

Attention and Opinion Content in “Great Era” and the Cross-Section of Chinese A-Share Returns: Out-of-Sample Evidence from 6.26 Million LLM-Annotated Posts

Shiyu Guan^a, IRIS^a

^a*EROS Quantitative Laboratory, China*

Abstract

This paper asks whether online discussion by Chinese retail investors contains tradable information and, if so, whether that information lies in what investors say or in how their attention is distributed. We construct a floor-level archive of the “Great Era” stock board on NGA, a large Chinese gaming community, covering 2016–2026 (6.47 million posts, 342 thousand threads, and 128 thousand users), and use large language models to produce structured labels for 6.09 million posts. The annotation system combines closed-set entity linking with stock-level stances, sentiment direction and intensity, panic and euphoria, irony, and speech acts. Under preregistration, frozen acceptance thresholds, within-family false-discovery control, mandatory style orthogonalization, and a code-enforced out-of-sample holdout, we test 20 hypotheses in seven families. Every surviving signal measures the distribution of attention: stock-level discussion heat, bull–bear disagreement, and sector-level heat. In the 2024–2026H1 holdout, these signals retain 80–95% of their exploration-period strength and remain negative predictors: crowded stocks and sectors underperform in the following week. By contrast, opinion content—sentiment direction, emotion intensity, self-reported trades, and poster track records—has no incremental predictive power. Three candidates with raw $|t|$ statistics of 3.9–5.1 lose 82–92% of their information after style regression and are therefore redundant expressions of turnover and short-term momentum. Two independent measures also reject the popular claim that retail capitulation marks a bottom: retail panic predicts continuation, not reversal. The results show why text-based factor research requires both a style gate and a genuinely protected holdout.

Keywords: investor attention, retail investors, social media text, large language models, cross-sectional returns, preregistration
JEL: G12, G14, G40

1. Introduction

Retail investors account for a substantial share of trading in China’s A-share market. Their online discussions therefore provide a direct view of how attention, beliefs, and trading narratives form outside institutional channels. This paper asks two nested questions. First, do these discussions contain information about the cross-section of future returns? Second, if they do, is that information in the *content* of opinions—what investors say, whether they are bullish or bearish, and how strongly they feel—or in the *distribution* of attention—how many people are speaking, which stocks and sectors they discuss, and whether bulls and bears are simultaneously present?

Prior evidence suggests that attention-induced buying can raise the prices of salient stocks and depress subsequent returns^[1], while concentrated retail trading can itself move prices in the short run^[2]. With short-sale constraints, disagreement can produce overvaluation and lower subsequent returns^[3, 4, 5]. Message-board and forum text has repeatedly been associated with trading and returns^[6, 7, 8]; media content has causal price effects^[9]; and search volume has become a standard attention proxy^[10]. A persistent identification concern remains: how much apparent text-factor predictability is merely the shadow of size, turnover, and recent momentum? Multiple testing and post-publication decay are well-established problems in factor discovery^[11, 12, 13]. Without style regression and a protected holdout, the concern cannot be resolved.

The NGA corpus is unusually demanding and therefore informative. NGA is one of China’s largest gaming communities; its stock board attracts young, internet-native retail investors. Irony is routine, and securities are often referred to through community slang. Dictionary methods are structurally ill-suited to this language. We archive the board at floor level over 2016–2026 and use large language models to label 6.09 million posts^[14, 15]. The central annotation design is *closed-set entity linking*: the model may select a security only from a locally generated candidate list that includes manually reviewed slang. Each stance is therefore linked to an exchange code, creating the cross-sectional dimension used in the tests.

The empirical design is enforced in code. Each hypothesis family is pre-registered before execution; signal definitions, directional priors, acceptance thresholds, and prohibited scans are frozen; false discoveries are controlled within family; and every stock-level survivor must retain information after removing size, turnover, recent return, and industry exposure. Return data from 2024-01 through 2026-06 are protected by a guard that permits one budgeted query per final candidate. Beginning in 2026-07, a forward shadow ledger records signals every week without reading returns.

The main result is concise: *retail attention contains information, but retail opinion content does not*. Three negative attention signals survive all gates. Stock-level discussion heat has an exploration-period rank IC of -0.032 and a holdout IC of -0.031 ; bull–bear disagreement moves from -0.023 to -0.018 ; sector-level heat moves from -0.060 to -0.054 and is the strongest signal in the corpus. Net sentiment, sentiment revisions, emotion intensity, irony, self-reported trades, poster skill, and newcomer share add no information. Three conventionally significant candidates are rejected as style-redundant, and the hypothesis that retail capitulation marks a bottom is rejected by both behavioral and emotional measures.

The paper contributes a decade-long, floor-level retail-forum archive; a large set of exchange-linked structured annotations; out-of-sample evidence that predictability lies in attention rather than opinion content; and a portable testing protocol for LLM-based text factors. The one unresolved observation is reported rather than rationalized: all three surviving signals flatten in 2026H1. The holdout is not reopened to explain the anomaly; that question is delegated to the forward ledger.

2. Data

2.1. Forum archive

NGA’s stock board (forum id 706) is a general A-share discussion venue inside a gaming community. The harvester archives pages through three fallbacks—a mobile endpoint, web parsing, and a headless browser—records every floor of every thread, and stores page-level archive status for resumability and audit. Table 1 summarizes the corpus, and Figure 1 shows its time profile. Pre-2020 coverage largely consists of a few long-lived megathreads that survived years of deletion and is not representative; all cross-sectional tests therefore begin in 2020-01.

Table 1: Corpus and annotation summary.

Archive window	2016-01 – 2026-07 (rolling)
Posts (floor level)	6,469,601
Threads	341,629
Distinct users	128,148
Posts submitted for annotation	6,263,930
Valid annotations	6,088,848 (97.2%)
Date-join loss against archive	0
Entity-linked stances	5,193,986
Posts with an investment-relevant stance	71.2%
Speech acts (opinion/question/trade/joke/news)	69.6 / 12.0 / 9.8 / 6.6 / 2.0%
Irony flagged	19.7%
Panic ≥ 1 (0–3 scale)	24.6%
Euphoria ≥ 1 (0–3 scale)	27.5%

Counts come from the v2 label store and the audited feature mart. Stance counts are obtained by expanding each post-level stance array.

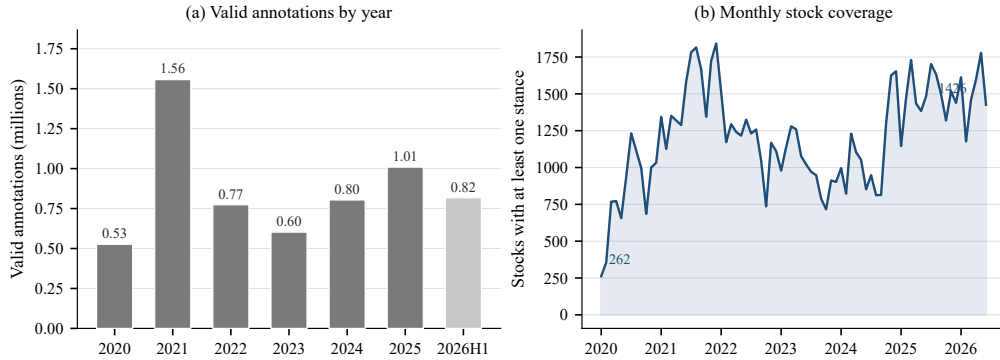


Figure 1: Corpus scale and stock coverage. Panel (a) reports valid annotations by year; panel (b) reports the number of stocks receiving at least one stance in each month. Coverage rises from 262 stocks in 2020-01 to more than 1,800 at the 2021 peak and equals 1,426 in 2026-06. Source: Feature Mart V3.

2.2. LLM annotation and closed-set entity linking

Each post is submitted with its thread title, quoted parent post, and a local candidate list. The model returns relevance; a stance array containing

target type, target code, aspect, direction, and horizon; speech act; panic and euphoria scores; irony; and confidence. Mechanical validation checks batch length, echoed post identifiers, schema validity, and membership of every target code in the candidate set. A failed check retries the whole batch.

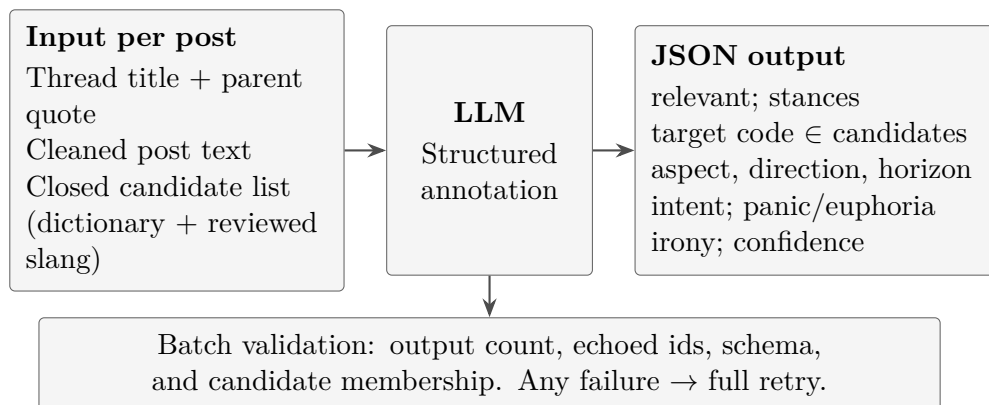


Figure 2: Annotation pipeline. The model cannot supply a security code from memory; entity linking is entirely constrained by the candidate set.



Figure 3: An anonymized real input and output, translated into English. With an empty candidate list, the model must return an unknown target rather than invent a stock code. The champagne phrase is labeled by intended meaning as negative and ironic.

Two decisions are important. Irony is labeled by intended meaning rather than surface polarity. Closed-set linking also prevents the model from inventing codes: slang enters the candidate list only after manual review. Table 2 gives four examples covering irony, slang-based entity linking, a self-reported trade, and multiple stances in one post.

Table 2: Real annotation examples, translated and lightly shortened.

Date	Post excerpt	Key annotation and use case
2021-06-22	“Champagne sales have collapsed. Save the children.”	Direction -1 ; panic 2; irony true; confidence 85. Irony recognition.
2025-09-26	“When I see another ‘Haigou hit limit-up’ post, I know it is probably up only one percent.”	Target 601919.SH (COSCO Shipping Holdings); joke; irony true. Slang-based entity linking.
2026-03-03	“I have held Haigou for more than five years. Even I cut most of the position today and kept only a token lot.”	Target 601919.SH; direction -1 ; self-reported trade.
2025-09-26	“Bullish for the long run, but the original poster sold half yesterday.”	Target 601869.SH; long-run fundamental stance $+1$ and short-run price stance -1 . Multiple stances and a belief-action gap.

Post identifiers and authors are omitted. “Haigou” is a community nickname for COSCO Shipping Holdings and enters the candidate set through a manually reviewed slang dictionary.

2.3. Returns, universe, and aggregation

Signals are aggregated by ISO week. A stock must receive at least three stances in the signal week and belong to the point-in-time main-board universe: at least 60 trading days since listing, 20-day average turnover of at least RMB 20 million, and neither ST status nor suspension. Weeks with fewer than 50 eligible stocks are skipped. The outcome is close-to-close return over the next five trading days. The median weekly universe contains 149 stocks in the exploration period and 186 in the holdout. Sector tests use a frozen mapping from 18 forum sectors to exchange industries; broad narratives without a clean mapping are excluded and the coverage loss is recorded.

2.4. Biases and where they are removed

Three biases affect historical estimates. Returns from 2020–2023 had been used in earlier research workflows; the annotation models’ training data may cover the sample period, although the models see text rather than returns; and pre-archive history is a survivor sample after years of deletion. The 2024–2026H1 holdout removes prior return exposure. The forward period from 2026-07, collected and labeled in real time, removes all three. This hierarchy determines the strength of evidence assigned to each result.

3. Research design

3.1. Sample partitions and testing budget

All hypotheses are first tested in 2020–2023. A code guard protects 2024-01–2026-06: any endpoint that touches the period raises an exception unless an environment confirmation matches a preregistered entry in an append-only budget ledger. The initial budget is six, each final candidate may be queried once, and consumption is irreversible. From 2026-07, the forward ledger records signal values and top-quintile names without reading returns.

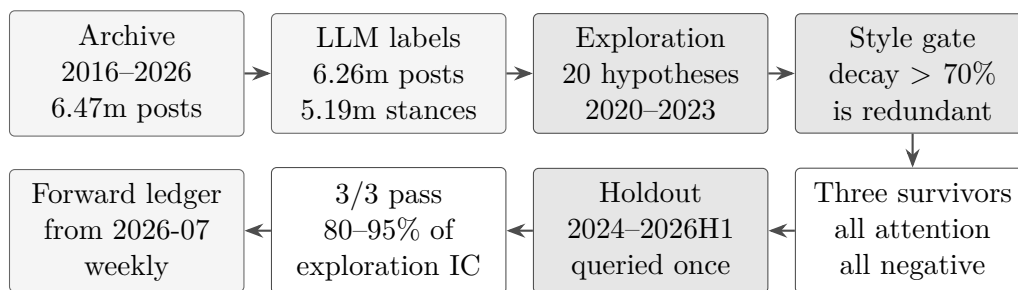


Figure 4: Research workflow and its ex ante constraints: family-level preregistration, within-family false-discovery control, style orthogonalization, a holdout budget ledger, and a forward shadow ledger.

3.2. Preregistration and acceptance thresholds

Each family freezes signal definitions, directional priors, thresholds, and a no-scan list before execution. Ex post searches over aggregation windows, thresholds, aspects, or horizons are prohibited. Acceptance requires: C1, $|IC| \geq 0.010$ with the registered sign; C2, the top quintile’s weekly excess return relative to the pool is at least 5 bp in the registered direction; and C3, at least three of four exploration years have the correct IC sign. Newey–West t statistics with four lags enter a within-family Benjamini–Hochberg procedure at $q = 0.10$. A wrong sign is a rejection, not an invitation to reverse the hypothesis. Twenty of a maximum 24 primary-test budget entries are consumed. The only structured direction check is itself registered and uses non-overlapping direction and evaluation periods.

3.3. Style orthogonalization

Retail discussion naturally concentrates in smaller, recently rising, high-turnover stocks. Each stock-level survivor is retested on the weekly cross-

sectional residual from

$$\begin{aligned} \text{signal} \sim & \ln(\text{float market cap}) + \ln(1 + \text{weekly turnover}) \\ & + \text{prior weekly return} + \text{industry fixed effects.} \end{aligned} \quad (1)$$

A loss of more than 70% of $|\text{IC}|$ implies style redundancy and rejection regardless of the raw t statistic. A second rule rejects a candidate as a mechanical heat variant when its mean weekly rank correlation with pure discussion heat exceeds 0.70.

4. Main results: out-of-sample validation of attention signals

Table 3: The three surviving signals in exploration, style regression, and the out-of-sample holdout.

Signal	Exploration: 2020–2023			Style gate	Holdout: 2024–2026H1	
	IC	Top 20%	Years	Orthogonal IC	IC	Top 20%
Stock heat	−0.032	−6.7 bp	4/4	−0.011 (66%)	−0.031	−22.5 bp
Disagreement	−0.023	−12.8 bp	4/4	−0.010 (57%)	−0.018	−28.0 bp
Sector heat	−0.060	−28.9 bp ^a	4/4	— ^b	−0.054	−20.7 bp ^a

Weekly rank IC predicts return over the next five trading days. “Top 20%” is the top signal quintile’s mean weekly excess return relative to the eligible pool. Each holdout candidate is queried once.

^a For the sector signal, the column reports the hot-minus-cold sector portfolio spread.

^b A stock-level style regression is not applicable; the preregistered time-series redundancy test against aggregate turnover is passed.

Stock heat is $\ln(1+n)$ for the number of stock-level stances in a week. Disagreement is the smaller of bullish and bearish stance counts divided by their sum. Sector heat is $\ln(1+n)$ for stances mapped to a sector, evaluated against the sector’s next-week equal-weighted return. All three registered signs are negative. The results are consistent with attention-driven overpricing^[1] and disagreement under short-sale constraints^[3, 5]. The sector signal is strongest, suggesting that crowding is completed more fully at the narrative level than at the ticker level.

The style gate removes roughly half of the two stock signals but does not eliminate them. Heat retains an orthogonal IC of -0.011 with the correct sign in all four exploration years; disagreement retains -0.010 with the correct sign in three. The raw effects therefore combine a turnover-measurable

component with an independent component associated with which stocks attract retail attention.

The three frozen candidates are then submitted together to the code-protected 2024–2026H1 holdout. All pass the registered thresholds and retain 80–95% of exploration strength. Figures 5 and 6 show that the result is not driven by a small number of weeks.

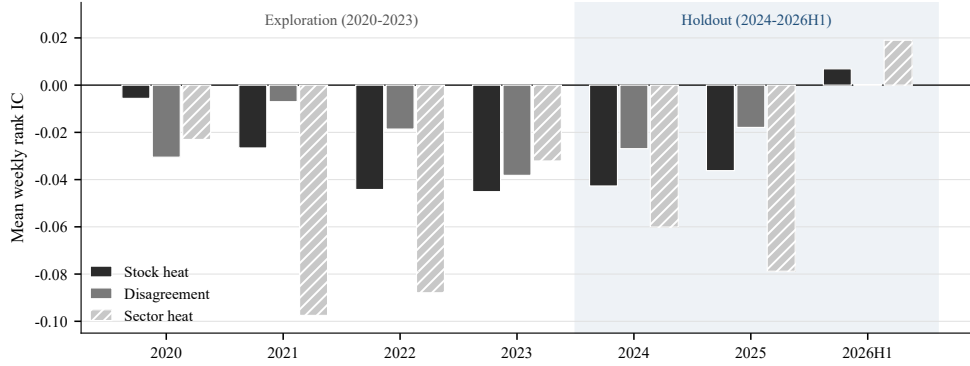


Figure 5: Annual mean weekly rank IC for the three surviving signals. The shaded region is the code-protected holdout, queried once. All three reproduce in 2024–2025 and flatten together in 2026H1.

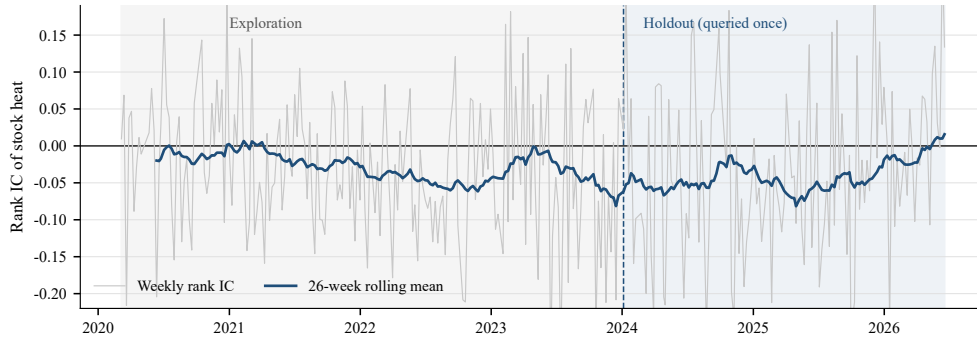


Figure 6: Weekly rank IC and its 26-week rolling mean for stock heat, 2020-01 through 2026-06 (317 weeks). The dashed line separates exploration from the holdout.

4.1. The 2026H1 flattening

In 2026H1 all three signals flatten or turn slightly positive: heat +0.0069, disagreement +0.0002, and sector heat +0.0190. Joint failure in the same

half-year may reflect low power in a 26-week sample, a momentum-dominated market regime, or drift in the language and therefore in the annotation instrument. The protected period cannot be reopened to discriminate among these explanations. The forward ledger supplies the only admissible answer: continued flattening will classify the mechanism as weakened; recovery of negative IC will classify 2026H1 as short-sample noise.

5. Identifying style redundancy: three counterexamples

Table 4: Significant raw candidates rejected as style-redundant.

Signal	Raw IC	NW t	Orthogonal IC	Decay
Buy-report crowding	−0.029	−4.16	+0.002	92%
Cold-start heat burst	−0.047	−5.07	+0.009	82%
Euphoria share	−0.029	−3.86	+0.003	90%

All three candidates pass the ordinary raw-IC thresholds, show the correct sign in every year, and have $|t|$ statistics between 3.9 and 5.1. After residualizing size, turnover, recent return, and industry, each becomes statistically indistinguishable from zero. Apparent reversal after reported buying is therefore reversal in recently rising, high-turnover stocks rather than incremental information from the trade or emotion label. Any text factor built from attention-heavy retail data should be presumed to repeat liquidity styles until it passes this regression.

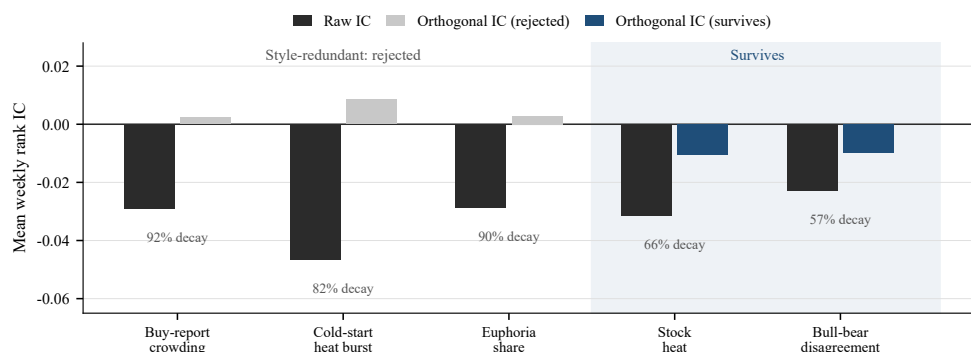


Figure 7: Raw and style-orthogonalized IC. The first three signals are rejected as style-redundant; the two surviving stock-level attention signals lose about half their strength but remain informative.

6. Systematic null results for opinion content

Net sentiment has no cross-sectional information (IC +0.004), and neither does the change in sentiment relative to its four-week baseline. Panic, post-orthogonalization euphoria, and irony also add no information. The gap between stated opinion and self-reported action has an IC of +0.008.

Poster-level skill is tested by estimating a Bayesian-shrunk 52-week rolling directional hit rate for each user. The resulting signal covers 4,679 eligible users over 180 weeks and has an IC of -0.0001 : predictive ability does not persist at the individual level. Newcomer share reaches significance in the middle of its distribution but its highest quintile earns +11.7 bp, opposite the registered negative direction. Question share fails in the same way. Breadth and author concentration are mechanically entangled with heat, with rank correlations of +0.93 and -0.84 , and are rejected as non-independent.

The claim that retail capitulation marks a bottom receives two independent tests. Conditional on a 20-day drawdown of at least 10%, high self-reported selling or high panic intensity is registered as a positive reversal signal. Both estimates are negative. The behavioral measure has IC -0.032 and the correct registered sign in only one of four years; the emotional measure has the wrong sign in all four. Retail panic therefore confirms continuation in this sample rather than marking a bottom. The sign is not reversed and retested.

Eight market-timing configurations—dictionary panic, euphoria, net sentiment, megathread share, LLM aggregate sentiment, new-account inflow,

and two megathread variants—all fail. Aggregate forum sentiment does not time the index under any admissible specification, consistent with the limited and unstable time-series predictability of aggregate sentiment measures [16, 17].

7. Discussion

7.1. Why attention survives and content does not

A noise-trader interpretation is parsimonious^[18]. Individual directional opinions are numerous and offset one another, but the concentration of opinion holders is itself a market event. A crowd around a stock indicates short-run demand pressure and later reversal; simultaneous bullish and bearish crowds indicate stronger disagreement; and a crowd around an entire narrative produces the strongest signal. That all three effects are negative, with no profitable configuration based on following opinion content, also accords with evidence that retail investors as a group underperform^[19].

7.2. Implications for text-factor research

The annotation instrument matters. The four dictionary-based configurations all fail on the same corpus, whereas every survivor requires LLM labels. In a community where irony is ordinary and tickers are replaced by slang, general dictionaries are misspecified by construction^[20]. Yet the economic value of the expensive annotation layer concentrates in one field: closed-set entity linking. Fine-grained sentiment fields add no tradable information. Teams allocating LLM budgets to factor research may therefore benefit more from entity resolution and validation than from increasingly fine emotion taxonomies^[21]. Finally, the rejected significant candidates show that style regression and a protected holdout are substantive identification tools, not administrative formalities.

7.3. Scope and limitations

A deep-value benchmark consisting of the top 50 stocks by earnings yield has only 6.8% weekly coverage in the forum. Attention signals therefore apply to the high-attention thematic universe, where they may serve as filters or short candidates, rather than as overlays on deep value. Because the signals are negative, implementation requires avoidance or a costly short leg; the paper makes no claim about net-of-cost returns.

The evidence comes from one forum, one market, and weekly data, with a median eligible cross-section of 149 stocks in exploration. Historical biases are disclosed in Section 2.4; only forward accumulation removes all of them. The simultaneous 2026H1 flattening remains unresolved by design.

8. Conclusion

Six million annotations, 20 preregistered hypotheses, and a code-protected holdout support a simple conclusion: in a decade of Chinese retail-forum data, tradable information lies in the distribution of attention—which stocks and sectors attract crowds and how intensely they are disputed—not in opinion content. The broader lesson is methodological. Without style orthogonalization and a protected holdout, this study would have reported three statistically strong but economically redundant sentiment factors.

Appendix A. Complete hypothesis registry

Table A.5: All preregistered tests in execution order. The cap was 24 primary tests and 20 were consumed. QUARREL was refunded when its instrument proved absent and was later closed on mechanism grounds.

Round	Family / signal (registered sign)	Verdict
IT17	Dictionary timing: PANIC / EUPHORIA / NET-SENT / MEGA (+/-/-/-)	All four rejected
IT18	LLM cross-section: HEAT / SENT / DIS-AGREE (-/-/-)	Pass / reject / pass
F8	Style audit of IT18 survivors	Both survive
IT19	TA-CROWD / TA-CAPITULATION / TALK-ACT-GAP (-/+/-)	Redundant / wrong sign / no information
	BREADTH / AUTHOR-HHI / NEWCOMER / NEW-ACCOUNTS (-/+/-/-)	Two heat variants; two rejected
IT20	HEAT-BURST / HEAT-PERSIST / SENT-REVERSAL (-/+/-)	Redundant / reject / reject
	EUPHORIA / PANIC-REVERSAL / IRONY (-/+/-)	Redundant / wrong sign / reject
IT21	SECTOR-HEAT / SPILLOVER / MEGA-MOMENTUM (-/check/+)	Pass / reject / reject
IT22	USER-ALPHA / QUARREL / QUESTION / ORGANIC-HEAT (+/-/-/-)	Reject / absent / reject / redundant
IT23	Holdout: HEAT / DISAGREE / SECTOR-HEAT (-/-/-)	Three passes

Note: Style redundancy means more than 70% IC decay under the regression in Section 3.3; heat redundancy means mean weekly rank correlation with HEAT above 0.70; a wrong sign is rejected without reversal and retesting.

Appendix B. Detailed round-level results

Table B.6: IT17: dictionary-based market timing ($n = 203$ weeks).

Signal	ρ (expected sign)	Q5–Q1	Annual sign	Verdict
PANIC	-0.002 (+)	-26 bp	2/4	No information
EUPHORIA	-0.101 (-)	-91 bp	2/4	Unstable
NET_SENT	-0.078 (-)	+15 bp	4/4	Intensity and monotonicity fail
MEGA_SHARE	+0.131 (expected -)	+50 bp	1/4	Wrong sign

Table B.7: IT18 and F8: initial LLM cross-section and style audit.

Signal	Raw IC	Top 20%	Years	Orthogonal IC	Decay	Verdict
NGA_HEAT	−0.0316	−6.7 bp	4/4	−0.0106	66%	Pass
NGA_SENT	+0.0040	+35.9 bp	2/4	—	—	Reject
NGA_DISAGREE	−0.0230	−12.8 bp	4/4	−0.0098	57%	Pass

Table B.8: IT19: trading behavior and user structure.

Signal	IC	NW t	Top 20%	Years	Heat corr.	Orth. decay	Verdict
TA_CROWD	−0.0291	−4.16	−26.3 bp	4/4	+0.54	92%	Style-redundant
TA_CAPITULATION	−0.0315 (expected +)	−1.17	−63.4 bp	1/4	+0.36	—	Wrong sign
TALK_ACT_GAP	+0.0083	+1.15	+36.9 bp	1/4	−0.01	—	No information
NGA_BREADTH	−0.0279	−4.34	−7.6 bp	4/4	+0.93	55%	Heat-redundant
AUTHOR_HHI	+0.0239	+3.67	+8.9 bp	4/4	−0.84	48%	Heat-redundant
NEWCOMER_SHARE	−0.0157	−2.45	+11.7 bp	3/4	−0.21	—	Non-monotone
NEW_ACCOUNT_INFLUX	+0.084 (expected −)	—	+60 bp	3/4	—	—	Wrong sign

Table B.9: IT20: dynamic patterns and emotion extremes.

Signal	IC	NW t	Top 20%	Years	Heat corr.	Orth. decay	Verdict
HEAT_BURST	−0.0466	−5.07	−15.6 bp	4/4	+0.81	82%	Style-redundant
HEAT_PERSIST	+0.0091	+1.32	+14.8 bp	3/4	+0.36	—	Too weak
SENT_REVERSAL	+0.0040	+0.56	+11.3 bp	1/4	−0.09	—	No information
STOCK_EUPHORIA	−0.0288	−3.86	−8.7 bp	4/4	+0.25	90%	Style-redundant
STOCK_PANIC_REV	−0.0091 (expected +)	−0.50	+16.1 bp	0/4	+0.32	—	Wrong sign
IRONY_DENSITY	−0.0062	−0.97	−1.4 bp	3/4	+0.34	—	Too weak

Table B.10: IT21 and IT22: sector, user skill, and discourse structure.

Round	Signal	Main statistic	Years	Verdict
IT21	SECTOR_HEAT_ROT	IC −0.0603; NW $t = -3.33$; hot-minus-cold −28.9 bp	4/4	Pass
IT21	SECTOR_SPILLOVER	Evaluation spread −25.8 bp; $t = -1.43$	2/3	Unstable
IT21	MEGA_SHARE_MOM	$\rho = +0.139$; Q5−Q1 +54 bp	2/4	Unstable
IT22	USER_ALPHA	IC −0.0001; NW $t = -0.01$	2/4	No persistence
IT22	QUARREL	Instrument absent	—	Closed
IT22	QUESTION_RATIO	IC −0.0135; top 20% +1.1 bp	4/4	Non-monotone
IT22	ORGANIC_HEAT	IC −0.0299; heat corr. +0.98	3/4	Heat-redundant

Appendix C. Holdout details

Table C.11: IT23 holdout results by subperiod (2024-01 – 2026-06).

Candidate	Exploration	Full holdout	2024	2025	2026H1	Verdict
NGA_HEAT_v1	−0.032	−0.0310	−0.0427	−0.0362	+0.0069	Pass
NGA_DISAGREE_v1	−0.023	−0.0183	−0.0269	−0.0179	+0.0002	Pass
SECTOR_HEAT_ROT_v1	−0.060	−0.0535	−0.0601	−0.0788	+0.0190	Pass

Three of six holdout budget entries are consumed. Passing requires the registered sign, absolute IC at least half the exploration benchmark, and the correct sign in at least two of three subperiods.

Table C.12: Exploration-period annual IC used in Figure 5.

Signal	2020	2021	2022	2023
NGA_HEAT	−0.0056	−0.0266	−0.0442	−0.0451
NGA_DISAGREE	−0.0305	−0.0070	−0.0186	−0.0382
SECTOR_HEAT_ROT	−0.0230	−0.0975	−0.0879	−0.0321

Appendix D. Research-discipline record

- The primary-test cap is 24 and 20 are consumed. QUARREL is re-funded because its instrument is absent.
- One direction check is registered for sector spillover, with non-overlapping direction (2020-01–2021-06) and evaluation (2021-07–2023-12) periods.
- The holdout begins with six entries; IT23 consumes three and leaves three. The append-only ledger stores preregistration hashes and verdicts.
- Four wrong-sign results are rejected without reversal and retesting.
- Retail capitulation as a bottom, market-level timing, individual poster skill, and both directions of megathread share are permanently closed.
- From 2026-07, the forward ledger records weekly signal values and top quintiles without accessing returns.

Appendix E. Reproducibility

The online replication package contains the annotation prompt and schema, 14 preregistration and result documents, eight probe and holdout scripts, per-week statistics for every test, the multiple-testing ledger, and the holdout budget ledger. The raw archive (17.8 GB) and label store (1.3 GB) are available from the authors. `build_mart_v3.py` reconstructs the feature mart from the two SQLite stores, and `make_figs_en_v2.py` generates every data figure in this version.

References

- [1] Barber, B.M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies* 21, 785–818.
- [2] Barber, B.M., Odean, T., Zhu, N., 2009. Do retail trades move markets? *Review of Financial Studies* 22, 151–186.
- [3] Miller, E.M., 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32, 1151–1168.
- [4] Cookson, J.A., Niessner, M., 2020. Why don’t we agree? Evidence from a social network of investors. *Journal of Finance* 75, 173–228.
- [5] Diether, K.B., Malloy, C.J., Scherbina, A., 2002. Differences of opinion and the cross section of stock returns. *Journal of Finance* 57, 2113–2141.
- [6] Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59, 1259–1294.
- [7] Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms’ fundamentals. *Journal of Finance* 63, 1437–1467.
- [8] Sabherwal, S., Sarkar, S.K., Zhang, Y., 2011. Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting* 38, 1209–1237.

- [9] Engelberg, J.E., Parsons, C.A., 2011. The causal impact of media in financial markets. *Journal of Finance* 66, 67–97.
- [10] Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. *Journal of Finance* 66, 1461–1499.
- [11] McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.
- [12] Harvey, C.R., Liu, Y., 2015. Backtesting. *Journal of Portfolio Management* 42, 13–28.
- [13] Harvey, C.R., Liu, Y., Zhu, H., 2016. ...and the cross-section of expected returns. *Review of Financial Studies* 29, 5–68.
- [14] Gilardi, F., Alizadeh, M., Kubli, M., 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, e2305016120.
- [15] Lopez-Lira, A., Tang, Y., 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. Working paper, arXiv:2304.07619.
- [16] Da, Z., Engelberg, J., Gao, P., 2015. The sum of all FEARS: Investor sentiment and asset prices. *Review of Financial Studies* 28, 1–32.
- [17] Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance* 62, 1139–1168.
- [18] Foucault, T., Sraer, D., Thesmar, D.J., 2011. Individual investors and volatility. *Journal of Finance* 66, 1369–1406.
- [19] Barber, B.M., Odean, T., 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.
- [20] Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66, 35–65.
- [21] Korinek, A., 2023. Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature* 61, 1281–1317.