Lagged Retail Sentiment Rankings and Weekly Stock Return Predictability:

Panel Evidence from China's Equity Markets and EastMoney Guba

by

Haozhi Zheng

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2025

Professor Marti G. Subrahmanyam Professor Christina Wang Professor Wendy Jin	Professor Chen Zhao
Faculty Advisers	Thesis Adviser

Abstract

This paper empirically investigates how retail investor sentiment predicts weekly returns on constituent stocks of the FTSE China A50 index. It finds a significant inverse relationship between the cross-sectional ranking of a stock's LLM-derived sentiment score and its subsequent return — All else equal, the more highly a stock is regarded by retail investors relative to other stocks, the lower its return will be. In addition, the negative predictive effect is stronger during market downturns, but loses statistical significance otherwise — The effect is asymmetrical. These findings strongly aligns with a series of conclusions from current literature (Baker and Wurgler 149; Baker et al. 281-283; Cui et al. 14-15; Bashir et al. 2-4), but features a novel non-market proxy variable for investor sentiment and large-cap stocks listed in Mainland China exchanges. It highlights the potential for LLM-enabled investment strategies and underscores the importance of investor psychology in managing market risk and renewing regulation as retail investors gain access to more asset classes. Future research can build on these findings to identify specific mediators and examine similar phenomena in other retail-heavy markets.

1. Introduction

On Nov 6, 2024, as Donald Trump marches toward 270, Wisesoft Co., Ltd.'s (川大智胜) share price rallied. The Sichuanese air traffic control software company, with little ties to the US market or its politics, saw its shares surge by 10%, hitting the Shenzhen Stock Exchange's daily limit-up once again after a prolonged period of speculative growth (Yang and Cha). The surge is largely attributable to the company's name. To mandarin speakers, the company's Chinese name would vaguely mean "Trump wins big with strategic intellect." With retail investors pouring in, the company experienced a market craze, while its income statements show a loss of ¥172

million in LTM Q3 2024 and more in the same order of magnitude over the past several years (S&P CapitalIQ). As a Wisesoft investor on EastMoney Guba ("EastMoney Stock Bar") put it on Nov 6, "we are investing in rubbish anyway, so why not invest in something fun? This stock is a pretty good option" (Kan Na Xiao Qiao Liu Shui).

For China's stock markets, this is not an isolated example. Wisesoft itself saw similar surges in price and volume after Trump's first victory in 2016 and the assassination attempt on him in July 2024. Among other characteristics, retail investors also look for fortune-bringing cultural symbols in company names, such as zodiac animals like bulls, dragons, or horses. Companies featuring these animals in their names often rally over the corresponding Spring Festival, forming what's dubbed the "zodiac-themed sector."

Despite many such episodes of events, traditional finance theory, building on top of the Efficient Market Hypothesis (Fama), leaves little room for retail investors and disregards such sentiment-driven trading behavior as noise. The recurring phenomenon of retail investors decoupling prices significantly away from fundamentals directly challenges the core assumptions of the EMH. As it demonstrated, in China's equity markets, prices could be easily swayed by cultural symbolism, coincidental puns, and superstitious narratives in addition to financial statements or policy directives. EMH, which in its semi-strong form posits that prices should have rationally reflected all obviously publicly available information (Fama 414), struggles to explain why companies like Wisesoft repeatedly see upward surges despite mediocre-at-best financial metrics and little real connection to the "concept" they are speculated on. While supporters of the EMH might attribute such anomalies to unresolved risk factors or dismiss them as outliers, the scale and recurrence of these events weaken EMH's explanatory power. This underscores a tension that finance theory must evolve to integrate psychological and social

media factors of market behavior, rather than treating them as occasional exceptions to the rational market paradigm.

In China, retail investors account for more than 60% of total stock turnover (Zhen and Zhou). Such recurring episodes of events epitomize a pervasive yet currently understudied phenomenon in behavioral finance: how retail investor sentiment can drive equity returns, or in some cases, completely decouple asset prices from economic reality. As this paper probes into this question in the context of China's equity markets, its findings will be of significance in improving risk management mechanisms and investor protection regulations alike.

2. Literature Review

Investor sentiment has attracted lots of attention from finance researchers. Earlier studies in the 80s and 90s mostly used sentiment as a convenient synonym for irrational trading behavior or non-fundamental idiosyncratic risk, leaving its implications implicit. Nevertheless, many findings remain relevant and fundamental. Black first defined noise in contrast to information (529) and theorized the enabling role of noise in financial markets in terms of providing liquidity and making information trading profitable (532). De Long et al. argued that the unpredictability of noise traders' beliefs deter rational arbitrageurs from aggressively betting against them, contributing prolonged mispricing even in the absence of fundamental risk (735) . Similarly, Shleifer and Vishny argued that noise trader risk and fundamental idiosyncratic risk alike deters arbitrageurs since they are few financially constrained institutions specializing in certain assets, highlighting the implausible assumption of many well-diversified arbitrageurs in the efficient market approach (52). In such theoretical frameworks, there always exists a dichotomy with unsophisticated, sentiment-driven noise traders on the one side and sophisticated, rational

arbitrageurs on the other. Mispricing arises from the unpredictability of sentiment-driven investors' trading behavior and the various constraints and risks facing the arbitrageurs preventing them from fully correcting the price to the asset's fundamental value.

Empirical investigations into investor sentiment levels followed in the 2000s. Notably, Baker and Wurgler constructed a sentiment index from six sentiment proxies, identifying a negative correlation between the sentiment index in the preceding month and average monthly return on a market portfolio (149). They also offered a comprehensive review of proxies for investor sentiment employed in previous empirical studies, including investor surveys, retail investor trades, mutual fund order flows, volume, dividend premium, closed-end fund discount, even climate and weather, etc. (135-138). These studies arrived at mixed conclusions. For example, Kamstra et al. found significant seasonal affective disorder (SAD) effects in multiple financial markets, with greater effects seen in higher latitude markets (340). Hirshleife and Shumway found a significant positive correlation between sunny weather and return on market indices across 26 exchanges worldwide (1028). On the other hand, Frazzini and Lamont discovered individual investors consistently tend to reallocate their money to mutual funds that do poorly in the following years (319). In this period, empirical studies on investor sentiment emerged with a wide array of proxy variables and mixed findings. Retail investor sentiment also started to gain traction as some studies distinguished their contribution from the general sentiment.

More recent empirical studies have seen more investigations into global markets outside the U.S. and an increasing focus on retail investor sentiment. Baker et al. again constructed sentiment indices using four metrics and found global and local sentiment to be contrarian predictors of cross-sectional and time-series returns in respective markets (281-283). Cui et al. found strong retail investors' bullish sentiment contributes to lower stock returns by affecting analyst attention

and liquidity, building a sentiment index using order flows of CSI 300 stocks (14-15). Bashir et al. identified a significant positive correlation between investor sentiment and stock price crash risk in Chinese exchanges using a sentiment index built from five market metrics (2-4).

A significant research gap can be identified among the current body of literature. Despite the use of various creative proxies — ranging from volume, order flows, and dividend premium to measures such as weather and climate variables — their validity as pure indicators of "sentiment" remains debatable. For those proxies constructed from market data, the true effect of sentiment can hardly be adequately captured: first, metrics such as volume and order flows may actively influence market prices in addition to reflecting underlying sentiment, making it unclear whether their contribution to the market is a result of sentiment shift or fund flows. Second, their use implicitly equates sentiment to trading decisions, where sentiment is only observed conditional on monetary action, preventing a more comprehensive review on this intrinsic psychological activity. For more exogenous proxies such as weather or sunlight exposure, while arguably having enough psychological influence, they may also pose fundamental risk by affecting business operations. They also do not conveniently limit their psychological influence to investors.

This paper empirically investigates how retail investors sentiment can impact equity returns in China's equity markets. It bridges the gap in current literature in three ways. First, its LLM-derived scores and ranking, as a proxy for retail investor sentiment, are enabled only by recent technology and more directly represent sentiment by analyzing investors' speech rather than inferring from market data. Second, while most prior studies are conducted on US markets, this paper contributes to the understanding of the more retail-heavy and yet-to-mature Chinese A-shares market. Third, this paper sources data from EastMoney Guba (EastMoney Stock Bar),

which is a dedicated stock discussion forum and segregates discussions by stock, allowing full and pure retrieval of discussion posts on individual stocks and novel research into the impacts of dedicated investment discussion platforms and online communities. Moreover, its findings will have policy implications as China's financial markets mature. They underscore the importance of enhancing investor protection frameworks, such as curbing misinformation on forums and improving financial literacy to mitigate speculative risks.

The rest of the paper will go on to introduce the data collection and manipulation methods employed to collect, quantify, and aggregate retail investors' sentiment. It will then demonstrate an econometric model that predicts real-world market movements with quantified sentiment. Finally, this paper will examine and interpret the findings before concluding and pointing out directions for future work.

3. Hypothesis & Methodology

3.1 Hypothesis

Following previous research and general knowledge on China's stock market, this paper hypothesizes the following:

H_0: High retail investor sentiment cross-section ranking predicts low stock returns.

3.2 Data

This paper investigates FTSE China A50 constituent stocks (FTSE Russell). It derives retail investor sentiment from discussion posts on Eastmoney Guba (Eastmoney Stock Bar, or "Guba" in short). Guba is one of the most popular stock discussion forums in China, and conveniently

organizes threads in subforums, each corresponding to a specific ticker. Around 1.49 million posts were retrieved via a crawler from the 50 subforums, with dates ranging from January 2024 to March 2025. Anti-scraping measures were carefully checked for and bypassed.

The posts were then fed to a locally deployed DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al.; DeepSeek-AI) reasoner LLM for sentiment evaluation. The model was asked to assign a sentiment score to each post. The score was on a scale of 0 (very negative) to 5 (very positive). The model's parameters were tuned as suggested by deepseek (DeepSeek-AI), and no output token limit was placed. User prompt and detailed configurations are available in Appendix 1.

A Python script then checks all output strings for validity of sentiment scores. After removing empty and problematic, the dataset was able to retain 1.4 million posts with full content and metadata, as well as sentiment ratings that fit all requirements. The removed posts' positions were also double checked to ensure they do not concentrate on a single or a few tickers.

A full list of features in the post dataset with sentiment scores is available in Appendix 2.

3.3 Data Manipulation and Feature Engineering

This paper primarily concerns weekly returns. Therefore, it aggregates sentiment & market data in that fashion. That is, each "observation" represents a particular stock i in week t, forming a balanced panel. For each observation, all relevant posts are grouped to calculate an average sentiment score. Then, observations are temporarily grouped by week to produce cross-sectional and sentiment ranking for each observation.

Additionally, with complete metadata on each post, a wide array of features were engineered for each stock/week observation. A full list of features in the final dataset is available in Appendix 3.

3.4 Model

With panel data, this paper employs a fixed effects model. It uses two-way clustered standard errors for robust inference and controls for time- and entity-fixed effects. The model is as follows:

$$SR_{i,t} = RSR_{i,t-1} + \Sigma EngagementMetrics_{i,t-1} + \Sigma RiskFactors_{i,t} + EntityEffects_i + TimeEffects_t + \epsilon$$

where

$SR_{i,t}$: the dependent variable, stock i's simple return in week t, "stock return"

$RSR_{i,t-1}$: the variable of interest, stock i's sentiment score ranking percentile among all stocks last week, 1 being having the strongest sentiment, "relative sentiment ranking"

$EntityEffects_i$: entity fixed effects for stock i

$TimeEffects_t$: time fixed effects for week t

$$\Sigma EngagementMetrics_{i,t-1} = BIGV_{i,t-1} + CLK_{i,t-1} + LIKE_{i,t-1} + FWD_{i,t-1} + CMT_{i,t-1}$$

- $BIGV_{i,t-1}$: percentage of posts authored by self-media as opposed to ordinary users for stock i in week t-1, "big V percentage"
- $CLK_{i,t-1}$: total clicks over post count for stock i in week t-1, in thousands, "average clicks per post"
- $LIKE_{i,t-1}$: total likes over total clicks for stock i in week t-1, "like rate"
- $FWD_{i,t-1}$: total forwards over total clicks for stock i in week t-1, "forward rate"
- $CMT_{i,t-1}$: total comments over total clicks for stock i in week t-1, "comment rate"

$$\Sigma RiskFactors_{i,t} = \beta_i \cdot IR_t + VOL_{i,t} + VLM_{i,t}$$

- β_i: beta of stock i calculated by regressing stock i's return on returns of the Shanghai Composite Index using weekly market data over the past 5 years, "5-year weekly beta"
- IR_t: return on the Shanghai Composite Index in week t, "index return"
- $VOL_{i,t}$: volatility of stock i in week t, "volatility"
- $VLM_{i,t}$: average daily volume of stock i in week t, in millions, "average daily volume"

In summary, the model for this paper controls for a) last week's social media engagement metrics b) this week's systematic and idiosyncratic risk factors. This approach peels away the effects of a) retail investor attention and engagement b) proper reward for bearing risk. The model also considers time and entity fixed effects to account for economic fluctuations and firm characteristics. In doing so, it distills the pure effect of content sentiment scores.

4. Analysis

4.1 Results

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
$RSR_{i,t-1}$	0.0086	0.0031	2.7289	0.0064	0.0024	0.0148
$BIGV_{i,t-1}$	0.0021	0.0099	0.2084	0.8349	-0.0173	0.0214
$CLK_{i,t-1}$	-0.0743	0.0223	-3.3284	0.0009	-0.1181	-0.0305
$LIKE_{i,t-1}$	0.4946	0.7012	0.7054	0.4806	-0.8803	1.8695
$FWD_{i,t-1}$	0.8219	0.3858	2.1302	0.0332	0.0653	1.5784
$CMT_{i,t-1}$	-0.3224	0.1909	-1.6888	0.0914	-0.6967	0.0519
$\beta_i \cdot IR_t$	0.8970	0.2023	4.4337	0.0000	0.5003	1.2937
$VOL_{i,t}$	0.1464	0.0632	2.3146	0.0207	0.0224	0.2704
$VLM_{i,t}$	0.0080	0.0013	6.2751	0.0000	0.0055	0.0105

Fig. 1. Results from all observations

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
$RSR_{i,t-1}$	0.0031	0.0036	0.8732	0.3827	-0.0039	0.0101
$BIGV_{i,t-1}$	-0.0211	0.0147	-1.4368	0.1510	-0.0499	0.0077
$CLK_{i,t-1}$	0.0031	0.0863	0.0358	0.9714	-0.1662	0.1724
$LIKE_{i,t-1}$	0.0122	1.1752	0.0104	0.9917	-2.2931	2.3175
$FWD_{i,t-1}$	1.3136	0.5031	2.6111	0.0091	0.3267	2.3006
$CMT_{i,t-1}$	0.4178	0.5886	0.7098	0.4780	-0.7369	1.5725
$\beta_i \cdot IR_t$	1.0166	0.2380	4.2712	0.0000	0.5497	1.4835
$VOL_{i,t}$	0.1039	0.0278	3.7418	0.0002	0.0494	0.1583
$VLM_{i,t}$	0.0071	0.0011	6.3970	0.0000	0.0049	0.0093

Fig. 2. Results from positive IR_{t-1} observations

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
$RSR_{i,t-1}$	0.0128	0.0045	2.8665	0.0042	0.0041	0.0216
$BIGV_{i,t-1}$	0.0220	0.0155	1.4161	0.1570	-0.0085	0.0524
$CLK_{i,t-1}$	-0.1071	0.0191	-5.5996	0.0000	-0.1447	-0.0696
$LIKE_{i,t-1}$	0.2639	0.7990	0.3303	0.7412	-1.3035	1.8313
$FWD_{i,t-1}$	0.4564	0.5855	0.7796	0.4357	-0.6921	1.6050
$CMT_{i,t-1}$	-0.3507	0.0724	-4.8428	0.0000	-0.4927	-0.2086
$\beta_i \cdot IR_t$	0.7515	0.2792	2.6920	0.0072	0.2039	1.2991
$VOL_{i,t}$	0.2840	0.1947	1.4588	0.1448	-0.0979	0.6658
$VLM_{i,t}$	0.0091	0.0019	4.6898	0.0000	0.0053	0.0130

Fig. 3. Results from negative IR_{t-1} observations

Details on regression results are available in Appendices 4, 5, and 6.

4.2 Interpretation

Controlling for social media engagement metrics and asset pricing risk factors, for every percentile a stock's retail sentiment is ranked higher among others in the prior week (decrease of 1 in $RSR_{i,t-1}$), its return this week is expected to decrease by 0.0086 percentage points (coefficient = 0.0086, p = 0.0064) (see figure 1), rejecting H_0. The effect becomes stronger

(coefficient = 0.0128, p = 0.0042) when the regression is performed only on observations with negative lagged index returns (IR_{t-1}) (see figure 3), but loses significance (coefficient = 0.0031, p = 0.3827) when the regression is performed only on observations with positive lagged index returns (IR_{t-1}) (see figure 2).

5. Conclusion & Discussion

Drawing from market data and Guba retail investors's discussion on constituent stocks of the FTSE China A50 index, this study finds that retail investor sentiment, measured through the cross-sectional ranking of a stock's LLM-derived sentiment score, exhibits a robust inverse relationship with the stock's future returns — all else equal, the more highly a stock is regarded by retail investors relative to other stocks in the prior week, the lower its return. In addition, the negative predictive effect is stronger when retail investor sentiment coincides with a market downturn, but loses significance otherwise — high retail sentiment may systematically prolong mispricing in a market downturn.

The baseline observation aligns strongly with conclusions posited by Black, Baker and Wurgler, Baker et al., and Cui et al. (Black 532, Baker and Wurgler 149; Baker et al. 281-283; Cui et al. 14-15). It also aligns with Frazzini and Lamont and Bashir et al. (Frazzini and Lamont 319; Bashir et al. 2-4) to a lesser extent. Despite different methods, all studies, this study included, highlights the contrarian nature of retail investor sentiment.

While this study primarily explores short-term predictive effects, given its rather granular timeframe setting, it may also shed light on the real contemporaneous effect of retail investor sentiment on stock price movement. In fact, by employing lagged non-market data, it rules out the cyclic causality problem that may exist in many prior proxy variables, such as trading volume

or scores calculated from market data (e.g. buy vs. sell orders). In this respect, this study aligns most closely with Baker and Wurgler's framework on investor sentiment's cross-sectional influence on stocks with different risk profiles and ease of valuation. Baker and Wurgler posit "the sentiment seesaw" (see figure 4), suggesting that investor sentiment should adversely affect "safe, easy-to-arbitrage, bond-like stocks" (132-133, 148). This study validates the argument in the context of China's A-shares market through identifying an inverse relationship between retail investor sentiment and Chinese large-cap stock returns. In other words, the left part of the sentiment seesaw is likely true in the Chinese stock market.

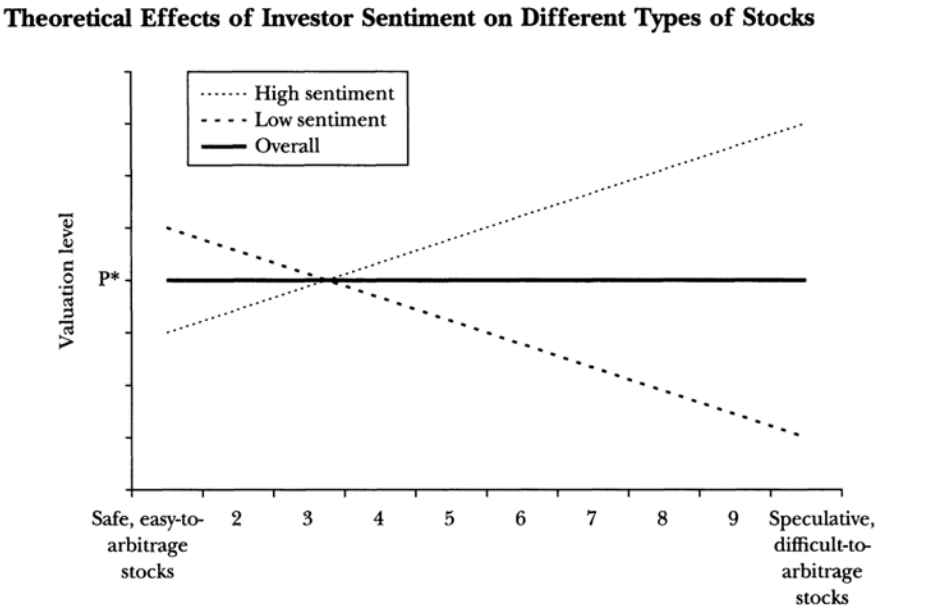
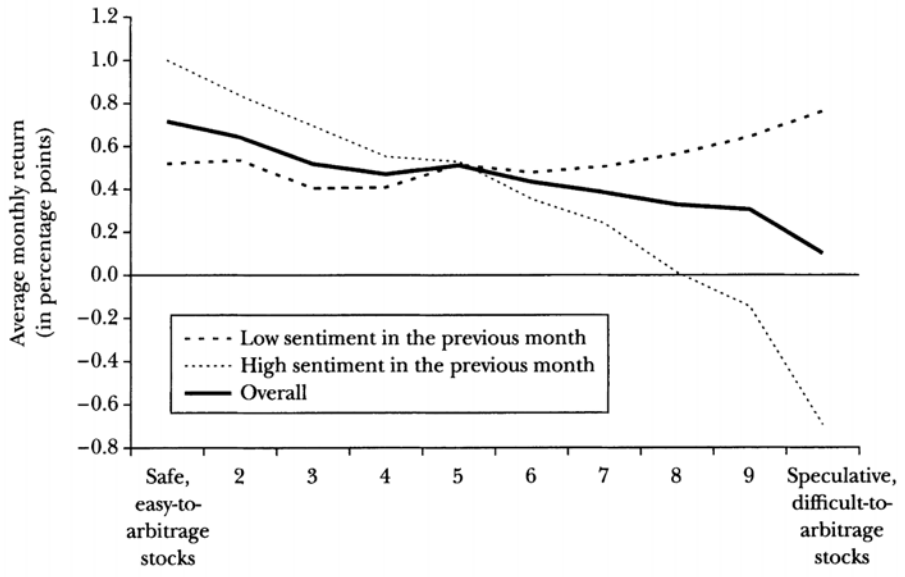


Fig. 4. Baker and Wurgler's "Sentiment Seesaw." Malcolm Baker, and Jeffrey Wurgler. "Investor Sentiment in the Stock Market." *The Journal of Economic Perspectives*, vol. 21, no. 2, 2007, pp. 129–51.

To briefly extrapolate, short-term price movements due to sentiment also create opportunities for gains in the longer term, where fundamentals dominate. As shown by Baker and Wurgler, for

"safe, easy-to-arbitrage, bond-like" stocks, high sentiment predicts high returns in the subsequent month (see figure 5) (147-148). The short-term contrarian nature of retail investor sentiment identified in this study may in part explain this phenomenon, as high sentiment is expected to push prices down in the subsequent week, causing undervaluation before being corrected over a longer time period.

Sentiment and Future Returns



Note: This figure shows average monthly returns over the value-weighted market index for volatility sorted portfolios following a month with a positive sentiment level (short dash) and with a negative sentiment level (long dash), as well as the overall average market-adjusted return (solid line). The latter are positive because of the small firm effect over this period. The volatility portfolios are equal weighted while the market index is value weighted.

Fig. 5. "Predictive Effect of Investor Sentiment in Real Markets. Malcolm Baker, and Jeffrey Wurgler. "Investor Sentiment in the Stock Market." *The Journal of Economic Perspectives*, vol. 21, no. 2, 2007, pp. 129–51.

In summary, with regard to the baseline finding, this study confirms a short-term contrarian role of retail investor sentiment in China's A-shares market, which largely agrees with previous studies' conclusions on, more broadly, investor sentiment. Meanwhile, this study features a

lagged non-market proxy variable and a very granular timeframe rarely seen in previous studies, which may enable a clean contemporaneous interpretation of its results, shedding light on the immediate causal effect of retail investor sentiment.

More interestingly, this study finds that the negative effect is stronger when relative sentiment coincides with a downward market, but insignificant when coinciding with an upward market. Following the sentimental investors vs. rational arbitrageurs dichotomy, this study poses interesting questions on the assumptions about retail investors' behavior. As often seen in such frameworks, retail investors, as "noise traders," indiscriminately inject random noise into asset prices for rational arbitrageurs to continuously correct and profit from. This empirical study, however, suggests that retail investor sentiment is asymmetric and context-dependent. On the theoretical front, this study suggests that retail investor sentiment is more effectively mitigated by arbitrage when market goes up, but retail optimism may lengthen mispricing in a market downturn.

In an upward market environment, statistical analysis indicates that retail investor sentiment, despite its prevalence and potential noise generation, does not significantly predict subsequent returns. One plausible interpretation of this finding is that during bullish periods, sentimental traders as a whole do not perceive any negative consequences from trading on non-information, while rational arbitrageurs find it easier and less risky to correct sentiment-induced mispricing swiftly. Rising markets typically feature higher liquidity and more robust investor confidence, conditions that reduce the costs and risks associated with arbitrage activities. Consequently, sentiment-driven mispricing might be short-lived or negligible in such favorable conditions.

However, the situation differs notably during downturns. In bearish periods, the analysis reveals a statistically significant and even stronger inverse relationship between retail investor sentiment and subsequent returns. In this context, stocks favored by retail sentiment continue to underperform, suggesting persistent mispricing. This result directly supports De Long et al.'s model of noise trader risk, which posits that arbitrageurs, deterred by potentially worsening sentiment-driven mispricing, may not aggressively correct prices during turbulent market periods (735). Their hesitation can extend the duration of sentiment-driven pricing errors, reinforcing the market inefficiency in downturn conditions.

The finding that retail investor sentiment operates asymmetrically according to market conditions provides valuable insights into the behavioral finance literature. Traditional finance frameworks typically characterize retail investors as uniformly prone to irrational exuberance or undue pessimism, independent of broader market contexts. This study exposes a more dynamic behavioral pattern, suggesting retail investors exhibit greater persistent loss-incurring biases during adverse market conditions. This finding has significant implications for theories on cognitive biases, particularly those related to loss aversion or disposition effects. During market downturns, retail investors may disproportionately anchor on optimistic narratives or selective attention biases, hoping desperately for mean reversion, thereby contributing to persistent mispricing.

From a practical standpoint, these findings imply meaningful implications for portfolio management and trading strategies. Investors and fund managers could potentially enhance returns by implementing contrarian strategies specifically during downturn periods, capitalizing on persistent mispricing driven by retail sentiment. Conversely, strategies based on retail sentiment in bullish periods might require additional signals or confirmation due to the reduced

predictive power of sentiment alone. This asymmetric characteristic demands adaptive strategies that consider prevailing market conditions when utilizing sentiment measures. More generally, when executing regular trading strategies, if more than one equity fulfills requirements for beta, volatility, liquidity (volume), and other fundamental considerations (region, sector, finances, etc.), investors can additionally take into account retail investor sentiment to enhance returns. This practice also reduces the investor's exposure to unforeseeable volatility caused by non-fundamental factors, such as those seen in Wisesoft and Gamestop.

Policy implications also arise from these insights. Regulators aiming to enhance market stability might reconsider policies related to retail short-selling and disclosure, particularly during downturns. If arbitrage is constrained precisely when it is most needed, sentiment-driven mispricing may unnecessarily last even longer, exacerbating market volatility. A balanced regulatory framework that facilitates prudent arbitrage while protecting against market manipulation may mitigate these inefficiencies.

The methodological approach adopted here, leveraging LLM-analyzed retail discussion posts, represents a novel application of emerging technologies to behavioral finance research. By systematically quantifying sentiment from large-scale, unstructured text data, this study demonstrates the efficacy and reliability of advanced NLP tools in capturing investor sentiment. Future research could explore extending this analytical framework to other markets or sentiment sources, such as news content or institutional analyst reports, to test the robustness and generalizability of these findings across diverse markets, settings and investor types.

In conclusion, this study empirically reaffirms retail investor sentiment — and investor sentiment more generally — as a contrarian predictor of equity returns. Importantly, it extends existing

theories by highlighting the asymmetric and context-dependent nature of sentiment effects. These insights not only refine theoretical understandings of sentiment-driven market phenomena but also offer practical guidance for market participants and policymakers navigating sentiment-induced market inefficiencies, especially during periods of market stress.

6. Limitations & Future Work

This study examined little about other markets or timeframes, nor did it nearly go through the full extent of collected data. Future research could adopt similar methods in exploring investor sentiment for other markets, and investigate its impact over longer investment horizons. In doing so, they may also incorporate more sophisticated risk factors to further single out sentiment's effect, define novel conditions on which to split the dataset into different market scenarios, or empirically investigate possible ways in which arbitrageurs are deterred from aggressively correcting mispricing from retail sentiment. This study is also limited in its capacity to efficiently collect discussion posts, restraining its scope to about 60 weeks' worth of data. Future research could find alternative web scraping sources to construct larger datasets and search for long-lived and systematic patterns.

This study also explored the use of LLMs in generating proxy variables, which leaves much to be desired. First, more accurately carving out sentiment would require significantly more computing power. Future research could deploy larger LLMs to better understand slangs and metaphors, as well as modifying inputs to provide a more explicit business/finance context. Second, the "black box" nature of LLMs remains concerning. While ranking sentiment scores would remove consistent model biases, and aggregation and averaging can mitigate outlier effects, LLMs' proficiency in assessing human sentiment remains fundamentally unknown.

Future research could use RLHF to fine-tune a potent pre-trained LLM for more accurate and human-like sentiment analysis results, aligning the proxy variable more closely to sentiment.

7. Acknowledgements

I'd first like to thank my advisor, Professor Zhao Chen from NYUSH's CSDSE Department for his guidance on choice of language models. This project has worked through bag-of-words models, fine-tuned FinBERT, OpenAI's API, before eventually settling on locally deploying DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI). Prof. Zhao also inspired me in creating relative sentiment measures to account for model biases and systematic sentiment trends. I'd also like to thank Professor Geoffery Zheng, Professor Christina Dan Wang, and Professor Ye Jin from NYUSH's Finance Department for inspiring talks we've had and keeping me up to progress throughout the year. I also sincerely appreciate my friends and family for supporting me through this somehow tortuous and rather expensive process. Much thanks! I couldn't have done it without you.

Appendix 1

Prompt and model configuration for sentiment analysis

```
async def analyze_sentiment(session, url, ticker, content):
    if content is None or content == '':
        return url, np.nan
    prompt = f'''你是一个专业的中文金融情感分析助手，你正在逛股票代码{ticker}的论
    坛。 请你在了解了这家公司的业务范围以及大致股价范围的前提下，根据以下的帖子内容判断
    其对于该股票下一周内走势情绪倾向。请注意中国散户口语化,情绪化的表达方式以及一些网络
    用词。

    你必须以JSON格式返回一个字典，包含两个字段：
    1. "sentiment_score"：整数，表示情绪倾向，范围为 0（非常负面）到 5（非常正面）；
    2. "sentiment_cat"：字符串，表示情绪类别,只能为 "positive"，"neutral"，或
    "negative"。

    帖子的内容如下：
    {content}
    '''
    payload = {
        "model": "deepseek-ai/DeepSeek-R1-Distill-Qwen-7B",
        "messages": [{"role": "user", "content": prompt}],
        "temperature": 0.6,
        "max_tokens": None,
        'response_format':{'type': 'json_object'}
    }
    headers = {
        "Content-Type": "application/json",
        # "Authorization": f"Bearer {openai.api_key}"
    }
```

async def analyze_sentiment()

Appendix 2

List of features in the raw post dataset mined from Guba

```
Index(['ID', 'url', 'ticker', 'page', 'clicks', 'likes', 'forwards',
       'comments', 'post_type', 'post_datetime', 'author', 'title', 'content',
       'url_type', 'pic', 'vid', 'sentiment_score', 'sentiment_cat'],
      dtype='object')
```

print(df.columns)

Appendix 3

List of features in final panel dataset after aggregation & feature engineering

```
Index(['count', 'zmt', 'clicks', 'likes', 'fwds', 'cmts', 'pics', 'vids',
       'sent_scr', 'sent_pos_pct', 'sent_neu_pct', 'sent_neg_pct',
       'stock_rank_in_week', 'week_rank_in_stock', 'open_price', 'close_price',
       'high_price', 'low_price', 'avg_daily_volume', 'volatility',
       'index_open', 'index_close', '5Y_weekly_beta', 'stock_return',
       'index_return', 'stock_return_lag1', 'stock_expected_return',
       'bigV_pct', 'avg_clicks_1000', 'like_pct', 'fwd_pct', 'cmt_pct',
       'bigV_pct_lag1', 'avg_clicks_1000_lag1', 'like_pct_lag1',
       'fwd_pct_lag1', 'cmt_pct_lag1', 'stock_rank_in_week_pctl',
       'stock_rank_in_week_pctl_lag1', 'week_rank_in_stock_pctl',
       'week_rank_in_stock_pctl_lag1', 'pos_over_neg', 'pos_over_neg_lag1',
       'avg_daily_volume_mil'],
      dtype='object')
```

print(panel_df.columns)

Appendix 4

Detailed regression results: all observations

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:           stock_return   R-squared:                        0.1553
Estimator:                   PanelOLS   R-squared (Between):             -6.6284
No. Observations:                2898   R-squared (Within):               0.4524
Date:                Mon, May 12 2025   R-squared (Overall):              0.3678
Time:                        16:31:09   Log-likelihood                   -7529.5
Cov. Estimator:            Clustered
                                        F-statistic:                      56.830
Entities:                          50   P-value                           0.0000
Avg Obs:                       57.960   Distribution:                 F(9,2782)
Min Obs:                       56.000
Max Obs:                       58.000   F-statistic (robust):            -11.007
                                        P-value                           1.0000
Time periods:                      58   Distribution:                 F(9,2782)
Avg Obs:                       49.966
Min Obs:                       49.000
Max Obs:                       50.000

                                Parameter Estimates
================================================================================================
                              Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------------------------
stock_rank_in_week_pctl_lag1     0.0086     0.0031     2.7289     0.0064      0.0024      0.0148
bigV_pct_lag1                    0.0021     0.0099     0.2084     0.8349     -0.0173      0.0214
avg_clicks_1000_lag1            -0.0743     0.0223    -3.3284     0.0009     -0.1181     -0.0305
like_pct_lag1                    0.4946     0.7012     0.7054     0.4806     -0.8803      1.8695
fwd_pct_lag1                     0.8219     0.3858     2.1302     0.0332      0.0653      1.5784
cmt_pct_lag1                    -0.3224     0.1909    -1.6888     0.0914     -0.6967      0.0519
stock_expected_return            0.8970     0.2023     4.4337     0.0000      0.5003      1.2937
volatility                       0.1464     0.0632     2.3146     0.0207      0.0224      0.2704
avg_daily_volume_mil             0.0080     0.0013     6.2751     0.0000      0.0055      0.0105
================================================================================================

F-test for Poolability: 3.5368
P-value: 0.0000
Distribution: F(106,2782)

Included effects: Entity, Time
```

print(fe_results.summary)

Appendix 5

Detailed regression results: bullish periods (positive IR_{t-1})

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:           stock_return   R-squared:                        0.2272
Estimator:                   PanelOLS   R-squared (Between):             -3.6743
No. Observations:                1450   R-squared (Within):               0.5816
Date:                Tue, May 13 2025   R-squared (Overall):              0.5409
Time:                        14:12:09   Log-likelihood                   -3764.4
Cov. Estimator:            Clustered
                                        F-statistic:                      44.523
Entities:                          50   P-value                           0.0000
Avg Obs:                       29.000   Distribution:                  F(9,1363)
Min Obs:                       29.000
Max Obs:                       29.000   F-statistic (robust):             10.030
                                        P-value                           0.0000
Time periods:                      29   Distribution:                  F(9,1363)
Avg Obs:                       50.000
Min Obs:                       50.000
Max Obs:                       50.000

                                Parameter Estimates
================================================================================================
                              Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------------------------
stock_rank_in_week_pctl_lag1     0.0031     0.0036     0.8732     0.3827     -0.0039      0.0101
bigV_pct_lag1                   -0.0211     0.0147    -1.4368     0.1510     -0.0499      0.0077
avg_clicks_1000_lag1             0.0031     0.0863     0.0358     0.9714     -0.1662      0.1724
like_pct_lag1                    0.0122     1.1752     0.0104     0.9917     -2.2931      2.3175
fwd_pct_lag1                     1.3136     0.5031     2.6111     0.0091      0.3267      2.3006
cmt_pct_lag1                     0.4178     0.5886     0.7098     0.4780     -0.7369      1.5725
stock_expected_return            1.0166     0.2380     4.2712     0.0000      0.5497      1.4835
volatility                       0.1039     0.0278     3.7418     0.0002      0.0494      0.1583
avg_daily_volume_mil             0.0071     0.0011     6.3970     0.0000      0.0049      0.0093
================================================================================================

F-test for Poolability: 2.6056
P-value: 0.0000
Distribution: F(77,1363)

Included effects: Entity, Time
```

print(fe_results_up.summary)

Appendix 6

Detailed regression results: bearish periods (negative IR_{t-1})

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:           stock_return   R-squared:                        0.1138
Estimator:                   PanelOLS   R-squared (Between):             -3.3506
No. Observations:                1398   R-squared (Within):               0.2348
Date:                Tue, May 13 2025   R-squared (Overall):             -0.0108
Time:                        14:12:09   Log-likelihood                   -3558.9
Cov. Estimator:            Clustered
                                        F-statistic:                      18.728
Entities:                          50   P-value                           0.0000
Avg Obs:                       27.960   Distribution:                  F(9,1312)
Min Obs:                       26.000
Max Obs:                       28.000   F-statistic (robust):             48.867
                                        P-value                           0.0000
Time periods:                      28   Distribution:                  F(9,1312)
Avg Obs:                       49.929
Min Obs:                       49.000
Max Obs:                       50.000

                                Parameter Estimates
================================================================================================
                              Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------------------------
stock_rank_in_week_pctl_lag1     0.0128     0.0045     2.8665     0.0042      0.0041      0.0216
bigV_pct_lag1                    0.0220     0.0155     1.4161     0.1570     -0.0085      0.0524
avg_clicks_1000_lag1            -0.1071     0.0191    -5.5996     0.0000     -0.1447     -0.0696
like_pct_lag1                    0.2639     0.7990     0.3303     0.7412     -1.3035      1.8313
fwd_pct_lag1                     0.4564     0.5855     0.7796     0.4357     -0.6921      1.6050
cmt_pct_lag1                    -0.3507     0.0724    -4.8428     0.0000     -0.4927     -0.2086
stock_expected_return            0.7515     0.2792     2.6920     0.0072      0.2039      1.2991
volatility                       0.2840     0.1947     1.4588     0.1448     -0.0979      0.6658
avg_daily_volume_mil             0.0091     0.0019     4.6898     0.0000      0.0053      0.0130
================================================================================================

F-test for Poolability: 3.5861
P-value: 0.0000
Distribution: F(76,1312)

Included effects: Entity, Time
```

print(fe_results_dn.summary)

Works Cited

Baker, Malcolm, et al. "Global, Local, and Contagious Investor Sentiment." *Journal of Financial Economics*, vol. 104, no. 2, May 2012, pp. 272–87. *ScienceDirect*, https://doi.org/10.1016/j.jfineco.2011.11.002.

Baker, Malcolm, and Jeffrey Wurgler. "Investor Sentiment in the Stock Market." *The Journal of Economic Perspectives*, vol. 21, no. 2, 2007, pp. 129–51. https://www.jstor.org/stable/30033721.

Bashir, Usman, et al. "Investor Sentiment and Stock Price Crash Risk: The Mediating Role of Analyst Herding." *Computers in Human Behavior Reports*, vol. 13, Mar. 2024, p. 100371. *ScienceDirect*, https://doi.org/10.1016/j.chbr.2024.100371.

Black, Fischer. "Noise." *The Journal of Finance*, vol. 41, no. 3, 1986, pp. 528–43. *Wiley Online Library*, https://doi.org/10.1111/j.1540-6261.1986.tb04513.x.

Cui, Jiayi, et al. "How Retail vs. Institutional Investor Sentiment Differ in Affecting Chinese Stock Returns?" *Journal of Risk and Financial Management*, vol. 18, no. 2, 2025, p. 95. *ProQuest*, https://doi.org/10.3390/jrfm18020095.

DeepSeek-AI, et al. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." *arXiv*, 22 Jan. 2025. *arXiv.org*, https://doi.org/10.48550/arXiv.2501.12948. Accessed 12 May 2025.

Deepseek-AI. "Deepseek-Ai/DeepSeek-R1 · Hugging Face." *Hugging Face*, 21 Jan. 2025, https://huggingface.co/deepseek-ai/DeepSeek-R1. Accessed 12 May 2025.

De Long, J. Bradford, et al. "Noise Trader Risk in Financial Markets." *Journal of Political*

Economy, vol. 98, no. 4, 1990, pp. 703–38. https://www.jstor.org/stable/2937765.

Fama, Eugene F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance*, vol. 25, no. 2, 1970, pp. 383–417. *JSTOR*, https://doi.org/10.2307/2325486.

Frazzini, Andrea, and Owen A. Lamont. "Dumb Money: Mutual Fund Flows and the Cross-Section of Stock Returns." *Journal of Financial Economics*, vol. 88, no. 2, May 2008, pp. 299–322. *ScienceDirect*, https://doi.org/10.1016/j.jfineco.2007.07.001.

FTSE Russell. "FTSE China A50." *LSEG*, 25 Mar. 2025, https://research.ftserussell.com/analytics/factsheets/Home/DownloadConstituentsWeights/?indexdetails=XINA50.

Hirshleifer, David, and Tyler Shumway. "Good Day Sunshine: Stock Returns and the Weather." *The Journal of Finance*, vol. 58, no. 3, 2003, pp. 1009–32. https://www.jstor.org/stable/3094570.

Kamstra, Mark J., et al. "Winter Blues: A SAD Stock Market Cycle." *The American Economic Review*, vol. 93, no. 1, 2003, pp. 324–43. https://www.jstor.org/stable/3132178.

看那小乔流水 [Kan Na Xiao Qiao Liu Shui]. "都是抄垃圾，还不如弄点好玩的。炒这个挺好" ["We are investing in rubbish anyway, so why not invest in something fun? This is a fairly good option"]. 东方财富股吧 [*Eastmoney Stock Bar*], 6 Nov. 2024, 6:20 p.m., https://guba.eastmoney.com/news,002253,1480577652.html. Accessed 23 March 2025.

Shleifer, Andrei, and Robert W. Vishny. "The Limits of Arbitrage." *The Journal of Finance*, vol. 52, no. 1, 1997, pp. 35–55. *JSTOR*, https://doi.org/10.2307/2329555.

S&P CapitalIQ. “Wisesoft Co., Ltd. (SZSE:002253) Financials > Income Statement.” *S&P CapitalIQ*, https://www.capitaliq.com/CIQDotNet/Financial/IncomeStatement.aspx?CompanyId=46234348. Accessed 23 March 2025.

Zhou, Winni, and Summer Zhen. “China Retail Investors from Gen-Z to Retirees Sit out Stock Rally.” *Reuters*, 15 Feb. 2023. *www.reuters.com*, https://www.reuters.com/markets/asia/china-retail-investors-gen-z-retirees-sit-out-stock-rally-2023-02-15/. Accessed 12 May 2025.