157	0.098	0.157	0.140	0.139

Table H.6: Within-Date Standardized GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow

The table reports the specifications in Table 3, with GainTalk and CGO standardized within date rather than over the full sample. Everything else follows Table 3. *t*-statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0173*** (20.24)	+0.0150*** (22.05)	+0.0176*** (19.52)	+0.0152*** (21.10)		
BMO GainTalk					+0.0009 (1.36)	
Overnight GainTalk						+0.0004 (0.64)
GainTalk $\times$ CGO			+0.0119*** (14.26)	+0.0131*** (18.73)		
BMO GT $\times$ CGO					+0.0153*** (16.77)	
Overnight GT $\times$ CGO						+0.0124*** (17.05)
CGO	-0.0529*** (26.14)	-0.0486*** (19.70)	-0.0517*** (25.54)	-0.0473*** (19.19)	-0.0470*** (19.05)	-0.0468*** (18.92)
Forward sentiment	-0.0101*** (8.22)	-0.0055*** (8.47)	-0.0102*** (8.29)	-0.0054*** (8.43)	-0.0055*** (8.48)	-0.0053*** (8.35)
Log social-media volume	+0.0862*** (14.78)	+0.2551*** (43.06)	+0.0858*** (14.76)	+0.2547*** (43.05)	+0.2540*** (42.92)	+0.2544*** (42.98)
RavenPack sentiment	+0.0326*** (8.30)	+0.0220*** (8.44)	+0.0309*** (7.88)	+0.0202*** (7.80)	+0.0205*** (7.88)	+0.0200*** (7.75)
RavenPack log news count	+0.0935*** (22.17)	+0.0272*** (9.73)	+0.0936*** (22.22)	+0.0273*** (9.77)	+0.0276*** (9.82)	+0.0270*** (9.70)
Same-day return	+0.0113*** (26.76)	+0.0093*** (25.62)	+0.0114*** (26.80)	+0.0094*** (25.70)	+0.0093*** (25.58)	+0.0094*** (25.65)
Cum. return $t-5$ to $t-1$	+0.0731*** (4.58)	+0.0799*** (7.51)	+0.0740*** (4.63)	+0.0802*** (7.54)	+0.0788*** (7.42)	+0.0795*** (7.48)
Cum. return $t-21$ to $t-6$	+0.1132*** (10.58)	+0.0994*** (9.10)	+0.1129*** (10.58)	+0.0988*** (9.10)	+0.0975*** (8.98)	+0.0982*** (9.05)
Volatility ($t-5$ to $t-1$)	+0.4748*** (9.19)	+0.1316*** (4.28)	+0.4743*** (9.16)	+0.1323*** (4.29)	+0.1335*** (4.32)	+0.1318*** (4.26)
Lagged AbnLogSell	+0.6623*** (208.87)	+0.5803*** (165.94)	+0.6622*** (208.82)	+0.5802*** (165.94)	+0.5795*** (165.75)	+0.5798*** (165.82)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,891,685	1,238,017	1,697,519
Adj. R^2	0.504	0.568	0.504	0.568	0.571	0.571

Table H.7: Firm-Equal-Weighted GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow

The table reports the specifications in Table 3, estimated by firm-equal-weighted WLS. Each observation is weighted by $w_{it} = 1/n_i$, where n_i is the number of estimation stock-days contributed by firm i . Everything else follows Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0200*** (6.29)	+0.0191*** (10.42)	+0.0256*** (8.34)	+0.0236*** (13.12)		
BMO GainTalk					−0.0008 (0.46)	
Overnight GainTalk						+0.0018 (1.11)
GainTalk × CGO			+0.0187*** (7.09)	+0.0162*** (10.80)		
BMO GT × CGO					+0.0126*** (7.29)	
Overnight GT × CGO						+0.0114*** (7.12)
CGO	−0.0568*** (15.41)	−0.0550*** (18.01)	−0.0550*** (15.06)	−0.0530*** (17.36)	−0.0525*** (17.22)	−0.0522*** (17.10)
Forward sentiment	+0.0018 (0.77)	+0.0068*** (4.53)	+0.0017 (0.71)	+0.0068*** (4.50)	+0.0069*** (4.55)	+0.0067*** (4.45)
Log social-media volume	+0.1587*** (23.88)	+0.3646*** (68.96)	+0.1578*** (23.81)	+0.3637*** (68.91)	+0.3628*** (68.75)	+0.3632*** (68.82)
RavenPack sentiment	+0.0491*** (5.14)	+0.0166*** (2.83)	+0.0456*** (4.80)	+0.0134** (2.29)	+0.0138** (2.35)	+0.0131** (2.22)
RavenPack log news count	+0.1046*** (20.39)	+0.0041 (1.19)	+0.1042*** (20.49)	+0.0038 (1.09)	+0.0042 (1.15)	+0.0045 (1.22)
Same-day return	+0.0149*** (15.55)	+0.0099*** (14.70)	+0.0151*** (15.69)	+0.0101*** (14.82)	+0.0100*** (14.75)	+0.0099*** (14.68)
Cum. return $t-5$ to $t-1$	+0.0237 (0.97)	+0.0290 (1.64)	+0.0253 (1.01)	+0.0307* (1.73)	+0.0295* (1.68)	+0.0302* (1.71)
Cum. return $t-21$ to $t-6$	+0.0843*** (6.37)	+0.0907*** (7.01)	+0.0837*** (6.30)	+0.0897*** (6.97)	+0.0885*** (6.88)	+0.0892*** (6.94)
Volatility ($t-5$ to $t-1$)	+0.5556*** (6.63)	+0.2309*** (4.48)	+0.5546*** (6.54)	+0.2294*** (4.42)	+0.2310*** (4.45)	+0.2288*** (4.38)
Lagged AbnLogSell	+0.6295*** (111.59)	+0.5075*** (128.75)	+0.6295*** (111.96)	+0.5074*** (128.62)	+0.5068*** (128.48)	+0.5070*** (128.55)
Firm FE	No	Yes	No	Yes	Yes	Yes
Date FE	No	Yes	No	Yes	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,891,685	1,238,017	1,697,519
Adj. R^2	0.470	0.586	0.471	0.587	0.590	0.590

Table H.8: Fama-MacBeth GainTalk, Capital Gain Overhang, and Abnormal Log Retail Sell Flow

The table reports Fama-MacBeth estimates of same-day abnormal log retail sell flow on full-sample standardized GainTalk, full-sample standardized CGO, and their interaction. Each trading day the outcome is regressed cross-sectionally by OLS with a day-specific intercept; reported coefficients are averages of the daily slope estimates. Col. (1) includes GainTalk and CGO; col. (2) adds their interaction; cols. (3) and (4) replace GainTalk with BMO GainTalk and overnight GainTalk, as in Table 3. Firm fixed effects are not included, because each daily cross-section has one observation per firm. Abnormal log retail sell flow, GainTalk, CGO, controls, and sample are as defined in that table. Days is the number of daily cross-sections. t -statistics in parentheses; Newey-West standard errors on the daily coefficient series with lag $\lfloor 4(T/100)^{2/9} \rfloor$. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)
GainTalk	+0.0137*** (22.71)	+0.0128*** (19.38)		
BMO GainTalk			−0.0008 (0.77)	
Overnight GainTalk				+0.0025*** (3.27)
GainTalk × CGO		+0.0174*** (16.54)		
BMO GT × CGO			+0.0235*** (16.56)	
Overnight GT × CGO				+0.0136*** (13.87)
CGO	−0.0749*** (37.26)	−0.0729*** (35.77)	−0.0725*** (35.55)	−0.0722*** (35.42)
Forward sentiment	−0.0102*** (13.23)	−0.0102*** (13.32)	−0.0103*** (13.38)	−0.0101*** (13.25)
Log social-media volume	+0.0916*** (55.97)	+0.0911*** (55.86)	+0.0905*** (55.62)	+0.0908*** (55.72)
RavenPack sentiment	+0.0371*** (14.01)	+0.0356*** (13.54)	+0.0359*** (13.62)	+0.0354*** (13.48)
RavenPack log news count	+0.0863*** (44.24)	+0.0865*** (44.68)	+0.0868*** (44.82)	+0.0862*** (44.58)
Same-day return	+0.0154*** (25.34)	+0.0156*** (25.76)	+0.0155*** (25.68)	+0.0155*** (25.62)
Cum. return $t-5$ to $t-1$	+0.1490*** (13.64)	+0.1521*** (13.70)	+0.1505*** (13.62)	+0.1512*** (13.66)
Cum. return $t-21$ to $t-6$	+0.1536*** (21.16)	+0.1535*** (20.98)	+0.1528*** (20.92)	+0.1530*** (20.95)
Volatility ($t-5$ to $t-1$)	+0.5090*** (15.67)	+0.5119*** (15.74)	+0.5150*** (15.82)	+0.5095*** (15.70)
Lagged AbnLogSell	+0.6389*** (234.97)	+0.6385*** (233.60)	+0.6380*** (233.35)	+0.6382*** (233.48)
Days	2,331	2,331	2,331	2,331
Avg. Adj. R^2	0.508	0.510	0.519	0.518

Table H.9: GainTalk $\times$ CGO with Nonlinear Capital Gain Overhang

The table reports stock-day OLS estimates of same-day abnormal log retail sell flow on full-sample standardized GainTalk, full-sample standardized CGO, and their interaction, adding nonlinear terms in CGO. Col. (1) is column (4) of Table 3. Col. (2) adds the square of standardized CGO. Cols. (3) and (4) replace the linear CGO term with a linear spline in CGO with a knot at zero, $\max(\text{CGO}, 0)$ and $\min(\text{CGO}, 0)$; col. (4) also includes $1[\text{CGO} > 0]$. Col. (5) keeps the linear CGO term, adds the square of standardized CGO, and includes $1[\text{CGO} > 0]$. Col. (6) replaces the linear CGO term with indicators for deciles of CGO. GainTalk $\times$ CGO is the product of full-sample standardized GainTalk and full-sample standardized CGO in every column. In this sample the correlation between standardized GainTalk and standardized CGO is +0.139. Everything else follows Table 3. *t*-statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1) Linear	(2) + CGO ²	(3) Spline	(4) Spline + jump	(5) CGO ² + jump	(6) Decile FE
GainTalk	+0.0152*** (20.69)	+0.0136*** (19.55)	+0.0130*** (19.01)	+0.0132*** (19.36)	+0.0136*** (19.61)	+0.0134*** (19.41)
GainTalk $\times$ CGO	+0.0143*** (19.26)	+0.0113*** (16.10)	+0.0087*** (12.87)	+0.0088*** (12.99)	+0.0113*** (16.07)	+0.0125*** (17.54)
CGO	-0.0592*** (22.50)	-0.0218*** (6.17)			-0.0241*** (6.22)	
CGO ²		+0.0123*** (16.37)			+0.0119*** (14.86)	
$\max(\text{CGO}, 0)$			+0.2941*** (16.52)	+0.3543*** (18.84)		
$\min(\text{CGO}, 0)$			-0.2232*** (29.85)	-0.2174*** (29.08)		
$1[\text{CGO} > 0]$				-0.0187*** (6.82)	+0.0038 (1.38)	
CGO decile FE	No	No	No	No	No	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,891,685	1,891,685	1,891,685
Adj. R^2	0.569	0.570	0.571	0.571	0.570	0.569

Table H.10: GainTalk, Capital Gain Overhang, and the Selling V-Shape

The table reports stock-day OLS estimates of abnormal log retail sell flow on CGO^- , CGO^+ , standardized GainTalk, and GainTalk $\times$ CGO-arm interactions. Cols. (1)–(3) use same-day AbnLogSell; col. (4) uses next-day AbnLogSell, measured on $t+1$, as in Table H.4. $AbnLogSell_{it}$ is $\log(1 + SellFlow_{it})$ minus its mean over $[t-140, t-20]$, where $SellFlow_{it}$ is defined in Table 3. $CGO_{it}^- = \min(CGO_{it}, 0)$ and $CGO_{it}^+ = \max(CGO_{it}, 0)$, where $CGO_{it} = (\hat{P}_{it} - RP_{it})/\hat{P}_{it}$ (Grinblatt and Han 2005). Controls follow Table 3. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)
CGO^-	−0.1057*** (8.47)	−0.2127*** (14.60)	−0.1799*** (13.48)	−0.2706*** (19.87)
CGO^+	+0.6742*** (18.11)	+0.5210*** (13.71)	+0.6531*** (18.30)	+0.5391*** (15.80)
GainTalk	+0.0256*** (13.40)	+0.0202*** (13.19)	+0.0212*** (14.81)	+0.0103*** (8.02)
GainTalk $\times$ CGO^-	−0.0402*** (9.80)	−0.0272*** (8.32)	−0.0279*** (9.11)	−0.0173*** (5.39)
GainTalk $\times$ CGO^+	+0.1128*** (9.59)	+0.1303*** (13.42)	+0.1207*** (13.56)	+0.0883*** (10.34)
Controls	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	Yes
Date FE	No	No	Yes	Yes
Observations	1,891,685	1,891,685	1,891,685	1,886,172
Adj. R^2	0.505	0.566	0.568	0.407

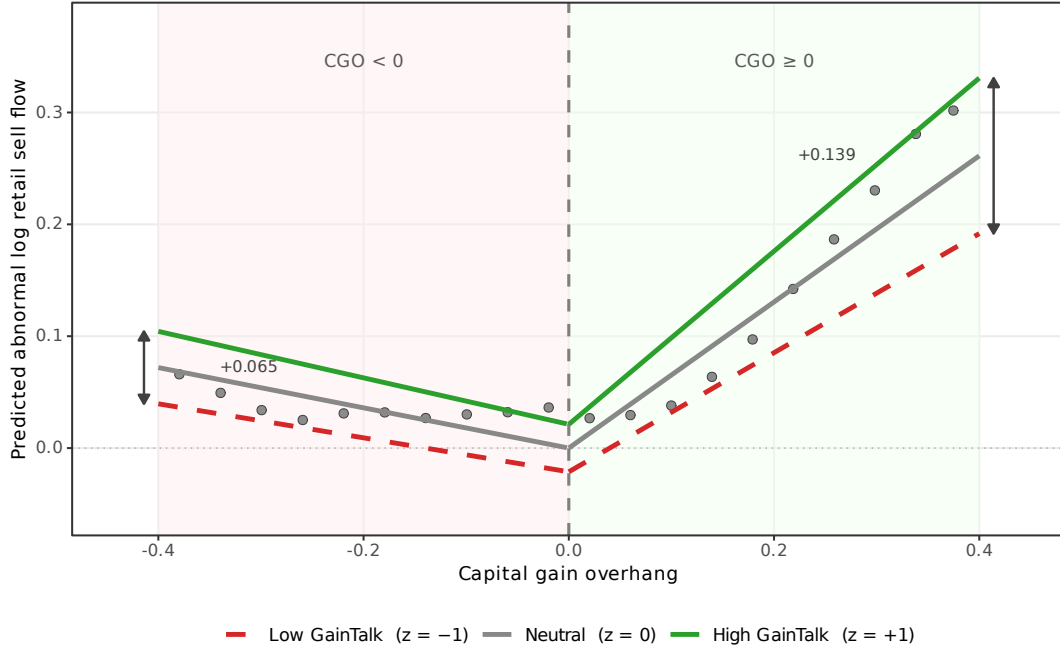


Figure H.1: GainTalk and the Selling V-Shape

The figure plots predicted same-day abnormal log retail sell flow from column (3) of Table H.10 against capital gain overhang from -0.40 to 0.40 , at standardized GainTalk of -1 , 0 , and $+1$ (red dashed, grey solid, and green solid). Predicted values use CGO^- , CGO^+ , GainTalk, and the GainTalk $\times$ CGO-arm interactions, with the remaining covariates and fixed effects set to zero. Dots are means of demeaned column (3) residuals plus the fitted CGO^- and CGO^+ terms within CGO bins of width 0.04 . Shaded regions are $CGO < 0$ and $CGO \geq 0$. Double-headed arrows mark the difference between GainTalk of $+1$ and -1 at $CGO = -0.40$ and $CGO = 0.40$. $\hat{\theta}^- = -0.0279$ ($t = 9.11$); $\hat{\theta}^+ = +0.1207$ ($t = 13.56$). Column (3) includes firm and date fixed effects. $N = 1,891,685$.

I Robustness and Extensions for Market-Level GainTalk and Abnormal Market Turnover

Table I.1: **GainTalk, Daily Change in Aggregate CGO, and Abnormal Log Trading**

The table reports the specifications in Table 5, with the same-day market return replaced by the standardized daily change in aggregate CGO. Cols. (1) and (2) add Δ aggregate CGO as a control. Cols. (3)–(6) add GainTalk $\times$ Δ aggregate CGO. Δ aggregate CGO is the day-to-day change in $100 \times \sum_i \text{CGO}_{it} \text{MktCap}_{it} / \sum_i \text{MktCap}_{it}$, standardized using the mean and standard deviation over the cols. (1)–(4) estimation sample; cols. (5)–(6) use the same standardized daily series on their shorter samples. $\text{CGO}_{it} = (\tilde{P}_{it} - \text{RP}_{it}) / \tilde{P}_{it}$ (Grinblatt and Han 2005). Everything else follows Table 5. t -statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	AbnLogTO				AbnLogSell	
	CRSP	S&P 500	CRSP	S&P 500	CRSP	S&P 500
	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0316*** (5.92)	+0.0277*** (4.93)	+0.0298*** (5.60)	+0.0260*** (4.61)	+0.0424*** (7.80)	+0.0310*** (5.05)
GainTalk $\times$ Δ Aggregate CGO			+0.0179*** (5.76)	+0.0184*** (5.42)	+0.0201*** (5.65)	+0.0206*** (5.06)
Forward market sentiment	−0.0969*** (8.52)	−0.1038*** (8.69)	−0.0895*** (7.95)	−0.0963*** (8.14)	−0.0793*** (8.54)	−0.0860*** (8.06)
Log social-media volume	+0.2304*** (10.70)	+0.1998*** (8.63)	+0.2265*** (10.71)	+0.1957*** (8.63)	+0.2072*** (10.56)	+0.1457*** (6.51)
RavenPack sentiment	+0.0211*** (6.31)	+0.0234*** (6.46)	+0.0206*** (6.16)	+0.0229*** (6.30)	+0.0114*** (3.84)	+0.0144*** (4.02)
RavenPack log news count	+0.0455*** (4.98)	+0.0444*** (4.64)	+0.0461*** (5.11)	+0.0451*** (4.78)	+0.0413*** (6.43)	+0.0452*** (5.89)
Δ Aggregate CGO	−0.0148*** (3.84)	−0.0160*** (3.86)	−0.0102*** (2.80)	−0.0113*** (2.84)	−0.0015 (0.23)	−0.0019 (0.24)
Cum. return $t-5$ to $t-1$	−0.0065*** (3.98)	−0.0067*** (3.83)	−0.0064*** (3.81)	−0.0066*** (3.65)	−0.0037** (2.10)	−0.0039* (1.89)
Cum. return $t-21$ to $t-6$	−0.0014 (1.29)	−0.0023** (1.96)	−0.0015 (1.35)	−0.0023** (2.02)	−0.0004 (0.45)	−0.0010 (1.09)
Volatility ($t-5$ to $t-1$)	+0.0391*** (2.79)	+0.0373** (2.41)	+0.0375*** (2.64)	+0.0357** (2.27)	+0.0242 (1.63)	+0.0214 (1.13)
Lagged dependent variable	+0.3596*** (16.32)	+0.3870*** (17.39)	+0.3572*** (16.29)	+0.3844*** (17.35)	+0.5261*** (26.36)	+0.5285*** (24.43)
Year-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,737	2,737	2,737	2,737	2,151	2,151
Adj. R^2	0.563	0.579	0.568	0.584	0.779	0.757

Table I.2: GainTalk, Market Return, and Δ Abnormal Log Trading Activity

The table reports the specifications in Table 5, with the dependent variable replaced by its one-day change. Cols. (1)–(4) use $\Delta \text{AbnLogTO}_t = \text{AbnLogTO}_t - \text{AbnLogTO}_{t-1}$; cols. (5)–(6) use $\Delta \text{AbnLogSell}_t = \text{AbnLogSell}_t - \text{AbnLogSell}_{t-1}$. The lagged dependent variable is the lag of this change. Everything else follows Table 5. t -statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$\Delta \text{AbnLogTO}$				$\Delta \text{AbnLogSell}$	
	CRSP	S&P 500	CRSP	S&P 500	CRSP	S&P 500
	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0229*** (4.72)	+0.0214*** (4.24)	+0.0211*** (4.27)	+0.0196*** (3.81)	+0.0212*** (4.94)	+0.0137*** (3.01)
GainTalk $\times$ Market return			+0.0117*** (3.59)	+0.0114*** (3.24)	+0.0197*** (5.70)	+0.0195*** (5.00)
Forward market sentiment	−0.0715*** (7.27)	−0.0786*** (7.42)	−0.0690*** (7.04)	−0.0761*** (7.21)	−0.0712*** (8.14)	−0.0780*** (7.52)
Log social-media volume	+0.1408*** (6.78)	+0.1436*** (6.53)	+0.1393*** (6.75)	+0.1423*** (6.50)	+0.1490*** (7.20)	+0.1471*** (6.37)
RavenPack sentiment	+0.0116*** (3.61)	+0.0098*** (2.86)	+0.0113*** (3.50)	+0.0095*** (2.76)	+0.0139*** (4.52)	+0.0117*** (3.38)
RavenPack log news count	+0.0096 (1.35)	+0.0034 (0.45)	+0.0095 (1.35)	+0.0033 (0.44)	+0.0010 (0.16)	−0.0052 (0.78)
Market return	−0.0185*** (4.55)	−0.0211*** (4.84)	−0.0143*** (3.25)	−0.0170*** (3.61)	+0.0123*** (2.97)	+0.0108** (2.34)
Cum. return $t-5$ to $t-1$	+0.0016 (1.08)	+0.0021 (1.34)	+0.0019 (1.27)	+0.0023 (1.52)	+0.0020 (1.32)	+0.0028* (1.67)
Cum. return $t-21$ to $t-6$	−0.0010 (1.11)	−0.0011 (1.10)	−0.0010 (1.06)	−0.0010 (1.05)	−0.0006 (0.80)	−0.0004 (0.53)
Volatility ($t-5$ to $t-1$)	−0.0429*** (4.19)	−0.0502*** (4.65)	−0.0433*** (4.22)	−0.0506*** (4.67)	−0.0298*** (3.06)	−0.0384*** (3.61)
Lagged Δ dependent variable	−0.3901*** (22.91)	−0.3811*** (22.48)	−0.3916*** (23.15)	−0.3827*** (22.69)	−0.3089*** (15.46)	−0.3033*** (13.85)
Year-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,736	2,736	2,736	2,736	2,150	2,150
Adj. R^2	0.217	0.217	0.220	0.220	0.178	0.155

Table I.3: GainTalk, Market Turnover, and Aggregate Retail Sell Flow: Extended Controls and Year-Month FE

The table reports the specifications in Table 5, with year-quarter fixed effects replaced by year-month fixed effects. GainTalk is the standardized gain-frame share in that table. Controls additionally include StockTwits disagreement, the message-weighted mean of firm-day $\sqrt{1 - \text{AvgSentiment}^2}$ (Cookson and Niessner 2020, 2023); the Baker-Bloom-Davis news-based Economic Policy Uncertainty index; the Aruoba-Diebold-Scotti business-conditions index; the log IBES earnings-announcement count (Gu et al. 2025); and lagged standardized GainTalk ($t-1$). Market return is the daily value-weighted return in percentage points. Everything else follows Table 5. t -statistics in parentheses; Newey-West standard errors with 5 lags. Cols. (1)–(4): $N = 2,737$. Cols. (5)–(6): $N = 2,151$. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	AbnLogTO				AbnLogSell	
	CRSP	S&P 500	CRSP	S&P 500	CRSP	S&P 500
	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0283*** (5.83)	+0.0272*** (5.36)	+0.0260*** (5.27)	+0.0249*** (4.83)	+0.0337*** (7.64)	+0.0262*** (5.22)
GainTalk $\times$ Market return			+0.0122*** (4.20)	+0.0122*** (3.94)	+0.0190*** (6.41)	+0.0196*** (5.95)
Forward market sentiment	−0.0950*** (8.48)	−0.1076*** (9.01)	−0.0930*** (8.34)	−0.1056*** (8.88)	−0.0896*** (9.00)	−0.1101*** (9.65)
Log social-media volume	+0.2428*** (10.03)	+0.2138*** (8.58)	+0.2394*** (9.91)	+0.2103*** (8.48)	+0.2507*** (11.65)	+0.1810*** (7.93)
RavenPack sentiment	+0.0232*** (6.12)	+0.0239*** (5.96)	+0.0231*** (6.13)	+0.0239*** (5.97)	+0.0113*** (4.03)	+0.0148*** (4.56)
RavenPack log news count	+0.1215*** (6.92)	+0.1141*** (6.70)	+0.1209*** (6.85)	+0.1135*** (6.64)	+0.0882*** (8.20)	+0.0879*** (7.12)
Market return	−0.0131*** (3.19)	−0.0137*** (3.11)	−0.0085* (1.94)	−0.0091* (1.93)	+0.0144*** (3.71)	+0.0168*** (3.91)
StockTwits disagreement	+0.0037 (0.70)	+0.0063 (1.08)	+0.0041 (0.78)	+0.0068 (1.15)	+0.0088** (2.00)	+0.0107** (2.04)
Baker-Bloom-Davis EPU	+0.0132*** (2.74)	+0.0143*** (2.85)	+0.0131*** (2.74)	+0.0142*** (2.83)	+0.0067* (1.80)	+0.0082** (2.00)
Aruoba-Diebold-Scotti ADS	+0.0033 (0.57)	+0.0050 (0.78)	+0.0030 (0.52)	+0.0047 (0.74)	−0.0090** (2.28)	−0.0086 (1.64)
Log earnings announcements	−0.0164*** (3.17)	−0.0141*** (2.75)	−0.0160*** (3.09)	−0.0136*** (2.67)	−0.0121*** (3.25)	−0.0096** (2.32)
Cum. return $t-5$ to $t-1$	−0.0075*** (3.74)	−0.0081*** (3.81)	−0.0073*** (3.63)	−0.0079*** (3.69)	−0.0009 (0.61)	−0.0014 (0.78)
Cum. return $t-21$ to $t-6$	−0.0030** (2.21)	−0.0040*** (2.76)	−0.0029** (2.18)	−0.0040*** (2.73)	−0.0010 (0.90)	−0.0021 (1.61)
Volatility ($t-5$ to $t-1$)	+0.0089 (0.55)	+0.0023 (0.14)	+0.0087 (0.54)	+0.0022 (0.13)	−0.0159 (1.16)	−0.0257 (1.58)
Lagged dependent variable	+0.2386*** (11.28)	+0.2674*** (12.46)	+0.2370*** (11.19)	+0.2657*** (12.37)	+0.3780*** (19.09)	+0.3742*** (18.40)
Lagged GainTalk	−0.0019 (0.51)	−0.0004 (0.09)	−0.0008 (0.22)	+0.0007 (0.19)	−0.0042 (1.16)	−0.0014 (0.36)
Year-month FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,737	2,737	2,737	2,737	2,151	2,151
Adj. R^2	0.601	0.615	0.604	0.617	0.807	0.786

Table I.4: GainTalk, Next-Day Market Turnover, and Next-Day Aggregate Retail Sell Flow: Extended Controls and Year-Month FE

The table reports the specifications in Table I.3, with the dependent variable replaced by next-day abnormal log trading activity. Cols. (1)–(4) use $\text{AbnLogTO}_{t+1} = \log(\text{TO}_{t+1}) - \frac{1}{121} \sum_{s=t-140}^{t-20} \log(\text{TO}_s)$; cols. (5)–(6) apply the same window to RetailSell_{t+1} . Controls additionally include StockTwits disagreement, the message-weighted mean of firm-day $\sqrt{1 - \text{AvgSentiment}^2}$ (Cookson and Niessner 2020, 2023); the Baker-Bloom-Davis news-based Economic Policy Uncertainty index; the Aruoba-Diebold-Scotti business-conditions index; the log IBES earnings-announcement count (Gu et al. 2025); and lagged standardized GainTalk ($t-1$). Market return is the daily value-weighted return in percentage points. Everything else follows Table I.3. t -statistics in parentheses; Newey-West standard errors with 5 lags. Cols. (1)–(4): $N=2,736$. Cols. (5)–(6): $N=2,150$. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	AbnLogTO _{t+1}				AbnLogSell _{t+1}	
	CRSP	S&P 500	CRSP	S&P 500	CRSP	S&P 500
	(1)	(2)	(3)	(4)	(5)	(6)
GainTalk	+0.0094* (1.87)	+0.0083 (1.61)	+0.0080 (1.61)	+0.0069 (1.35)	+0.0149*** (3.20)	+0.0089* (1.76)
GainTalk × Market return			+0.0072* (1.85)	+0.0072* (1.78)	+0.0110*** (3.31)	+0.0101*** (2.72)
Forward market sentiment	−0.0544*** (4.62)	−0.0656*** (5.24)	−0.0532*** (4.55)	−0.0644*** (5.19)	−0.0405*** (4.05)	−0.0608*** (5.47)
Log social-media volume	+0.1382*** (5.15)	+0.1286*** (4.51)	+0.1362*** (5.10)	+0.1265*** (4.46)	+0.1563*** (6.20)	+0.1202*** (4.27)
RavenPack sentiment	+0.0286*** (6.53)	+0.0288*** (6.21)	+0.0286*** (6.53)	+0.0288*** (6.21)	+0.0150*** (3.88)	+0.0177*** (3.92)
RavenPack log news count	+0.0544*** (3.45)	+0.0397** (2.40)	+0.0540*** (3.43)	+0.0393** (2.38)	+0.0729*** (5.23)	+0.0625*** (3.93)
Market return	−0.0122*** (2.92)	−0.0119*** (2.66)	−0.0095** (2.14)	−0.0091* (1.91)	−0.0046 (1.19)	−0.0037 (0.82)
StockTwits disagreement	−0.0026 (0.35)	−0.0017 (0.24)	−0.0023 (0.31)	−0.0014 (0.20)	+0.0041 (0.80)	+0.0027 (0.45)
Baker-Bloom-Davis EPU	+0.0127** (2.32)	+0.0126** (2.19)	+0.0126** (2.31)	+0.0126** (2.18)	+0.0066 (1.54)	+0.0074 (1.49)
Aruoba-Diebold-Scotti ADS	+0.0000 (0.01)	+0.0009 (0.15)	−0.0001 (0.02)	+0.0007 (0.12)	−0.0053 (1.15)	−0.0050 (0.88)
Log earnings announcements	+0.0036 (0.76)	+0.0063 (1.25)	+0.0038 (0.81)	+0.0065 (1.31)	+0.0022 (0.47)	+0.0051 (1.00)
Cum. return $t-5$ to $t-1$	−0.0088*** (3.64)	−0.0099*** (3.81)	−0.0087*** (3.59)	−0.0098*** (3.77)	−0.0029 (1.42)	−0.0037 (1.62)
Cum. return $t-21$ to $t-6$	−0.0042*** (2.66)	−0.0059*** (3.41)	−0.0042*** (2.65)	−0.0059*** (3.41)	−0.0020 (1.33)	−0.0036** (2.07)
Volatility ($t-5$ to $t-1$)	−0.0332 (1.51)	−0.0448* (1.85)	−0.0333 (1.52)	−0.0448* (1.86)	−0.0167 (0.87)	−0.0264 (1.13)
Lagged dependent variable	+0.1166*** (5.61)	+0.1392*** (6.65)	+0.1157*** (5.55)	+0.1382*** (6.58)	+0.1831*** (7.32)	+0.1682*** (6.76)
Lagged GainTalk	+0.0073 (1.63)	+0.0098** (2.06)	+0.0080* (1.76)	+0.0104** (2.19)	+0.0015 (0.34)	+0.0025 (0.53)
Year-month FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,736	2,736	2,736	2,736	2,150	2,150
Adj. R^2	0.501	0.517	0.501	0.517	0.724	0.706

Table I.5: Market-Wide GainTalk and Individual-Stock Abnormal Log Turnover

The table reports stock-day OLS estimates of same-day abnormal log turnover on leave-one-out market GainTalk, the value-weighted market return, and their interaction. Cols. (1)–(2) report the baseline specification; col. (3) adds own-stock GainTalk and its interaction with the market return. Leave-one-out GainTalk excludes messages about stock i and is full-sample standardized; own-stock GainTalk is within-date standardized. Col. (1) includes no fixed effects; cols. (2)–(3) include firm and year-quarter fixed effects. Controls are the stock-level counterparts of those in Table 5. Unlike the main stock-day panel in Section 4, this specification does not require retail sell flow or the trailing history needed to construct abnormal log retail sell flow. t -statistics in parentheses; standard errors are double-clustered by firm and date. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)
Market GainTalk $_t^{-i}$	+0.0181*** (5.58)	+0.0080** (2.26)	+0.0088** (2.50)
Market GainTalk $_t^{-i} \times$ Market return	+0.0098*** (3.03)	+0.0106*** (3.66)	+0.0106*** (3.64)
Stock GainTalk $_{i,t}$			+0.0275*** (35.08)
Stock GainTalk $_{i,t} \times$ Market return			+0.0025*** (4.32)
Forward sentiment	+0.0444*** (9.80)	+0.0050 (1.64)	−0.0107*** (3.44)
Log social-media volume	+0.0826*** (15.19)	+0.2468*** (46.85)	+0.2477*** (47.09)
RavenPack sentiment	+0.0738*** (15.14)	+0.0712*** (21.29)	+0.0638*** (19.32)
RavenPack log news count	+0.1061*** (23.20)	+0.0465*** (17.80)	+0.0467*** (17.91)
Market return	−0.0142*** (3.85)	−0.0103*** (2.99)	−0.0103*** (3.01)
Cum. return $t-5$ to $t-1$	−0.0723*** (5.09)	−0.0187* (1.87)	−0.0336*** (3.37)
Cum. return $t-21$ to $t-6$	+0.0515*** (6.82)	+0.0353*** (5.87)	+0.0325*** (5.67)
Volatility ($t-5$ to $t-1$)	+0.5230*** (7.14)	+0.0858*** (2.62)	+0.1130*** (3.48)
Lagged AbnLogTO	+0.7042*** (144.00)	+0.6133*** (123.83)	+0.6121*** (123.91)
Firm FE	No	Yes	Yes
Year-quarter FE	No	Yes	Yes
Observations	2,187,512	2,187,397	2,187,397
Adj. R^2	0.531	0.590	0.591

Table I.6: HAC Lag Sensitivity and Block Bootstrap: GainTalk $\times$ Market Return

This table reports the GainTalk $\times$ market return coefficient from Table 5, columns (3) and (4), and from Table I.2, columns (3) and (4). t -statistics in parentheses are Newey-West standard errors under alternative lag lengths; the last row reports moving-block bootstrap p -values. Cols. (1)–(2) use same-day AbnLogTO; cols. (3)–(4) use $\Delta\text{AbnLogTO}_t = \text{AbnLogTO}_t - \text{AbnLogTO}_{t-1}$. Within each block, columns report the CRSP universe and S&P 500 constituents. All columns include year-quarter and day-of-week fixed effects. The lagged dependent variable is lagged AbnLogTO in cols. (1)–(2) and lagged $\Delta\text{AbnLogTO}$ in cols. (3)–(4). Everything else follows Table 5. Moving-block bootstrap uses 21-day blocks and 2,000 replications with controls partialled out (FWL); bootstrap 95% percentile intervals for the interaction are $[+0.0063, +0.0191]$, $[+0.0058, +0.0196]$, $[+0.0053, +0.0190]$, and $[+0.0045, +0.0186]$ for cols. (1)–(4). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$Y = \text{AbnLogTO}_t$		$Y = \Delta\text{AbnLogTO}_t$	
	CRSP (1)	S&P 500 (2)	CRSP (3)	S&P 500 (4)
Newey-West, 5 lags	+0.0120*** (4.06)	+0.0118*** (3.72)	+0.0117*** (3.59)	+0.0114*** (3.24)
Newey-West, 10 lags	+0.0120*** (3.91)	+0.0118*** (3.59)	+0.0117*** (3.59)	+0.0114*** (3.23)
Newey-West, 21 lags	+0.0120*** (3.77)	+0.0118*** (3.45)	+0.0117*** (3.60)	+0.0114*** (3.24)
Newey-West, 63 lags	+0.0120*** (3.45)	+0.0118*** (3.21)	+0.0117*** (3.37)	+0.0114*** (3.08)
Moving-block bootstrap	+0.0120*** ($p < 0.001$)	+0.0118*** ($p < 0.001$)	+0.0117*** ($p < 0.001$)	+0.0114*** ($p < 0.001$)
Observations	2,737	2,737	2,736	2,736

Table I.7: GainTalk Index and Market Turnover: Split-Sample Test by Alternating Days

The table reports a split-sample test of a GainTalk sentiment index for daily abnormal log market turnover (CRSP). Trading days in the sample are numbered in calendar order and split into odd days ($N = 1,369$) and even days ($N = 1,368$). In the fitting half, AbnLogTO is regressed on Share, Total, and $\text{Share} \times \text{Total}$, where Share is the gain-frame share of reference-dependent posts (GainTalk), Total is the log number of reference-dependent posts, and both are standardized with fitting-half moments (cols. (1) and (4)). The GainTalk index is the fitted linear combination of the three terms, standardized with fitting-half moments. The index is then computed in the other half with the same coefficients and moments, and AbnLogTO is regressed on the index (cols. (2) and (5)) and on the index and index $\times$ market return (cols. (3) and (6)). Controls are those of Table 5. Everything else follows Table 5. The full sample covers April 2013 to September 2024 ($N = 2,737$). t -statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Fit on odd days, test on even days			Fit on even days, test on odd days		
	Fit	Test		Fit	Test	
	(1)	(2)	(3)	(4)	(5)	(6)
Share	+0.0232*** (3.46)			+0.0395*** (6.54)		
Total	+0.1154*** (3.60)			+0.1476*** (4.68)		
Share $\times$ Total	+0.0103*** (2.75)			+0.0100*** (2.88)		
GainTalk index		+0.0306*** (5.38)	+0.0306*** (5.36)		+0.0222*** (3.70)	+0.0220*** (3.67)
GainTalk index $\times$ Market return			+0.0087** (2.25)			+0.0108*** (2.96)
Market return	-0.0197*** (3.71)	-0.0147*** (3.10)	-0.0122** (2.47)	-0.0154*** (3.28)	-0.0197*** (3.75)	-0.0168*** (3.07)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes
Year-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,369	1,368	1,368	1,368	1,369	1,369
Adj. R^2	0.579	0.559	0.561	0.567	0.575	0.577

Table I.8: Influencer GainTalk and Abnormal Log Market Turnover

Cols. (1)–(2) report same-day abnormal log market turnover on influencer and other-poster GainTalk and the market return for the CRSP and S&P 500 universes. Cols. (3)–(6) report after-market-open GainTalk on overnight GainTalk; see column headers for influencer versus other-poster AMO GainTalk. A message is assigned to the influencer group when the poster is above the 99th percentile by number of followers (Cookson et al. 2024). Overnight GainTalk aggregates the evening session from 16:00 ET on the prior trading day and the pre-open session before 09:30 ET; AMO GainTalk uses messages posted from 09:30 ET to 16:00 ET. All columns include year-quarter and day-of-week fixed effects. $\beta_1 - \beta_2$ is the difference of the two GainTalk coefficients in cols. (1)–(2); $\gamma_1 - \gamma_2$ is the difference of the two overnight GainTalk coefficients in each column of cols. (3)–(6). Everything else follows Table 5. *t*-statistics in parentheses; Newey-West standard errors with 5 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Same-day AbnLogTO		AMO GainTalk			
	CRSP	S&P 500	CRSP		S&P 500	
			Influencer AMO	Other- poster AMO	Influencer AMO	Other- poster AMO
	(1)	(2)	(3)	(4)	(5)	(6)
Influencer same-day GainTalk (β_1)	+0.0278*** (5.12)	+0.0246*** (4.31)				
Other-poster same-day GainTalk (β_2)	+0.0118** (2.41)	+0.0098* (1.92)				
$\beta_1 - \beta_2$	+0.0160** (2.06)	+0.0148* (1.78)				
Influencer overnight GainTalk (γ_1)			+0.1585*** (3.38)	+0.2815*** (6.18)	+0.1602*** (3.41)	+0.2798*** (6.12)
Other-poster overnight GainTalk (γ_2)			+0.0095 (0.23)	+0.1465*** (3.48)	+0.0108 (0.26)	+0.1472*** (3.49)
$\gamma_1 - \gamma_2$			+0.1490** (2.12)	+0.1350** (2.14)	+0.1494** (2.08)	+0.1326** (2.11)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,737	2,737	2,737	2,737	2,737	2,737
Adj. R^2	0.558	0.574	0.346	0.389	0.344	0.387

J Earnings News, GainTalk, and Disposition Spreads: Additional Results

This Internet Appendix reports additional results for the earnings-news and pre-announcement GainTalk sorts in Section 6. It includes residual GainTalk spreads, post-announcement alphas within surprise quintiles, value-weighted spreads, and sensitivity to the pre-announcement measurement window.

Table J.1: Residual-GainTalk Spread Alphas

The table reports the specifications in Table 9, with pre-announcement GainTalk replaced by its residual from month-by-month cross-sectional regressions on capital gain overhang, the prior-year return, the return from month $t-36$ to $t-13$, log market capitalization at $t-1$, and average turnover over the prior 12 months. The return, size, and turnover controls follow the residual-overhang regression in Frazzini (2006). Everything else follows Table 9. t -statistics in parentheses; Newey-West standard errors with $K-1$ lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Rolling period	High-GainTalk spread			Low-GainTalk spread			High – low	N months
	Bad news	Good news	L/S	Bad news	Good news	L/S		
+1	−0.679*** (2.72)	+0.565*** (3.81)	+1.244*** (4.49)	−0.266 (1.21)	−0.014 (0.08)	+0.252 (0.95)	+0.992*** (2.67)	138
+2	−0.628*** (2.94)	+0.492*** (3.30)	+1.120*** (4.59)	−0.188 (0.84)	−0.086 (0.45)	+0.102 (0.37)	+1.019*** (2.64)	138
+3	−0.564*** (2.68)	+0.462*** (3.03)	+1.026*** (4.38)	−0.186 (0.86)	−0.055 (0.28)	+0.130 (0.49)	+0.896** (2.38)	138
+6	−0.463** (2.41)	+0.451*** (2.84)	+0.914*** (4.02)	−0.019 (0.10)	−0.046 (0.27)	−0.027 (0.13)	+0.941*** (2.80)	138
+12	−0.370* (1.84)	+0.373*** (3.99)	+0.742*** (3.68)	+0.032 (0.19)	+0.019 (0.11)	−0.013 (0.08)	+0.755*** (2.63)	138

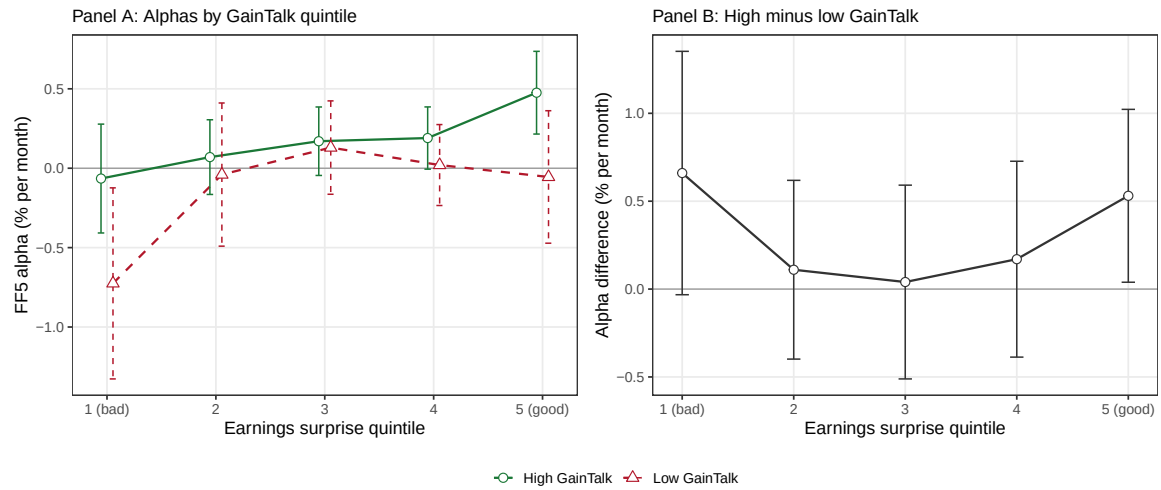


Figure J.1: Post-Announcement Alphas Within Earnings-Surprise Quintiles

The figure plots equal-weighted calendar-time Fama-French five-factor alphas of portfolios held for 1 month. Panel A plots alphas within each surprise quintile for the top and bottom GainTalk quintiles; Panel B plots the alpha of the difference between them. Whiskers are 95% confidence intervals. Everything else follows Table 9.

Table J.2: Value-Weighted High- and Low-GainTalk Spread Alphas

The table reports the specifications in Table 9, with portfolio returns value-weighted by market capitalization at the prior month-end. Everything else follows Table 9.

Rolling period	High-GainTalk spread			Low-GainTalk spread			High – low	<i>N</i> months
	Bad news	Good news	L/S	Bad news	Good news	L/S		
+1	−0.847*** (2.64)	+0.493*** (3.60)	+1.340*** (3.71)	+0.142 (0.68)	−0.158 (0.75)	−0.301 (1.02)	+1.641*** (3.44)	138
+2	−0.676** (2.34)	+0.452*** (3.51)	+1.128*** (3.71)	+0.209 (0.93)	−0.181 (0.85)	−0.390 (1.23)	+1.518*** (3.15)	138
+3	−0.625** (2.10)	+0.435*** (3.47)	+1.060*** (3.50)	+0.205 (0.93)	−0.095 (0.44)	−0.300 (0.92)	+1.359*** (2.71)	138
+6	−0.571** (2.05)	+0.405*** (2.78)	+0.976*** (3.31)	+0.242 (1.42)	−0.007 (0.04)	−0.249 (0.96)	+1.225*** (2.76)	138
+12	−0.430 (1.62)	+0.385*** (3.96)	+0.814*** (3.21)	+0.186 (1.41)	+0.049 (0.28)	−0.138 (0.77)	+0.952** (2.54)	138

Table J.3: GainTalk and Overhang Spread Alphas by Pre-Announcement Window

The table reports equal-weighted calendar-time FF5 alphas for GainTalk and overhang disposition spreads by pre-announcement window. GainTalk is measured in windows $[-15, -1]$, $[-30, -16]$, $[-45, -31]$, $[-60, -46]$; capital gain overhang is measured on stop days $-1, -15, -30, -45$. The overhang spread is long good-news/high-overhang and short bad-news/low-overhang. Rows are overlapping holding periods $K = 1, 3, 6, 12$ months. Everything else follows Table 9. t -statistics in parentheses; Newey-West standard errors with $K-1$ lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Hold	Pre-EA window	High-GT L/S	Low-GT L/S	Days rel. to EA	OH L/S
+1	$[-15, -1]$	+1.171*** (3.19)	-0.239 (0.85)	-1	+0.814*** (2.84)
+1	$[-30, -16]$	+1.143*** (3.11)	-0.228 (0.81)	-15	+0.709* (1.90)
+1	$[-45, -31]$	+1.139*** (2.98)	-0.410 (1.29)	-30	+0.889*** (3.02)
+1	$[-60, -46]$	+0.963** (2.52)	-0.414 (1.52)	-45	+0.640** (2.32)
+3	$[-15, -1]$	+0.932*** (2.80)	-0.354 (1.13)	-1	+0.586** (2.39)
+3	$[-30, -16]$	+0.897*** (2.79)	-0.237 (0.78)	-15	+0.620** (2.56)
+3	$[-45, -31]$	+0.862** (2.54)	-0.265 (0.86)	-30	+0.633*** (2.62)
+3	$[-60, -46]$	+0.758** (2.34)	-0.313 (1.03)	-45	+0.517** (2.33)
+6	$[-15, -1]$	+0.911*** (3.65)	-0.409 (1.64)	-1	+0.434* (1.92)
+6	$[-30, -16]$	+0.862*** (3.05)	-0.304 (1.25)	-15	+0.419* (1.88)
+6	$[-45, -31]$	+0.830*** (2.78)	-0.231 (1.01)	-30	+0.400* (1.90)
+6	$[-60, -46]$	+0.756** (2.51)	-0.253 (1.15)	-45	+0.392** (2.02)
+12	$[-15, -1]$	+0.633** (2.23)	-0.336 (1.56)	-1	+0.203 (0.93)
+12	$[-30, -16]$	+0.665** (2.53)	-0.273 (1.27)	-15	+0.140 (0.63)
+12	$[-45, -31]$	+0.660** (2.39)	-0.259 (1.18)	-30	+0.189 (1.02)
+12	$[-60, -46]$	+0.607** (2.20)	-0.261 (1.18)	-45	+0.232 (1.29)

K LLM Teacher-Stage Annotation Protocol

This section shows the annotation protocol used at the teacher stage of the two-stage pipeline in Section 2. The protocol is the complete coding rubric rather than a shortened instruction: the retail-investor persona, the three labeling tasks (whether a holder’s position is discussed, gain/loss frame, and reference anchor), the JSON schema, annotation principles, and worked examples. In the teacher stage the file is loaded in full, with no truncation, and placed in the system message of every labeling request. The user message contains only the post and a short context block (ticker and date). The OpenAI Batch API does not share a prompt prefix across requests, so the full protocol is repeated immediately before each post is labeled. Each request is independent: it contains only the protocol and the current post, with no prior messages or prior labels in the conversation. The teacher model therefore rereads the rubric before every annotation rather than relying on a cached or truncated summary.

Teacher labeling is submitted as sharded JSONL files to the chat-completions endpoint. Each line is a single request with a unique message identifier, temperature set to zero, and a required JSON-object response. Because the protocol is long and is duplicated on every line, shard size is kept at 500 requests so that the organization’s enqueued-token limit is not exhausted. Shards are uploaded and queued serially, with a 24-hour completion window and resume by identifier if a shard fails. This procedure yields labels on a representative sample of 100,000 messages. Those labels are then frozen with a SHA-256 hash of this protocol, so that the training artifact is bound one-to-one to the rubric below. This ensures that student-model results based on this protocol can be reproduced.

Running the same teacher calls over the full StockTwits corpus of 185.1 million messages is prohibitively costly. The pipeline therefore trains a DeBERTa-v3-base student on the teacher labels and applies that student to every post. The protocol below is the specification that generates the supervision, not the inference engine for the full corpus.

```

1 LLM Annotation Protocol
2
3 1. System Prompt (Persona)
4
5 You are a typical retail investor reading posts on StockTwits.
6 Your task is to interpret each message as a peer would, focusing on whether it conveys a sense of
7 winning or losing about a holder's position, and whether it is anchored to past outcomes or
8 future expectations.
9
10 Base your judgment on the natural, colloquial meaning the message would have for an ordinary investor
11 scrolling through social media.
12
13 This persona is designed to capture audience-perceived holder-dependent thinking, not expert-level
14 semantic correctness.
15
16 -----
17
18 2. User Prompt Template
19
20 You will be given a StockTwits message and limited context.
21
22 Input
23
24 MESSAGE:
25 "{text}"
26
27 SYMBOL: {symbol}
28 DATE (US/Eastern trading day): {date}
29
30 Task
31
32 Classify how a typical retail investor would interpret this message in terms of:
33 • Whether it is about a holder's position,
34 • Whether it conveys a gain frame, a loss frame, or neither,
35 • Whether it is anchored to past outcomes, future expectations, or neither.
36
37 -----
38
39 3. Definitions
40
41 3.1 Holder's position
42
43 Question: Is this message about a holder's position?
44 • HOLDER
45   The reader would feel the post is about a holder's position, for example
46   relative to a price such as their purchase price, average cost basis,
47   entry price, or break-even price.
48 Examples:
49   • "Up 18% from my entry"
50   • "Still above water"
51   • "Still underwater from my 35 entry... not sure I'll ever see that price again"
52
53 • NONE
54   The post is not about a holder's position.
55 Examples:
56   • "When is the SpaceX IPO?"
57   • "Revenue beat, margins expanding"
58   • "GREEN today"
59
60 -----
61
62 3.2 Gain/Loss Frame
63
64 Question: Does this message make the reader feel like the situation is framed as winning, losing, or
65 neither?
66 • GAIN
67   The reader would feel the post conveys a sense of being up, winning, or succeeding. At this stage,
68   the label captures the presence of gain-related language or framing, reflecting a perceived sense of
69   winning or improvement, not merely a neutral description of activity or attention.
70 Examples:
71   • "I'm printing"
72   • "Finally in the money"
73   • "To the moon"
74
75 • LOSS
76   The reader would feel the post conveys a sense of being down, losing, or getting hurt. At this
77   stage, the label captures the presence of loss-related language or framing, reflecting a perceived
78   sense of losing or deterioration, not merely a neutral or critical opinion.
79 Examples:
80   • "Bagholding this"
81   • "This is trash, dumping hard"
82   • "Red again"
83
84 • NONE
85   No clear sense of winning or losing.
86 Examples:
87   • "What's the next earnings date?"

```

88 • "Big volume today"

89 Spam & Promotion Rule: Any post that reads like a structured advertisement, promotes a specific

90 external service/link with promised returns (e.g., "Make 40% annualized," "Join our discord"), or

91 uses bot-like templates must be labeled **NONE** (Gain Loss Frame).

92

93 -----

94

95 3.3 Reference Anchor

96

97 Question: Is the message anchored to a past reference point or to a future outcome?

98 • **PAST**

99 The reader would feel the post refers to a current or realized price relative to a prior price level

100 (e.g., a prior close, a stated entry, having sold, bagholding, or running up/down). Including

101 dynamic changes where the current state is defined by contrast to a prior state (e.g., "giving back

102 gains", "trapped", "bagholding").

103 Examples:

104 • "Finally above water"

105 • "Still bagholding this"

106 • "Back to breakeven"

107 • "I am in the green/red"

108 • "Gave back last week's gains"

109

110 • **FUTURE**

111 The reader would feel the post refers to what will happen next, not what has already and actually

112 happened.

113 Examples:

114 • "This is about to dump"

115 • "Next stop \$50"

116 • "Going to moon"

117 • "I believe this will turn around"

118 • "Feels like this is going lower"

119

120 • **UNSPECIFIED**

121 The reader cannot clearly infer a time anchor. When both past outcomes and future expectations are

122 mentioned, prioritize **PAST** if the framing emphasizes achieved status, unless the post explicitly

123 centers on what will happen next.

124 Examples:

125 • "Nice choice"

126 • "Strong stock"

127 • "Wild move"

128

129 -----

130

131 4. Output Format (Strict JSON)

132

133 Return only valid JSON in the following format:

134

135 {

136 "object": "HOLDER|NONE",

137 "gain_loss_frame": "GAIN|LOSS|NONE",

138 "reference_anchor": "PAST|FUTURE|UNSPECIFIED",

139 "rationale": "Short explanation (1-2 sentences) describing how a typical investor would interpret

140 this message."

141 }

142

143 No extra text. No markdown. No commentary outside JSON.

144

145 -----

146

147 5. Annotation Principles

148

149 Principle 1: Reader Impression

150

151 Label based on what a typical retail investor would feel, not whether the statement is financially

152 precise.

153

154 -----

155

156 Principle 2: Gain/Loss Language is a Necessary Signal, Not a Mechanism by Itself

157

158 If the post creates a sense of being up or being down, even implicitly, capture it. Gain/loss language

159 identifies whether a message is framed in terms of winning or losing, but by itself does not imply

160 holder-dependent thinking.

161

162 -----

163

164 Principle 3: Holder-Dependent Framing is the Mechanism Signal

165

166 The strongest evidence of holder-dependent thinking occurs when a post is about a holder's position

167 (object = **HOLDER**) and frames that position relative to a past anchor (reference_anchor = **PAST**).

168

169 This combination captures a holder's position relative to a reference point. Posts that

170 mention only gains/losses, without a past anchor, should be interpreted as general sentiment rather

171 than full holder-dependent thinking.

172

173 -----

174

175 Principle 4: Ambiguity is a Valid Label

```

176
177 If a normal reader could reasonably interpret the post in more than one way, use:
178 • NONE for object if no price level is clearly evaluated
179 • UNSPECIFIED for reference_anchor
180 • NONE for gain_loss_frame if needed
181
182 -----
183
184 Principle 5: Minimal Use of Context
185
186 You may use the provided symbol and date only to help interpret tone or meaning,
187 not to mechanically infer gains or losses.
188
189 -----
190
191 6. Worked Examples
192
193 Example 1
194
195 Message: "I finally sold my shares, locked in a 15% gain"
196
197 {
198   "object": "HOLDER",
199   "gain_loss_frame": "GAIN",
200   "reference_anchor": "PAST",
201   "rationale": "The message evaluates a completed price outcome relative to a past price level."
202 }
203
204 -----
205
206 Example 2
207
208 Message: "This thing is about to dump, watch out"
209
210 {
211   "object": "NONE",
212   "gain_loss_frame": "LOSS",
213   "reference_anchor": "FUTURE",
214   "rationale": "A reader would interpret this as a forecast that the price
215     will fall, not as a statement about a holder's position. The loss frame
216     refers to a future outcome rather than a past loss."
217 }
218
219 -----
220
221 Example 3
222
223 Message: "Still underwater from my 35 entry... not sure I'll ever see that price again"
224
225 {
226   "object": "HOLDER",
227   "gain_loss_frame": "LOSS",
228   "reference_anchor": "PAST",
229   "rationale": "The poster is describing their own position, entered at 35, and the
230     doubt of seeing that price again frames it as a past-anchored loss."
231 }
232
233 -----
234
235 7. One-Line Philosophy (Methods-Ready)
236
237 This annotation protocol approximates how retail investors perceive whether a post is about a
238 holder's position, whether it is gain- or loss-framed, and whether that framing is anchored to a
239 past price or to a future expectation.
240
241 -----
242
243 End of Protocol

```