

大时代注意力、观点内容与 A 股横截面收益： 基于 626 万条大模型标注的样本外证据

管世宇 IRIS

EROS 量化实验室

内部审阅稿

摘 要

本文研究中国散户投资者的网络讨论是否含有可交易信息；若有，信息存在于讨论的内容，还是讨论的分布。本文对国内最大游戏社区 NGA 的股票板块“大时代”完成了 2016–2026 年楼层级全量留档（647 万帖、34.2 万主题、12.8 万用户），并利用大语言模型对其中 609 万帖完成结构化标注。标注体系以闭集实体链接为核心，覆盖个股立场、情绪方向与强度、恐慌与亢奋、反讽及言语行为等维度。在预注册、门槛冻结、族内错误发现率控制、风格正交与代码强制的样本外保留段这一整套研究设计下，本文系统检验了七族共 20 个假设。结果显示：语料中全部可交易信息均来自“注意力的分布”——讨论热度、板块聚焦程度与多空并存度；三条信号在 2024–2026 年上半年的保留段内复现了探索段强度的 80%–95%，方向均为负，即讨论拥挤的股票与板块在次周跑输。与之相对，观点的内容——情绪方向、情绪强度、交易自述、个体历史战绩——均不具有预测力。本文还报告两组有方法意义的负结果：三个原始 $|t|$ 值达 3.9–5.1 的候选信号经风格回归后信息含量衰减 82%–92%，本质上是换手率与短期动量的冗余表达；“散户恐慌预示底部”的流行假设被行为与情绪两套独立指标一致证伪，A 股散户的恐慌确认下跌的延续而非底部。上述发现对基于文本的因子研究具有一般含义：缺少风格正交与严格样本外检验时，情绪类因子容易反复发现流动性风格的既有信息。

关键词：投资者注意力，散户情绪，社交媒体文本，大语言模型，横截面收益，预注册

1 引言

中国 A 股市场的成交主体长期以个人投资者为主，散户在网络社区中的发言因此成为观察其行为的一道独特窗口。本文回答两个层层递进的问题：第一，散户的网络讨论中是否含有对未来横截面收益的预测信息；第二，若有，信息究竟存在于讨论的**内容**——散户说了什么、看多还是看空、情绪几何——还是存在于讨论的**分布**——多少人在说、围绕哪些股票与板块说、分歧有多大。

既有文献对前一问题多持肯定态度。注意力驱动的买入会推高受关注股票的价格、压低其次期收益^[1]，个人投资者的集中交易本身就足以在短期内推动价格^[2]；卖空约束下的意见分歧导致高估，分歧越大、后续收益越低^[3,4,5]。论坛与留言板文本被反复证明含有可交易的信息^[6,7,8]，媒体内容对价格具有因果影响^[9]，搜索量则成为注意力的标准代理^[10]。然而这一文献传统中始终存在一个未决的质疑：文本因子看似稳健的显著性，有多少只是市值、换手率与短期动量的

影子？因子发掘中的多重检验与发表后的衰减已是常识性问题^[11,12,13]，不经风格回归与严格样本外检验，上述质疑无从回答。

本文所用的语料为回答这两个问题提供了可能。NGA 是国内最大的游戏社区，其股票板块聚集了大量年轻的互联网原生用户；该板块的通用语体是反讽——大跌被“开香槟”迎接，指数原地踏步被写作“大 A 牛逼，又双叒叕回到 3000 点了”。词表方法在此类语料上结构性失效，这很可能正是该板块从未进入因子研究视野的原因。本文对该板块进行了 2016–2026 年的楼层级全量留档，滚动采集使删帖一经抓取即被保留；并以大语言模型完成 609 万帖的结构化标注^[14,15]。标注管道的关键设计是**闭集实体链接**：模型只能从本地词典（含人工审核的社区黑话，如“海狗”对应中远海控）给出的候选中选择立场标的，每条立场因而天然携带交易所代码，截面维度由此得以建立。

在这份语料之上，本文执行一套以代码而非自律来强制的事前研究设计（图 4）：每个假设族先预注册再运行，冻结信号定义、方向先验、验收门槛与禁扫清单；族内以 Benjamini–Hochberg 方法控制错误发现率；每个截面幸存者必须在剔除市值、换手、短期动量与行业暴露后仍然存活，衰减超过 70% 即判定为风格冗余；2024 年 1 月至 2026 年 6 月的收益端数据由守卫代码封存，任何访问抛出异常，每个最终候选仅允许凭预算台账查询一次；2026 年 7 月起，每周影子账本自动记录信号取值，积累完全无偏的前向证据。

本文的发现可以概括为一句话：**散户的注意力有信息，散户的观点没有**。三条信号通过了全部检验，且均为注意力的度量、方向均为负：个股周度讨论热度（探索段 rank IC -0.032 ，保留段 -0.031 ）、多空分歧（ -0.023 ， -0.018 ）、板块热度轮动（ -0.060 ， -0.054 ，为全库最强信号）。而关于“说了什么”的一切维度均未通过：净情绪、情绪拐点、恐慌与亢奋强度、反讽浓度、交易自述与观点之差、用户历史战绩、新面孔占比，均无增量信息。两组负结果与幸存者同样重要。其一，三个在常规筛选流程下必然入选的信号——买入自述拥挤（ $t = -4.2$ ）、冷启动热度爆发（ $t = -5.1$ ）、亢奋浓度（ $t = -3.9$ ）——经风格回归后信息含量归零，它们是流动性风格的冗余表达；其二，“散户割肉标记情绪底”的流行假设被行为自述与情绪强度两套独立指标分别证伪：高“投降”比例的股票在随后继续下跌，散户恐慌在 A 股是下跌的确认信号而非反转信号。

本文的贡献有三。第一，数据资产：一份十年期、楼层级、滚动留档的散户论坛档案及 609 万条结构化标注，立场与交易所代码逐一链接（第 2 节）。第二，发现：为“信息在注意力而不在内容”这一命题提供了经严格样本外验证的证据，并显示该效应在板块尺度上进一步放大（第 4–6 节）。第三，方法含义：三个“表面显著、实为风格冗余”的完整标本，以及一套可移植的文本因子检验规程（第 5 节）。第 7 节讨论机制解释与局限，其中包括本文唯一未决的观察——三条幸存信号在 2026 年上半年同时走平，其原因本文刻意未在保留段内追查。

2 数据

2.1 论坛语料与档案建设

NGA 股票板块（版面编号 706）是游戏社区内的 A 股综合讨论区。采集器以页为单位归档，依次尝试移动端接口、网页解析与无头浏览器三级回退，记录每一层楼，并逐页登记归档状态以

支持断点续跑与事后审计。表 1 汇总语料概况，图 1 给出规模与覆盖的时间形态。2020 年以前的覆盖主要来自少数历经多年删帖仍存活的长寿集中帖（社区称“水楼”，本文称长寿集中帖），不能代表当年论坛总体，故全部截面研究自 2020 年 1 月开始。此点声明一次，后文不再重复。

表 1: 语料与标注概况

档案窗口	2016-01 – 2026-07（滚动增量）
帖子（楼层级）	6,469,601
主题	341,629
独立用户	128,148
送标帖数	6,263,930
有效标注（schema 通过）	6,088,848（97.2%）
与档案日期关联丢失	0
展开后实体链接立场	5,193,986
含投资相关立场的帖子	71.2%
言语行为（观点/提问/交易/玩梗/新闻）	69.6 / 12.0 / 9.8 / 6.6 / 2.0 %
反讽标记	19.7%
恐慌 ≥ 1 （0–3 级）	24.6%
亢奋 ≥ 1 （0–3 级）	27.5%

注：标注计数来自 v2 标注库；比例按全部有效标注帖计算；立场计数为帖内立场数组展开所得。数据来源见复现包中的档案审计与特征集市审计。

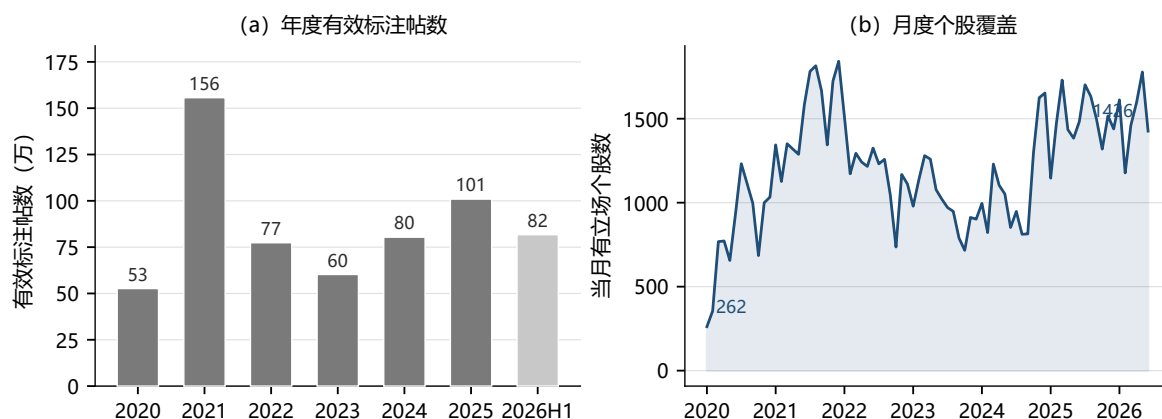


图 1: 语料规模与个股覆盖。(a) 年度有效标注帖数；(b) 当月获得至少一条个股立场的股票数。覆盖由 2020 年初的 262 只升至 2021 年末的逾 1,800 只，2026 年 6 月为 1,426 只。2016–2019 年仅覆盖少数长寿集中帖，不具代表性，未列入研究窗口。数据：特征集市 V3。

2.2 大语言模型标注与闭集实体链接

每帖连同主题标题、被引用的父帖（多数帖子为回复，语境不可或缺）以及由本地词典匹配出的**候选标的列表**一并送入标注模型。模型逐帖返回：相关性标记；立场数组，每条含标的类型（大盘、板块、个股、自身持仓、未知）、**必须取自候选列表**的标的代码、维度（基本面、

价格、资金面)、方向 (-2 至 +2) 与期限; 言语行为 (观点、提问、交易自述、玩梗、新闻转发); 恐慌与亢奋强度 (0-3); 反讽标记; 置信度。每个批次做机械校验 (返回条数、post_id 回显、schema 合法性、候选闭集), 任一失败整包重试; 97.2% 的帖子产出 schema 合规的标注。主力标注模型为一款紧凑型商用模型 (占 84.6%), 其余由四个备用模型在限流时段分担; 标注提示词、完整 schema 与校验规则见复现包。管道结构如图 2 所示, 图 3 给出一条真实帖子的送标输入与模型输出。

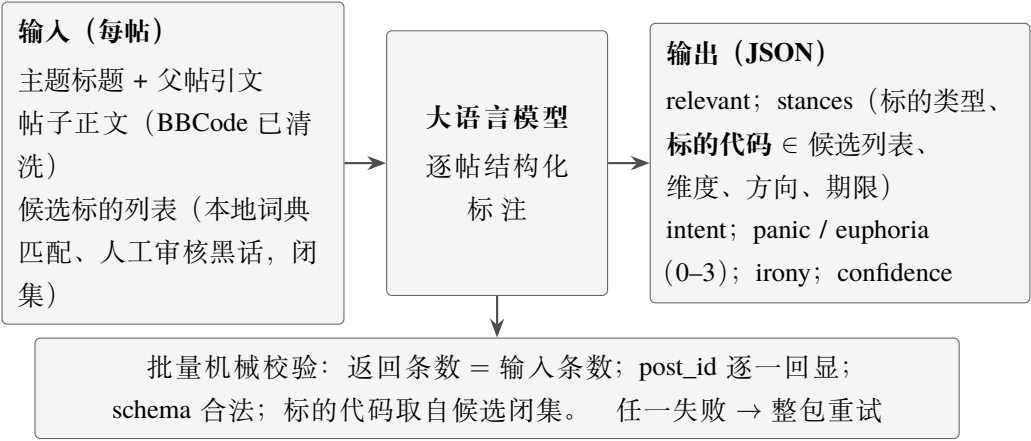


图 2: 标注管道。模型不得凭记忆填写标的代码，实体链接完全经由候选闭集完成；这一设计是截面维度得以建立的前提。



图 3: 真实标注的一条输入与输出（取自标注库，post_id 已匿名化）。候选列表为空时，模型只能将标的类型记为 unknown，无法杜撰个股代码——闭集规则的直接体现；“香槟滞销”被按反讽的真实意图判为负向。

两个设计值得强调。第一，反讽按**真实意图**而非表面方向标注：“大 A 牛逼，又回到 3000

EROS 量化实验室

4

点”标为负向且反讽为真。在这一社区，阴阳怪气是表达失望的默认语体，词表会把同一句话读成看多。第二，闭集候选意味着模型无法凭记忆杜撰代码；黑话只能经人工审核的词典进入候选。由此得到的“立场-代码”链接是整个数据集截面维度的根基。表 2 给出四条真实帖子的标注示例，分别对应反讽识别、黑话实体链接、交易自述与一帖多立场四种典型情形。

表 2: 真实标注示例（原文摘自标注库，BBCode 已清洗，仅作节选）

日期	帖子原文（节选）	关键标注输出	典型情形
2021-06-22	香槟滞销，救救孩子	方向 -1；恐慌 = 2；反讽 = 真；置信度 85	反讽识别：词表读不出真
2025-09-26	多看看海狗贴就知道了，现在我看到海狗涨停了的帖子不用点开软件就知道海狗涨了 1 个点	标的 = 601919.SH（中远海 控）；言语行为 = 玩梗；反讽 = 真	黑话实体链接
2026-03-03	我持股海狗 5 年多了，作为最忠诚的藤壶今天都大幅减仓了，只留了个底仓	标的 = 601919.SH；方向 -1；言语行为 = 交易自述	交易自述
2025-09-26	长期看好啊，但是楼主昨天卖了一半	标的 = 601869.SH（长飞光 纤）；立场一：基本面 +1、中长期；立场二：价格 -1、短期；言语行为 = 交易自述	一帖多立场（言与行的分离）

注：示例按帖子主键从标注库检索所得，未作任何筛选性修饰。“海狗”“藤壶”为中远海控股股东的社区自称，映射经人工审核词典进入候选列表。

2.3 收益数据、样本池与变量构造

收益端遵循标准周度协议。信号按 ISO 周聚合；个股当周须获得至少 3 条立场，且属于时点主板样本池（上市满 60 个交易日、20 日日均成交额不低于 2000 万元、剔除 ST 与停牌）；达标股票不足 50 只的周予以跳过。前瞻收益定义为信号周最后一个交易日起 +5 个交易日的收盘对收盘收益。探索段周达标池中位数为 149 只，保留段为 186 只。板块信号使用 18 个论坛板块名到交易所行业分类的事前冻结映射；“科技”“AI”“新能源”等宽叙事标签（约占板块立场的三分之一）没有干净的行业对应，直接放弃并如实记录覆盖损失，不做映射调整。

2.4 数据偏差及其消解位置

三类偏差影响历史段估计，本文逐一披露并给出消解位置。其一，2020–2023 年收益端数据此前已被其他研究流程使用，探索段结论的证据等级相应下调；其二，标注模型的训练数据覆盖样本期，其对 2020–2023 年帖子的解读可能带有事后视角——程度有限，因为模型只见文本、从不见收益；其三，采集开始之前的历史帖子是多年删改后的幸存样本。保留段（2024–2026H1）消除第一类偏差；前向段（2026 年 7 月起实时采集、实时标注）三类偏差全部消除。这一分级去偏构成本文证据定级的骨架。

3 研究设计

3.1 样本分区与检验预算

全部假设先在探索段（2020–2023）检验。保留段（2024-01–2026-06）由代码强制封存：任何终点触及该区间的数据访问都会抛出异常，除非环境确认变量与追加式预算台账中的预记账条目匹配；初始预算 6 次，每个最终候选仅允许查询一次，消耗不可逆。前向段（2026-07 起）由每周影子账本写入，只记录信号取值与前 20% 名单，不做任何交易假设；账本从不读取行情，在构造上即无前视偏差。整体流程见图 4。

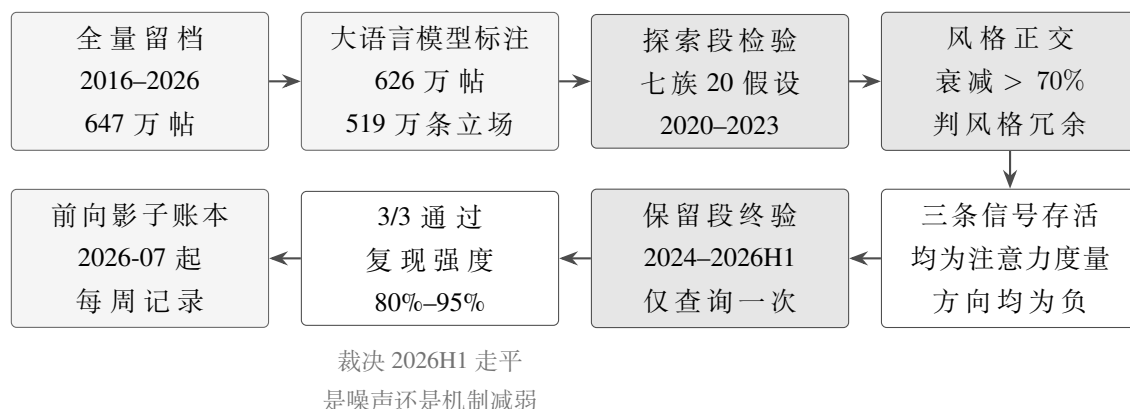


图 4: 研究流程。每一环节均有对应的事前约束：假设族预注册、族内错误发现率控制、风格正交规则、保留段预算台账与前向账本。

3.2 预注册、验收门槛与多重检验控制

每个假设族在任何代码运行之前完成预注册，冻结信号定义、带机制说明的方向先验、验收门槛与禁扫清单。禁扫清单针对本数据最大的扫参诱惑设立：聚合窗口、阈值、维度与期限子集一律禁止事后枚举。验收门槛为：C1，周均 rank IC 满足 $|IC| \geq 0.010$ 且符号与注册一致；C2，信号最高五分位相对池均的周均超额不小于 $|5| \text{ bp}$ 且方向一致；C3，探索段四个年度中至少三年 IC 符号一致。Newey–West t 值（4 阶滞后）进入族内 Benjamini–Hochberg 控制（ $q = 0.10$ ）。符号与注册相反即判定不通过，无论幅度；看过数据后翻转符号按数据挖掘论处，一律禁止。全部主检验预算上限 24 次，实际消耗 20 次。唯一一次结构化方向预检（一个没有可信方向先验的板块溢出假设）本身亦经预注册：方向在 2020-01–2021-06 上固定，评估仅在 2021-07–2023-12 上进行，两段不重叠。

3.3 风格正交规则

散户讨论天然集中于小市值、近期上涨、换手放大的股票，任何注意力度量生来与公开风格相关。每个截面幸存者须在逐周横截面回归

$$\text{signal} \sim \ln(\text{流通市值}) + \ln(1 + \text{周换手}) + \text{前周收益} + \text{行业哑元} \quad (1)$$

的残差上重新估计 IC。若正交化摧毁超过 70% 的 $|IC|$ ，判定为 **风格冗余**、不予通过——无论 t 值多高，它只是公开风格的重复度量。另设第二道冗余规则：与纯热度的周均秩相关超过 0.70

的候选，判定为热度的构造性变体，不再计入独立信号。

4 主要结果：注意力信号的样本外验证

表 3: 三条幸存信号：探索段、风格正交与保留段终验

信号	探索段 2020–2023			风格正交	保留段 2024–2026H1	
	IC	前 20%	年度	正交 IC (衰减)	IC	前 20%
个股热度 HEAT	−0.032	−6.7 bp	4/4	−0.011 (66%)	−0.031	−22.5 bp
多空分歧 DISAGREE	−0.023	−12.8 bp	4/4	−0.010 (57%)	−0.018	−28.0 bp
板块热度 SECTOR-HEAT	−0.060	−28.9 bp ^a	4/4	— ^b	−0.054	−20.7 bp ^a

注：周度 rank IC，对 +5 交易日前瞻收益；“前 20%”为信号最高五分位相对池均的周均超额。探索段 191–203 周；保留段 126–127 周，各候选凭预算台账查询一次。三条信号保留段裁决均为通过：符号正确、 $|IC|$ 不低于探索段基准之半、子段符号 2/3 一致。

^a 板块信号为热度前 1/3 减后 1/3 板块组合的周均价差。

^b 板块级信号不适用个股风格回归；替代冗余检查（信号与全市场换手变化的时序相关）通过。

4.1 信号定义与探索段证据

个股热度 (HEAT) 为个股周立场数的 $\ln(1 + n)$ ；**多空分歧 (DISAGREE)** 定义为看多与看空立场数中较小者占两者之和的比例，度量多空双方同时在场的强度；**板块热度 (SECTOR-HEAT)** 为映射板块的周立场数 $\ln(1 + n)$ ，对板块成员次周等权收益评估。三者的注册方向均为负，即注意力拥挤；三者全部兑现（表 3）：讨论最多的股票次周跑输，与 Barber 和 Odean 的注意力效应一致^[1]；分歧最大的股票跑输更多，与 Miller 的分歧–高估机制一致^[3,5]；而最热的**板块**跑输最甚。板块结果是全库最强信号（IC −0.060；热减冷板块价差 −28.9 bp/周），与散户资金以叙事为单位轮动的观察相符：拥挤在故事层面比在代码层面完成得更彻底。一个同族的未通过假设从反面印证了同一点：热板块内**未被论坛点名的**成员股，仍较冷板块同类股每周跑输 25.8 bp（方向由事前注册的方向预检确定；该信号年度稳定性 2/3 未达门槛，按规则判定不通过）——注意力的影响半径覆盖整个板块。

4.2 风格正交检验

风格规则（第 3.3 节）将两条个股信号的强度削去约一半，但未能将其消除：HEAT 的正交 IC 为 −0.011，四个年度符号全部保持；DISAGREE 为 −0.010 (3/4)。两条信号对换手 ($\rho \approx +0.10$) 与市值 (HEAT 为 +0.14) 的原始暴露并不高，衰减主要来自与行业哑元的联合回归。本文的解读是：原始效应约一半是用换手率即可度量的拥挤，另一半则来自“谁在关注”这一维度的独立信息；后者正是风格中性组合能够变现的部分。

4.3 保留段终验

三条信号连同冻结定义与事前固定的通过阈值（符号正确； $|IC|$ 不低于探索段之半；子段符号 2/3 一致）一次性提交给代码封存的 2024-01 – 2026-06 保留段，消耗六次预算中的三次。三条全部通过，强度为探索段的 80%–95%（表 3、表 13）；SECTOR-HEAT 的 2025 年 IC（ -0.079 ）甚至高于其探索段均值。图 5 给出年度纹理，图 6 给出 HEAT 的周度时序：效应并非由个别极端周驱动，滚动均值在整个样本期稳定位于零以下。

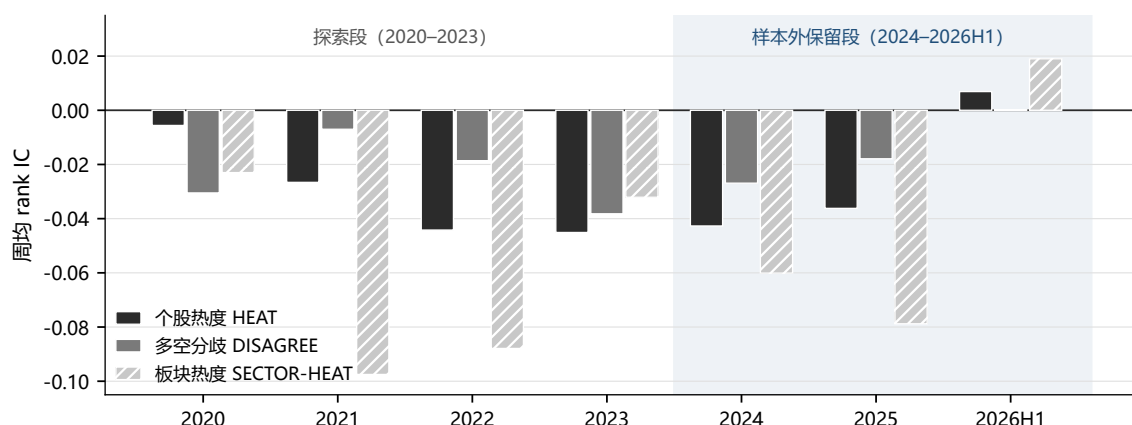


图 5：三条幸存信号的年度周均 rank IC。浅色区为代码封存的样本外保留段（2024–2026H1），仅查询一次。三条信号在 2024–2025 年复现，2026H1 同时走平——本文唯一未决观察，交由前向影子账本裁决（第 4.4 节）。

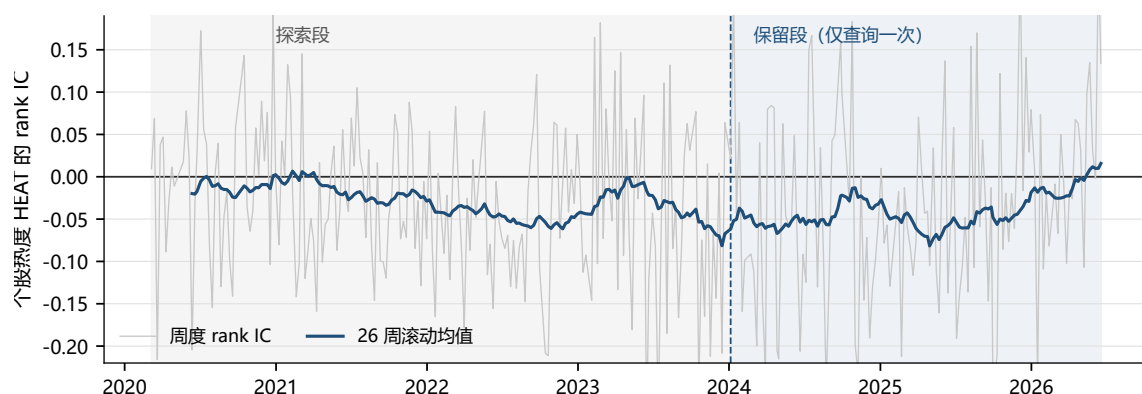


图 6：个股热度 HEAT 的周度 rank IC 与 26 周滚动均值，2020-01 – 2026-06，共 317 周。虚线为探索段与保留段分界。

4.4 2026 年上半年的信号走平

比三项通过更值得如实记录的观察是：2026 年上半年，三条信号同时走平或微幅翻正（HEAT $+0.007$ ，DISAGREE $+0.0002$ ，SECTOR-HEAT $+0.019$ ）。三条机制相互独立的信号在同一个半年内集体失效，更像市场结构的变化而非随机噪声。候选解释有三种：约 26 周的短样本统计功效有限；注意力反向效应在偏动量的市况下真实减弱；社区语言演化带来的标注仪器漂移。在不二次触碰保留段（为规则所禁止）的前提下，三者无法区分。本文选择等待：前向账本自 2026

年7月起每周积累无偏证据，处置规则事前确定——若走平延续，三条信号转入“机制减弱”观察，不得进入生产使用；若负IC恢复，2026H1 归档为短样本噪声。

5 风格冗余的识别：三个反例

表 4: 被风格正交规则判为冗余的三个显著原始信号

信号	原始 IC	NW- <i>t</i>	正交 IC	衰减
买入自述拥挤 TA-CROWD	-0.029	-4.16	+0.002	92%
冷启动热度爆发 HEAT-BURST	-0.047	-5.07	+0.009	82%
个股亢奋浓度 EUPHORIA	-0.029	-3.86	+0.003	90%

注：探索段，周度截面，样本池与门槛同表 3。三者原始形态均通过 C1-C3 与族内错误发现率控制，在常规流程下都会被采纳；对市值、换手、动量、行业残差化后与统计零无异。

表 4 与图 7 报告本文认为最具方法警示意义的一组结果。三个信号——用户自述实际买入的帖数占比、本周热度相对近零基数的倍数（“突然走红”形态）、亢奋强度不低于 2 的立场占比——全部通过常规验收门槛，原始 *t* 值介于 3.9 与 5.1 之间，年度符号四年全对；经风格回归后，三者全部归于统计零。表面上的“散户集中买入后反转”，在机制上实为“换手激增且近期上涨的股票回落”，情绪维度没有贡献任何增量。其一般含义在于：任何以原始 IC 为唯一裁决标准、构建在与散户注意力相关语料上的文本因子，在通过这一回归之前都应被推定为流动性风格的重复度量。本文的研究设计若缺少这道规则，将被同样的假象误导三次。

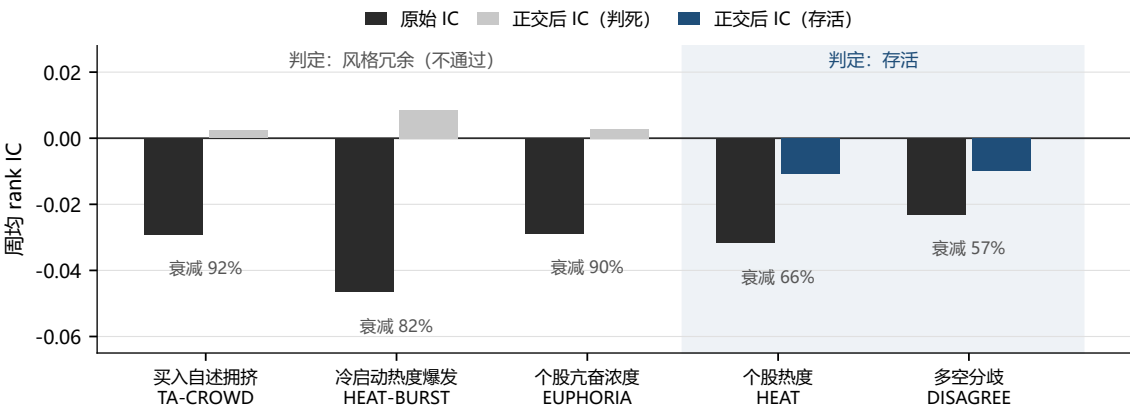


图 7: 风格正交前后的 IC 对照。左侧三个信号原始形态显著 ($|t|$ 3.9–5.1)，正交后归零，判定为风格冗余；右侧两条幸存信号强度约减半但保持存活。

6 观点内容的系统性负结果

其余各假设族的结局平淡，但失败的**模式**构成本文的第二项发现。分维度报告如下（完整统计见附录 B）。

情绪方向。净情绪没有截面信息 (IC +0.004)；情绪拐点 (本周均值相对四周基线的变化) 同样没有 (+0.004)。水平与转折两种构型均告失败。

情绪强度。个股恐慌浓度、反讽密度以及正交后的亢奋浓度，均无增量信息。

言与行。嘴上观点与自述交易之间的缺口无信息 (+0.008)：自述实际买卖的帖子并不比口头观点更有含金量。

个体能力。对每位用户以滚动 52 周方向命中率做贝叶斯收缩打分，构造“论坛中存在持续预测能力者”的组合信号。这是本研究工程量最大的一项检验 (4,679 名合格用户、180 周)，所得 IC 为 -0.0001：用户能力没有任何可识别的延续性，论坛中不存在可跟踪的聪明资金。

结构变量。个股新面孔占比在分布中段达到族内显著，但最高五分位反而**跑赢** (+11.7 bp, 注册方向为负)，单调性断裂，归为不可交易；提问密度的失败方式相同。作者广度与作者集中度与热度高度纠缠 (秩相关 +0.93 与 -0.84)，按冗余规则不再计为独立信号；但保留一条记注：广度的正交 IC (-0.0124) 略优于热度 (-0.0106)，若未来生产化，“多少个不同的人”是同一效应更干净的度量口径。

“散户投降与底部”假设的证伪。该假设认为，20 日回撤不超过 -10% 的深度回撤条件下，卖出自述占比高 (或恐慌强度不低于 2 的占比高) 应标记情绪底部、预示反转，注册方向为正。两套独立指标各自得到一次**负号**：行为版 IC -0.032 (年度符号 1/4, 2020 年条件 IC 低至 -0.21)，情绪版年度符号 0/4。换言之，散户“投降”比例高的股票随后继续下跌：在本市场，散户恐慌确认的是下跌的延续，人群在仍会继续的下跌中过早割让。按事前纪律，符号相反即判定不通过、不得翻转重测；但机制记注予以保留，“散户投降等于底部”一族永久关闭。

市场层面择时。词表版恐慌、亢奋、净情绪与长寿集中帖份额四个构型，大语言模型聚合情绪，新开户涌入，以及长寿集中帖份额的动量确认构型——累计八个市场级构型无一通过检验。论坛总体情绪在本文允许自己使用的任何口径下都不具备对指数的择时能力，这与既有文献中聚合情绪的时序预测力有限且不稳定的证据一致^[16,17]。

7 讨论

7.1 为什么注意力有信息而内容没有

噪声交易者框架给出了近乎工整的解释^[18]：个体方向观点数量巨大且相互抵消——71% 的帖子带有立场，其净和一无所有；但**观点持有者的聚集**本身即是市场事件。一只股票上的人群预示短期供需扭曲及其后的回落，与人群的信念内容无关；双向同时在场的人群 (多空分歧) 预示更强，与卖空约束下的分歧高估一致；聚集在同一个**叙事**上的人群 (板块热度) 则是最彻底的拥挤。三条信号一致为负——不存在任何“跟随论坛发言获利”的构型——本身就是对散户观点内容质量最干净的概括，也与个人投资者整体难以战胜市场的证据相符^[19]。

7.2 对文本因子研究方法的启示

第一，仪器的必要性。本文全部幸存者都产自 LLM 标注；同一语料上唯一跑过的词表族四个构型无一通过。在以反讽为默认语体、标的以黑话指称的社区里，词表方法在构造上就是失明的——这与财经文本分析中通用词表会系统性误判语义的结论一致^[20]。第二，标注预算的投

向。标注层的价值最终集中在一个字段：闭集实体链接。情绪字段——成本最高的部分——没有贡献任何可交易信息。为因子研究配置 LLM 标注的团队，或许应当倒转常规配比：把预算投向实体链接与校验，而非日益精细的情绪分类^[21]。第三，检验规程的必要性。三个风格冗余标本表明，缺少风格回归与样本外封存时，基于散户语料的因子研究很容易反复发现换手率；本文的协议若缺这道规则，亦不能幸免。

7.3 适用边界

本文的对照组合（盈利收益率前 50 的深度价值组合）在论坛中的周覆盖率仅 6.8%：价值股几乎无人讨论，注意力信号无法过滤那个宇宙。本文将这一不重叠解读为价值端“被忽视溢价”的旁证，同时也划定了注意力因子的适用边界：它们是高关注题材股宇宙的过滤器与空头候选来源，而非深度价值策略的叠加层。由于信号方向为负，直接变现需要采取回避（宇宙过滤）或在 A 股微观结构下成本较高的空头腿，本文不对费后收益作任何断言。

7.4 局限

单一论坛、单一市场、周度频率，以及中位 149 只的周达标池所隐含的有限容量。历史段的三类偏差已于第 2.4 节披露，保留段消除了其中的数据消耗层，完全去偏有赖前向积累。2026 年上半年的信号走平仍然悬置（第 4.4 节）；本文认为，事前约定的等待是唯一站得住的处置，并如实注记于此。

8 结论

六百万条标注、二十个预注册假设、一段代码封存的保留期，支撑一句话的发现：十年的中国散户论坛讨论中，一切可交易的信息都在注意力的分布里——哪些股票、哪些板块、争论多激烈；关于观点内容的一切都是噪声。对文献更有普遍意义的教训是方法上的：若没有风格正交规则与代码强制的样本外保留段，本文将发表三个实为流动性风格的“情绪因子”，并就此收工。

A 完整假设登记表

表 5: 全部预注册检验（按执行顺序）。主检验预算上限 24，实际消耗 20；QUARREL 因仪器缺失退还预算，后按机制判断永久关闭。

轮次	假设族 / 信号	注册符号	裁决
IT17	词表择时：恐慌 / 亢奋 / 净情绪 / 长寿集中帖 份额	+/-/-/-	4 条均未通过
IT18	LLM 截面：热度 / 净情绪 / 多空分歧	-/-/-	通过 / 未通过 / 通过
F8	IT18 幸存者风格正交审计	(审计)	双双存活
IT19	买入拥挤 / 割肉反转 / 言行缺口	-/+/-	风格冗余 / 方向相反 / 无信息
IT20	作者广度 / 集中度 / 新面孔 / 新开户	-/+/-/-	2 条热度冗余，2 条未通过
	热度爆发 / 慢性热度 / 情绪拐点	-/+/-	风格冗余 / 未通过 / 未通过
	亢奋 / 恐慌反转 / 反讽	-/+/-	风格冗余 / 方向相反 / 未通过
IT21	板块热度 / 板块溢出 / 集中帖动量	-/预检/+	通过 / 未通过 / 未通过
IT22	用户战绩 / 多空对喷 / 提问 / 自发热度	+/-/-/-	未通过 / 无仪器 / 未通过 / 冗余
IT23	保留段终验：HEAT、DISAGREE、SECTOR- HEAT	-/-/-	3 条通过

注：“风格冗余”指被第 3.3 节规则判定（衰减 > 70%）；“热度冗余”指与 HEAT 周秩相关超过 0.70；“方向相反”指与注册方向相反，判定不通过且不翻转重测，机制注记保留。完整预注册、冻结门槛、禁扫清单、逐周统计与保留段预算台账见复现包。

B 各轮详细试验记录

以下各表为每轮探测的完整统计，全部来自复现包随附的各轮结果（原始 JSON/CSV），未做任何事后筛选。NW- t 为 4 阶滞后 Newey–West t 值；“热度相关”为信号与 HEAT 的周均秩相关；正交衰减按第 3.3 节流程计算。

表 6: IT17：词表版市场时序探测（周频， $n = 203$ 周，全部未通过）

信号	ρ （预期符号）	Q5–Q1	年度符号	裁决
PANIC	-0.002（+）	-26 bp	2/4	未通过（无信息）
EUPHORIA	-0.101（-）	-91 bp	2/4	未通过（稳定性）
NET_SENT	-0.078（-）	+15 bp（方向错）	4/4	未通过（强度与单调性）
MEGA_SHARE	+0.131（预期 -）	+50 bp	1/4	未通过（方向相反）

表 7: IT18: LLM 标注版个股截面首轮探测 (2020–2023, 周池中位 149)

信号	IC	最热 20% 超额	年度	裁决
NGA_HEAT	−0.0316	−6.7 bp	4/4	通过
NGA_SENT	+0.004	+35.9 bp (方向反)	2/4	未通过
NGA_DISAGREE	−0.0230	−12.8 bp	4/4	通过

表 8: F8: IT18 幸存者的风格正交审计 (191 周)

信号	原始 IC	正交 IC	衰减	NW- <i>t</i> (正交)	正交年度	换手暴露	市值暴露	动量暴露
NGA_HEAT	−0.0316	−0.0106	66%	−1.86	4/4	+0.10	+0.14	+0.02
NGA_DISAGREE	−0.0230	−0.0098	57%	−1.55	3/4	+0.10	+0.03	+0.00

表 9: IT19: 交易行为族 + 用户结构族 (7 检验)

信号	IC	NW- <i>t</i>	前 20%	年度	热度相关	正交衰减	裁决
TA_CROWD	−0.0291	−4.16	−26.3 bp	4/4	+0.54	92%	风格冗余
TA_CAPITULATION	−0.0315 (预期 +)	−1.17	−63.4 bp	1/4	+0.36	—	方向相反
TALK_ACT_GAP	+0.0083	+1.15	+36.9 bp	1/4	−0.01	—	无信息
NGA_BREADTH	−0.0279	−4.34	−7.6 bp	4/4	+0.93	55%	热度冗余
AUTHOR_HHI	+0.0239	+3.67	+8.9 bp	4/4	−0.84	48%	热度冗余
NEWCOMER_SHARE	−0.0157	−2.45	+11.7 bp	3/4	−0.21	—	单调性断裂
NEW_ACCOUNT_INFLUX	$\rho = +0.084$ (预期 −)	—	+60 bp	3/4	—	—	方向相反

注: NEWCOMER_SHARE 的前 20% 为正 (C2 方向不符); NEW_ACCOUNT_INFLUX 为市场时序信号, 列值为 Q5–Q1。

表 10: IT20: 动态形态族 + 情绪极端族 (6 检验, 全部未通过)

信号	IC	NW- <i>t</i>	前 20%	年度	热度相关	正交衰减	裁决
HEAT_BURST	−0.0466	−5.07	−15.6 bp	4/4	+0.81	82%	风格冗余
HEAT_PERSIST	+0.0091	+1.32	+14.8 bp	3/4	+0.36	—	强度不足
SENT_REVERSAL	+0.0040	+0.56	+11.3 bp	1/4	−0.09	—	无信息
STOCK_EUPHORIA	−0.0288	−3.86	−8.7 bp	4/4	+0.25	90%	风格冗余
STOCK_PANIC_REV	−0.0091 (预期 +)	−0.50	+16.1 bp	0/4	+0.32	—	方向相反
IRONY_DENSITY	−0.0062	−0.97	−1.4 bp	3/4	+0.34	—	强度不足

表 11: IT21: 板块族 + 长寿集中帖重开 (3 检验, 含 1 次注册内方向预检)

信号	统计	年度	裁决
SECTOR_HEAT_ROT	IC -0.0603 , NW- $t - 3.33$, 热 - 冷价差 -28.9 bp/周	4/4	通过
SECTOR_SPILLOVER	预检定方向 -; 评估段 -25.8 bp, $t = -1.43$	2/3	稳定性不足
MEGA_SHARE_MOM	$\rho = +0.139$, Q5-Q1 $+54$ bp (方向正确)	2/4	稳定性不足

注: MEGA_SHARE 两个方向 (IT17 反向、本轮正向) 各检验一次, 均因年度稳定性不足未通过, 该线永久关闭。

表 12: IT22: 用户战绩 + 话语结构族 (4 检验)

信号	IC	NW- t	前 20%	年度	热度相关	裁决
USER_ALPHA	-0.0001	-0.01	-8.1 bp	2/4	-0.08	无信息 (战绩无延续性)
QUARREL	—	—	—	—	—	仪器缺失, 后按机制关闭
QUESTION_RATIO	-0.0135	-2.57	$+1.1$ bp (C2 反)	4/4	$+0.09$	单调性断裂
ORGANIC_HEAT	-0.0299	-4.44	-8.2 bp	3/4	$+0.98$	热度冗余 (构造性)

C 保留段终验明细

表 13 给出三条候选在保留段的分段 IC 与裁决; 表 14 给出探索段年度 IC (即图 5 的数据)。预算消耗与台账见附录 D。

表 13: IT23 保留段终验: 分段 IC 明细 (2024-01 - 2026-06, 126-127 周)

候选	探索基准	保留段全期	2024	2025	2026H1	裁决
NGA_HEAT_v1	-0.032	-0.0310	-0.0427	-0.0362	$+0.0069$	通过
NGA_DISAGREE_v1	-0.023	-0.0183	-0.0269	-0.0179	$+0.0002$	通过
SECTOR_HEAT_ROT_v1	-0.060	-0.0535	-0.0601	-0.0788	$+0.0190$	通过

注: 预算消耗 3/6, 剩余 3; 预注册文件 SHA-256 与逐候选记账见台账。通过阈值: 符号正确、 $|IC|$ 不低于探索基准之半、子段符号 2/3。2026H1 三条集体走平为本文唯一未决观察 (正文第 4.4 节)。

表 14: 三条幸存信号的探索段年度 IC (图 5 数据)

信号	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

D 研究纪律执行记录

- 检验预算: 主检验上限 24, 消耗 20 (IT17: 4, IT18: 3, IT19: 7, IT20: 6, IT21: 3, IT22: 4, 减 QUARREL 退还 1; 另有 IT18 前置 3 条按口径不计入本计划预算, 详见试验台账逐

行记录)。

- **方向预检登记**：全程 1 次 (IT21 板块溢出，预检段 2020-01 – 2021-06 仅定方向，评估段 2021-07 – 2023-12，两段不重叠)。
- **保留段预算**：初始 6 次；IT23 消耗 3 次 (三候选各一)，剩余 3 次；台账为追加式 CSV，含预注册文件 SHA-256、逐候选记账与裁决回填。
- **符号纪律**：四例方向相反者 (TA_CAPITULATION、STOCK_PANIC_REV、NEW_ACCOUNT_INFLUX、MEGA_SHARE (IT17 方向)) 一律判定不通过、不翻转重测；其中“投降”类两例的机制标记 (散户恐慌确认下跌延续) 以文字保留。
- **永久关闭清单**：散户投降等于底部 (双指标证伪)；市场层面择时 (八个构型无一通过)；个体用户战绩；长寿集中帖份额双向；QUARREL (机制判断：本社区股票版“空谈无成本”文化下多空对喷不成立)。
- **前向影子账本**：2026-07 起每周记录三条幸存信号的全池取值与前 20% 名单 (个股轨与板块轨)，脚本从不读取行情，构造上无前视偏差。

E 复现说明

复现包 (随文在线附件) 包含：标注提示词与 schema、14 份预注册/结果文档、8 个探测与终验脚本、每个检验的逐周统计、试验台账 (多重检验计数) 与保留段预算台账。原始档案 (17.8 GB) 与标注库 (1.3 GB) 可向作者索取；前向影子账本自 2026-07 起每周累积。特征集市由 build_mart_v3.py 从两个 SQLite 库重建 (626 万条标注展开为 519 万条立场，日期关联零丢失)；本文全部数据图由 make_figs_v2.py 生成。

参考文献

- [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.