Talking About Stocks: Pathways into Financial Engagement

Olga Balakina* and Jonathan Seim†

September 2026

Abstract

Before anyone buys a stock, they start paying attention. Following 6.8 million Reddit users, we find distinct routes into stock-market discussion, visible a year before entry, that open under different conditions. When policy uncertainty is high, a larger share of entrants arrive from politics and news and no larger share from gambling or crypto, and those arriving during a major event write about it in their first post. Composition relates to market reactions only through which securities draw attention: entrants from gambling forums choose more volatile stocks, but the same company reacts alike whoever discusses it.

Keywords: attention, stock-market participation, uncertainty, social finance, Reddit, online communities

*University of Bonn. Email: obalakin@uni-bonn.de

†University of Cologne. Email: jonathan.seim@wiso.uni-koeln.de. Jonathan Seim acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866.

1 Introduction

Household-finance research explains differences in stock-market participation through participation costs, information frictions, background risk, ambiguity, trust, financial literacy, advice, and social interactions (Haliassos and Bertaut, 1995; Vissing-Jørgensen, 2002; Campbell, 2006; Guiso et al., 2008; Guiso and Jappelli, 2005; van Rooij et al., 2011). That work begins with people who already hold accounts or make investment decisions, and it explains why they differ from one another. Before anyone opens a brokerage account or buys a stock, however, they must first begin paying attention to financial markets, seeking information, weighing potential investments, and talking with others about what they might do.

Yet, far less is known about this earlier stage, and in particular about who starts paying attention and when. Surveys and administrative records capture decisions rather than the process that leads to them, leaving the formation of financial interest largely outside the evidence base. The gap matters for how episodes of heightened retail attention are read, since the people who begin following markets during a policy crisis need not resemble those who begin during a speculative mania, and a surge in attention says nothing on its own about who is behind it.

To address this gap, we exploit the structure of Reddit, where the same accounts participate across thousands of topic communities. We use a public archive covering roughly 40,000 subreddits from June 2005 through July 2026 and identify 6.8 million users whose first contribution to a stock-market subreddit falls between January 2015 and July 2026, reconstructing for each what they discussed in the year before they first said anything about stocks. We define entry as that first contribution. It is not a measure of stock ownership, since we do not observe whether a contributor invests or trades. Writing a post or comment is nonetheless an active choice, placing public discussion between passive attention, as measured by searches or browsing, and financial action, as measured by trading or portfolio choice (Barber and Odean, 2008; Da et al., 2011).

Two boundaries on this evidence should be clear from the outset. Reddit users are not a representative sample of households, and the platform's own composition shifted over our sample, most visibly around the GameStop episode, making our findings a description of how interest in markets forms among people who discuss them in public online rather than among households in general. The timing we study is not peculiar to the platform, however, since the months of unusually high entry are also the months of unusually high trading and account opening at two large retail brokerages.[1] Moreover, a pre-entry route can be read

[1]Appendix I compares monthly entry with Interactive Brokers' daily average revenue trades and Charles Schwab's new brokerage accounts. The correlations of the detrended series are 0.69 and 0.70, and 0.62 and 0.55 excluding the GameStop months. This speaks to when entry happens, not to whether our users resemble

only for users with enough prior activity. Just over a quarter of entrants receive one of our types, a further fifth have a legible history whose dominant interest lies elsewhere, and the remainder post too little beforehand to classify.[2]

We use these histories to address three related questions. Which interests lead users into stock-market discussion? Does the mix of users who enter change with economic and market conditions? And does it matter for markets which kinds of users enter, or only which securities they attend to? Ex ante, it is unclear in which direction conditions should move the mix. On the one hand, volatile and uncertain periods are when speculative bets are most visible, which could tilt entry toward users arriving from gambling and crypto. On the other hand, policy uncertainty is itself news, reaching first the users who already follow public events. The third question admits two answers that are equally hard to separate in advance, since the path into markets could shape how strongly a person responds to a given security, or shape only which securities that person attends to in the first place.

Our answers are as follows. Users arrive along distinct routes that are visible a year before they say anything about stocks, and those routes remain legible in what they write and in the securities they discuss. The mix of entrants changes with conditions. Notably, it does not change in the direction a story about speculation would suggest, since when policy uncertainty is high entry tilts toward users arriving from politics and news, no route associated with speculation gains, and users arriving during a major event write about that event in their first post. Importantly, the route a user took relates to market outcomes only through which securities draw attention, since we detect no difference in how the same company reacts whoever is discussing it.

We measure a user's route with the composition of their pre-entry activity, comparing each entrant's activity shares across subreddit categories during the twelve months before entry with the contemporaneous Reddit-wide shares and assigning the finance-adjacent category in which their standardized excess activity is largest. This yields nine types, *Politics and News*, *Personal Finance*, *Productivity and Career*, *Home and Housing*, *Crypto*, *Economics*, *Law and Legal*, *Non-Stock Investing*, and *Gambling*. Future entrants are unusually active in finance-adjacent topics well before their first stock-market contribution, and this concentration rises steadily as entry approaches, making entry the culmination of a broader shift in attention rather than an isolated first post. Relative to otherwise active users who have not yet entered, prior activity in *Non-Stock Investing*, *Economics*, *Personal Finance*, and *Crypto* predicts

retail investors in general.

[2]Among the 6,792,874 users entering between May 2015 and July 2026, 28.0 percent receive one of the nine types, 19.4 percent have at least ten posts in the year before entry concentrated outside the nine finance-adjacent categories, and 52.6 percent have fewer than ten. Section 4 treats the last group as an outcome in its own right rather than discarding it.

entry in the following month, while most non-financial categories carry small or negative associations.

The types persist once users arrive, appearing in the communities they enter through, in the motives their first posts express, and in the securities they go on to discuss. *Personal Finance* entrants concentrate on broad-market index funds, *Crypto* entrants on bitcoin proxies, and *Gambling* entrants on sportsbooks and casinos. Knowing which security a mention concerns resolves 27.5 times as much uncertainty about its author's type as chance would, a lower bound, since the types are assigned from pre-entry activity alone and misclassification can only weaken it.

Entry into stock-market discussion is highly uneven over time, surging most during the GameStop episode of early 2021 and less around the onset of the pandemic and the tariff announcements of 2025. To ask whether conditions change who arrives, we build a daily panel of first posts and relate it to the news-based index of economic policy uncertainty of Baker et al. (2016), to market-implied volatility, and to absolute market returns, separating two margins. The first is whether an entrant has a legible history at all. When uncertainty is high, entry rises both among users with no usable history and among users with one, and rises more among the latter, lowering the share of entrants without a usable history by close to a percentage point for a one standard deviation increase in the index. The second margin is the route among entrants whose history is legible. There, a one standard deviation increase in uncertainty is associated with a rise in the share arriving from politics and news of roughly a sixth of that category's mean share. The share arriving from budgeting and retirement communities falls by about half as much in percentage points, while no route associated with speculation gains.

A daily association between an index and the mix of arrivals cannot show that any individual arrived on account of the episode the index registers, so we also read what entrants write. We assemble twenty-two dated episodes spanning tariff announcements, Federal Reserve decisions, inflation releases, bank failures, elections and the pandemic. In the days following each, the rate at which first posts mention the episode's own topic rises by 93 percent, measured against the change in the other five topics on the very same days, a comparison that nets out whatever lifted attention to all topics together. A version built on the published decomposition of the uncertainty index into ten policy categories, which involves no dates chosen by us, points in the same direction. We nonetheless read the compositional results as associations rather than causal effects. The rise in the politics and news share survives dropping the pandemic year, the tariff period, or both, but it operates chiefly between months rather than within them, which is the frequency most exposed to slow-moving confounders.

We then ask what this heterogeneity means for markets. Identifying episodes in which a

security is discussed unusually heavily, and measuring returns and trading volume around them, we find that larger episodes go with larger absolute abnormal returns, higher abnormal volume, and mild subsequent reversal. Composition predicts these reactions across securities, with episodes dominated by *Gambling* entrants showing absolute abnormal returns larger than those dominated by *Personal Finance* entrants by about half the sample mean.

That gap, however, does not reflect a difference in how markets respond to a given company. Three observable security characteristics, return volatility above all, absorb about two thirds of it on their own. Comparing episodes within the same security shrinks it to about an eighth of its across-security size, with a confidence interval that excludes the across-security estimate and a design able to detect an effect less than half as large. The evidence therefore points toward selection rather than price impact, since entry types attend to different companies and those companies differ in volatility.

Taken together, these results suggest that who begins discussing stocks at any moment depends both on where users came from and on what made markets salient, and that this composition relates to market outcomes through where attention is directed rather than through how a given security responds.

Three objections deserve a direct answer. The first is that the types are an argmax over a standardized score, and that such a rule can manufacture apparent separation when categories differ in size on the platform. Near ties are uncommon, with the share of users whose top two scores differ by less than half a standard unit running from 1.1 percent to 7.6 percent across types. The labels are also stable, since splitting each user's pre-entry history at random into two halves and classifying each half separately yields agreement for 88.8 percent of users who receive a substantive label in both halves. Because each half uses only half of the available history, that figure understates the reliability of the labels we use.

The second objection is that our central market result is a null estimated on a constructed regressor, and that the within-security comparison removes the variation common to a security, which is the part most likely to be signal, raising the possibility that the disappearance of the composition effect is mechanical rather than economic. We address this by instrumenting one composition measure with another built from the two independent classifications just described, which cover the same authors over the same event windows and differ only in the labels they assign. The instrumented within-security estimate remains close to zero, its confidence interval still excludes the across-security estimate, and classification error turns out to attenuate the estimates by at most fifteen percent.

The third objection is that the entry results merely record the volume of news, or Reddit itself becoming more political in the same periods, rather than any change in who turns toward markets. The first-post evidence is built to rule out the former, comparing an episode's

own topic against every other topic on the same day and thereby differencing out anything that raised attention in general. For the latter, controlling for the share of all Reddit activity that falls in each category leaves the politics and news estimate within a fifth of its original value in either direction, and it remains significant at the one percent level.

Our analysis speaks first to the literature on household participation in financial markets (Mankiw and Zeldes, 1991; Haliassos and Bertaut, 1995; Vissing-Jørgensen, 2002; Hong et al., 2004; Campbell, 2006; van Rooij et al., 2011; Brown et al., 2008; Fagereng et al., 2017; Giannetti and Wang, 2016; Duraj et al., 2025), which explains why households differ in whether and how they invest, but which takes as its starting point people who already own stocks or hold brokerage accounts. We observe a margin that comes earlier, following users before their first public engagement with the stock market, and we show that among the users we observe this margin resolves into several distinct routes through which financial markets enter their attention.

Second, the paper connects to work on investor attention and social finance, in which attention, social interactions, communication, and narratives shape investor beliefs and behavior (Barber and Odean, 2008; Da et al., 2011; Hirshleifer, 2020; Shiller, 2019; Kuchler and Stroebel, 2021). Studies using social-media data have examined sentiment, disagreement, information transmission, and behavior among users already active in financial communities (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Bollen et al., 2011; Chen et al., 2014; Sprenger et al., 2014; Bartov et al., 2018; Cookson and Niessner, 2020; Giannini et al., 2019; Cookson et al., 2023, 2024), and related work documents attention-induced trading among retail brokerage users (Barber et al., 2022). We shift the object of study from communication within these communities to selection into them, which Reddit's multi-community structure makes observable.

Third, we show that the information environment relates not only to how many users turn to financial discussion but also to who does. Following the same users across communities makes this compositional margin visible, and reading their first posts ties the users who arrive during an episode to the content of that episode, giving evidence at the level of the individual entrant that aggregate series cannot supply.

Finally, we connect the composition of entry to market outcomes. A large literature relates social-media activity to prices, volume, and retail trading (Antweiler and Frank, 2004; Chen et al., 2014; Cookson et al., 2023, 2024), taking the membership of a financial community as given and studying what its members say. We show that this membership is itself systematic, and that it maps onto the securities the community discusses. Our results also caution against reading compositional shifts in a community's membership as differences in its price impact.

The remainder of the paper is structured as follows. Section 2 describes the data. Section 3

studies users' interests before they engage with stock-market discussions and defines the user types. Section 4 examines how the composition of entry varies with policy uncertainty and market conditions, and what entrants write when they arrive. Section 5 connects entry types to market outcomes, and Section 6 concludes.

2 Data

2.1 Reddit Archive and User Histories

We use a public archive of 26,121,553,873 comments and submissions posted to the 40,000 largest subreddits between June 2005 and July 2026 (Watchful1, 2026), a source that prior work has used to study social organization, user life cycles, movement across communities, and specific financial episodes.[3] For each contribution the archive records the public username, subreddit, timestamp, available text and a unique identifier, and we refer to comments and submissions jointly as posts, linking them over time by username.

Reddit's organization into topic-specific communities is the main advantage of the data for our setting, since the platform hosts stock-market forums alongside communities devoted to *Personal Finance*, *Economics*, politics, entertainment, local issues, and thousands of unrelated topics. The same account can participate in all of them. We can therefore reconstruct what users discussed elsewhere on Reddit before their first contribution to a stock-market community, observing a transition that would be unobservable in data collected only within financial forums such as StockTwits. Appendix A illustrates the platform's structure.

The main analysis covers users who entered a stock-market subreddit between January 2015 and July 2026. We nevertheless use all available pre-2015 activity to determine whether a contribution is a user's first observed stock-market post, preventing accounts already active before the window opens from being counted as new entrants. Stock-market activity before 2015 accounts for 5.3% of all stock-market posts in the archive.

2.2 Subreddit Taxonomy and Stock-Market Entry

To impose some structure on the data, we assign each of the 40,000 subreddits to one of 38 mutually exclusive categories, building on *r/ListOfSubreddits*.[4] The categories most relevant to our analysis are Stock Market, *Personal Finance*, *Economics*, *Crypto*, *Non-Stock Investing*,

[3] See Baumgartner et al. (2020); Waller and Anderson (2021); Hamilton et al. (2017); Zhang et al. (2017); Kumar et al. (2018); Semenova and Winkler (2021); Anand and Pathak (2022); Warkulat and Pelster (2024).

[4] https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits/

Gambling, *Politics and News*, *Productivity and Career*, *Law and Legal*, and *Home and Housing*. We refer to them as finance related or adjacent because they capture activity that has a direct relationship with the stock market (*Personal Finance*, *Economics*), represents relevant behavior (*Gambling*) or is related to a person's other financial decisions (*Productivity and Career*, *Law and Legal*, *Home and Housing*).

Each post inherits the category assigned to its subreddit, giving us a measure of participation in communities rather than of isolated references to stocks elsewhere on Reddit. A mention of a company in a general news forum does not count as stock-market participation, while a contribution to a stock-market subreddit does.

We use GPT-5.6 Luna to assign each subreddit to a category, giving the model the subreddit's name, the full taxonomy, the subreddit's self-description, and the title and text of eight randomly sampled submissions. It assigns exactly one category and reports a confidence score between zero and one. The final taxonomy includes all 40,000 subreddits, of which 320 are classified as Stock Market, and these communities account for 0.98% of all posts in the archive.

We assess the taxonomy against independent annotations of a fixed random sample of 250 subreddits, where GPT-5.6 Luna's agreement with the human annotator is substantial (Cohen's $\kappa = 0.734$) and the mean pairwise agreement across the human annotator and four language models is 0.795. Appendix B reports the full prompt, Table 15 lists category counts, post shares, and confidence scores, and Appendix D.3 describes the validation exercise.

We define entry as a user's first observed contribution to any of the 320 stock-market subreddits, and because the definition uses the full available history, each account has at most one first entry even if it later participates in several stock-market communities. The resulting ledger contains 6,804,621 users whose first observed entry occurs between January 2015 and July 2026. Entry therefore measures the beginning of public engagement with stock-market discussions on Reddit, not a user's first exposure to financial information or the date on which the user first invests. The months in which entry is unusually high are nonetheless unusual beyond the platform, since detrended monthly entry correlates at 0.69 with Interactive Brokers' trading activity and at 0.70 with Charles Schwab's new brokerage accounts, and both correlations survive removing the GameStop episode, as Appendix I reports.

2.3 Market Data and Security Mentions

We measure market outcomes with daily closing prices and trading volumes from LSEG, covering 5,686 securities between January 2015 and December 2025 in a panel of 10.7 million

security-day observations. We compute daily returns from closing prices and benchmark abnormal returns against the daily factor returns of Fama and French (1993) over the same period.

Market coverage ends in December 2025, while the Reddit archive runs through July 2026. The analyses of entry and its timing in Sections 3 and 4 therefore use the full Reddit window, while the market analysis in Section 5 covers January 2015 to December 2025.

We also build a panel of security mentions from the text of every submission and comment in the 320 stock-market subreddits, using the matching rule described in Appendix J. The rule gives priority to explicit cashtags and otherwise accepts uppercase tokens that are valid ticker symbols and are not common English words. The procedure yields 44.7 million mentions, each carrying the author and the posting date, which lets us merge mentions to the pre-entry user types. Appendix J.2 describes the screen that removes symbols used in this corpus as ordinary abbreviations rather than as securities.

3 Pre-Entry Activity and Stock-Market Engagement

This section studies users' activity in the year before their first engagement with a stock-market subreddit, tracing how that activity changes as entry approaches and which parts of it predict who enters. We then show that prior activity is substantially heterogeneous, with users approaching stock-market subreddits from different parts of the platform. We use this heterogeneity to define user types, and we show that the types differ in the way their members first engage with the stock-market part of Reddit.

3.1 How Activity Changes Before Entry

As entry approaches, users shift more of their activity toward finance-adjacent categories. For each user, category, and month in the year before entry, we divide the share of the user's activity in that category by the corresponding platform-wide share of Reddit activity. A value of one means that the user allocates the same share of their activity to the category as the average Reddit user, and values above one indicate relatively greater activity there.

Figure 1 shows that finance-adjacent activity among future entrants already exceeds the Reddit benchmark twelve months before entry and rises further as entry approaches, with *Non-Stock Investing* and *Personal Finance* carrying the largest ratios, followed by *Crypto*, *Economics*, and *Gambling*. The plot depicts the average pattern for the 47.4% of users in our sample for whom we observe sufficient prior activity, meaning at least ten non-stock posts in the year before entry. Table 16 formally tests whether this excess attention

changes systematically as entry approaches, relative to users' own average activity and contemporaneous Reddit-wide trends in each category.

One caveat governs how much of this pattern is behavioral. Entry is a user's first *observed* stock-market contribution and the archive records activity rather than account creation, leaving us with observed Reddit tenure rather than account age. Among the 6,804,621 entrants, 30.8 percent make their first stock-market contribution in the same month as their first observed activity anywhere on the platform, while 33.6 percent have at least 24 months of observed tenure at entry. For the first group the pre-entry window is mechanically short, making part of the rise in Figure 1 a matter of composition rather than a change in behavior, though the ten-post requirement removes most of these users from the figure. We cannot separate short observed tenure from a recently created account with these data, since the archive carries no account-creation date. These descriptive patterns show how activity changes before entry but not whether prior activity predicts who enters, which leads us to compare entrants with otherwise active users who have not yet entered.

3.2 Pre-Entry Activity Predicts Entry

We next test whether users' recent interests predict entry into stock-market discussion, estimating a monthly risk-set model that compares users who enter in a given month with otherwise similar users who do not. For computational tractability we draw deterministic samples of 300,000 eventual entrants and 300,000 non-entrants and follow them through July 2026. In each month the risk set contains sampled users who have not yet made an observed stock-market contribution and whose recent non-stock Reddit activity is observed, including both users who enter later in the sample and users for whom we never observe entry. A user leaves the risk set after their first observed stock-market contribution.

We estimate

$$\text{Entry}_{im} = \alpha_m + \sum_c \beta_c \, \text{PriorShare}_{ic,m-12:m-1} + X'_{i,m-1}\gamma + \varepsilon_{im}, \tag{1}$$

Here Entry_{im} equals one if user i makes their first observed contribution to any stock-market community in month m, and $\text{PriorShare}_{ic,m-12:m-1}$ is the share of the user's non-stock Reddit activity in category c during the preceding twelve months. The vector $X_{i,m-1}$ contains predetermined measures of the user's prior Reddit activity. Calendar-month fixed effects, α_m, absorb changes over time in aggregate Reddit activity and in the overall probability of entry.

Because category shares sum to one, coefficients are interpreted relative to the omitted category, Other Hobbies, and each coefficient measures how a ten-percentage-point reallocation of prior activity from Other Hobbies to category c is associated with the probability of entry

in the following month.

Entrants are deliberately oversampled relative to their frequency in the population, so all estimates are inverse-probability weighted to target population quantities. Eligible eventual entrants are sampled with probability 0.104 and carry a weight of 9.61, while never-entrants are sampled with probability 0.0437, with one quarter of each sampled user's months retained, giving an inclusion probability of 0.0109 and a weight of 91.62. The percentage-point magnitudes below are therefore population-targeting estimates rather than unweighted case-control slopes.

Figure 2 shows that prior activity in finance-adjacent categories strongly predicts subsequent entry. The largest association is for *Non-Stock Investing*, where a ten-percentage-point higher share of activity in the category is associated with a 1.3-percentage-point higher monthly probability of entering stock-market discussion. The corresponding estimates are also large for *Economics* and *Personal Finance*, at about one percentage point, and for *Crypto*, at about 0.8 percentage points. Prior activity in *Gambling*, *Home and Housing*, *Politics and News*, *Law and Legal*, and *Productivity and Career* also predicts entry, though less strongly, while most non-financial categories carry small or negative associations.

These results show that future entrants are not simply unusually active Reddit users, the composition of their earlier activity instead carrying substantial information about whether and when they first join stock-market discussion.

3.3 Classifying Entrants by Their Prior Activity

The analysis so far describes average patterns among future entrants, and those averages may conceal substantial heterogeneity in the interests that lead users to stock-market discussion. We therefore define user types to characterize the distinct pre-entry pathways through which users arrive.

We assign each entrant a type using only activity during the twelve months before their first stock-market contribution, comparing the user's activity share in each finance-adjacent category with the contemporaneous Reddit-wide share and taking the category with the largest standardized excess activity as the user's type.

$$R_{ic} = \frac{n_{ic} / \sum_d n_{id}}{(1/12) \sum_{s=-12}^{-1} q_{c,t_i+s}}, \tag{2}$$

A value of $R_{ic} = 1$ means that the user's share equals the platform benchmark, while values above one indicate overrepresentation. Here n_{ic} counts user i's posts in category c during months $t_i - 12$ through $t_i - 1$, $N_i = \sum_d n_{id}$ is the user's total activity in the non-stock categories, $q_{c,m}$ is the platform-wide share of category c in calendar month m, and t_i is the

user's first observed stock-market contribution month. For users with $N_i \geq 10$, we select the category with the largest standardized excess activity:

$$S_{ic} = \frac{p_{ic} - q_{ic}}{\sqrt{q_{ic}(1 - q_{ic})/N_i}}, \qquad p_{ic} = \frac{n_{ic}}{N_i}, \qquad q_{ic} = \frac{1}{12} \sum_{s=-12}^{-1} q_{c,t_i+s}.$$

The baseline rule also requires at least two posts in the category selected as a user's type. Users with at least ten pre-entry posts but fewer than two posts in the finance-adjacent category with the highest excess attention are classified as *Enough history, not finance-related*, while users with fewer than ten pre-entry posts are classified as *Not enough prior history*. These rules distinguish users for whom the data provide too little prior activity to assign a meaningful type from users with sufficient activity whose interests do not concentrate in the finance-adjacent categories used to define types.

Because the assignment is an argmax, two questions arise about how much it leaves to chance, the first being whether the winning category is typically clear. It is. Across the eight of nine substantive types for which the diagnostic is currently available, the share of users whose top two standardized scores differ by less than 0.50 runs from 1.1 percent for *Gambling* to 7.6 percent for *Politics and News*, against median gaps between 5.3 and 21.8.

The second is whether the label is stable, which we measure directly by splitting each user's pre-entry history at random into two halves and running the same assignment separately on each. The two labels agree for 88.8 percent of users who receive a substantive label in both halves. Each half uses only half of a user's history, making this a lower bound on the reliability of the labels we actually use, which are built from the full pre-entry window. Stability varies across types in a way that tracks the score gaps, running highest for *Crypto* at 93.3 percent and *Politics and News* at 93.1 percent, and lowest for *Economics* at 80.4 percent and *Law and Legal* at 80.0 percent. Appendix C reports both diagnostics by type.

Figure 3 describes both the prevalence of types and the shared activity between them. Panel A shows that a little over one quarter of the 6,804,621 entrants receive a substantive type, with *Politics and News* the largest, followed by *Personal Finance* and *Productivity and Career*. *Home and Housing*, *Crypto*, *Economics*, *Law and Legal*, *Non-Stock Investing*, and *Gambling* each account for smaller shares of entrants. The remaining users fall into one of the two residual groups, either lacking sufficient pre-entry history or holding sufficient activity that does not meet the minimum threshold in any finance-adjacent category.

Panel B shows that these labels capture dominant rather than exclusive interests, reporting for each assigned type the share of users whose activity is also elevated relative to Reddit-wide benchmarks in each other category.

Users assigned to one type often show elevated interest in several other categories as well,

an overlap that is natural because the defining category is simply the one that stands out most. Users assigned to the *Personal Finance* type, for example, are also unusually active in *Productivity and Career* and in *Home and Housing*, and there is substantial co-interest between the *Economics* type and the *Politics and News* category. The types therefore summarize the most distinctive component of users' pre-entry activity rather than partitioning users into mutually exclusive groups. Appendix C reports type coverage and label stability under alternative classification cutoffs.

3.4 Where User Types Enter

As a first test of whether our pre-entry types capture distinct online personas, we examine where users first enter stock-market discussion, since types that reflect meaningful differences in users' interests and motivations should predict which stock-market communities attract them. We focus on five large communities with distinct orientations: *r/WallStreetBets*, *r/Bogleheads*, *r/Options*, *r/StockMarket*, and *r/PennyStocks*.

Table 1 shows that entry is far from random, since the community a user first posts in lines up closely with what they were talking about beforehand, and the clearest split is between risk-seeking and risk-averse dispositions carried over from outside contexts. *Gambling* users, who are typically over-represented in casino and sports-betting content before entry, funnel into *r/WallStreetBets*, *r/Options*, and *r/PennyStocks*, three communities organized around leverage, volatility, or lottery-like payoffs, and they are the only type overrepresented in *r/PennyStocks*. *Personal Finance* users show the opposite pattern, entering *r/Bogleheads* at more than four times the average rate, essentially treating their first stock-market post as an extension of the budgeting and saving discussions they were already having. *Home and Housing* and *Productivity and Career* users, whose pre-entry interests also skew toward long-horizon financial planning, follow a similar though less pronounced path into *r/Bogleheads*.

A second pattern involves users drawn in by current events rather than by a settled investing philosophy. *Politics and News* users are overrepresented in both *r/WallStreetBets* and *r/StockMarket*, where the former likely reflects the wave of news-driven users who arrived during the GameStop squeeze, while the latter suggests that politically and economically engaged users are pulled into stock-market discussion more broadly whenever markets become newsworthy. *Economics* users follow a related but more specialized path, concentrating almost exclusively in *r/StockMarket*, a community whose framing is closer to market commentary than to portfolio strategy or speculation. *Crypto* users, meanwhile, pattern with the risk-tolerant group, appearing disproportionately in *r/WallStreetBets*, consistent with a shared taste for speculative assets that carries over from one domain to the other.

Together, these results offer an initial form of external validation, since types built solely from what users discussed before ever mentioning stocks nonetheless predict through which subreddits users enter.

3.5 Which Stocks User Types Discuss

The evidence in Section 3.4 is internal to Reddit, since entry community is itself a platform outcome, and a skeptical reader could argue that both the types and the communities reflect the same underlying activity data. A stronger test asks whether the types predict which individual securities users discuss, since securities are external to the platform and can be matched to market data.

This question does double duty, testing the classification against data the types never saw and forming the first stage of the market analysis in Section 5. If the path into stock-market discussion matters for markets, the first thing it must do is direct attention toward different securities.

We extract ticker mentions from every submission and comment in the stock-market subreddits over the sample period, using the matching rule and the screening described in Appendix J. Attaching to each mention the pre-entry type of its author yields 16.3 million mentions of 4,410 securities from users we can classify. The types were built from pre-entry subreddit activity alone and contain no information about securities.

Types Attend to Different Securities

We measure attention with the same standardized enrichment used in Section 3 to define types, now applied to securities rather than to subreddit categories. For type g and ticker k,

$$S_{gk} = \frac{p_{gk} - q_k}{\sqrt{q_k(1-q_k)/N_g}}, \qquad p_{gk} = \frac{n_{gk}}{N_g},$$

where n_{gk} counts mentions of ticker k by users of type g, N_g is type g's total mentions, and q_k is ticker k's share of all classified mentions, with positive values indicating that a type discusses a security more than the average classified user does.

Table 2 reports the most over-represented securities for each type, and the pattern follows the pre-entry interests closely. *Personal Finance* entrants concentrate on broad-market index funds, their four strongest securities being VXUS and BND, Vanguard's total international stock and total bond funds, and QQQM and QQQ, two Nasdaq-100 trackers. These are instruments of diversified long-horizon saving, aligning with the budgeting and retirement discussions these users came from.

Crypto entrants concentrate on the equity market's bitcoin proxies, with MicroStrategy their strongest security by a wide margin and the spot bitcoin fund IBIT also among their most over-represented, suggesting that these users are reaching for bitcoin exposure through instruments available in an ordinary brokerage account.

Gambling entrants concentrate on sports betting and casinos, with DraftKings and MGM Resorts their two strongest securities. This is the same disposition documented in Section 3.4, where these users funnel into *r/WallStreetBets* and *r/PennyStocks*, now visible in the individual companies they discuss.

Non-Stock Investing entrants stand apart in a different way, their strongest securities being small, little-known names such as FLNT, AEHR and PERI rather than any recognizable category of large company. *Productivity and Career* and *Home and Housing* entrants resemble *Personal Finance*, consistent with the co-interest structure in Figure 3.

Politics and News is the exception, and informatively so, being the least thematically coherent of the nine types. Its most over-represented securities mix mega-cap technology names, TSLA, PLTR and AAPL, with securities tied to particular episodes of retail attention, RKT, ICLN and CRSR, and no single category organizes the list in the way sportsbooks organize *Gambling* or index funds organize *Personal Finance*. This is what the label implies, since users who arrive from political and news communities attend to whatever is prominent at the time rather than to a standing interest.[5]

The Sorting Exceeds Chance

Enrichment rankings are informative but do not establish that the distributions differ, so we compare each pair of types by the total variation distance between their ticker distributions. This is the share of one type's mentions that would have to move to a different security for the two types to attend identically.

Any finite sample produces some separation by chance, and thin securities produce more of it, so we build a null by drawing mention counts under independence while preserving the expected margins, leaving each type with its total mentions and each security with its total mentions but the two unrelated. Table 3 reports the observed distances alongside this null.

The observed distances run from 3.6 to 7.3 times the null, with a median of about five. *Gambling* and *Personal Finance* are the most separated pair among the types, at 0.224 against a null of 0.047, meaning that nearly a quarter of gambling entrants' attention would have to

[5]One caveat applies to this type in particular. *Politics and News* contributes the largest share of classified mentions, and the benchmark q_k includes the type's own mentions, which shrinks its enrichment values relative to smaller types. Comparisons of enrichment magnitudes across types of very different size should be read with that in mind.

move to different securities to match where personal-finance entrants look. *Crypto* separates from both *Politics and News* and *Personal Finance* by similar margins.

We also summarize the whole table with the normalized mutual information between type and security, $I(g;k)/H(g)$. Knowing which security a mention concerns resolves 2.8 percent of the uncertainty about the author's entry type, against 0.102 percent under the null, making the observed value 27.5 times chance.

Taken together, these results establish the first stage of the market question. The types are constructed from what users discussed before they ever mentioned a stock, using no market data, yet they predict which securities users go on to discuss, and they predict them in the direction their labels imply. Savers discuss index funds, crypto entrants discuss bitcoin proxies, and gamblers discuss sportsbooks. Entry paths direct attention toward different parts of the market, and Section 5 asks what that does to prices.

These magnitudes are lower bounds. Types are assigned from pre-entry Reddit activity and use no market data, so classification error is independent of which securities a user later mentions, and a noisy label can therefore carry no more information about a security than the true label carries, leaving the measured association an understatement of the true one.[6] Misclassification works against these findings rather than for them, which is the opposite of its role in the within-security comparison of Section 5.

3.6 Entry Motives Across User Types

If prior activity reflects the concerns that bring users to financial discussion, users from different types should differ systematically in what they say, and in what they appear to seek, when they first arrive. We test this prediction by comparing the language of users' first stock-market posts across types.

For each user we analyze their first observed stock-market post, applying six short, concept-specific dictionaries to measure whether the post discusses *Economics*, politics, portfolio and household-investing decisions, trading and speculation, *Crypto*, or advice and help, and recording whether the post contains a question mark and how long it is. Our main text outcomes indicate whether a post contains at least one dictionary term or complete phrase, rather than relying on word densities that can be mechanically sensitive to the length of short comments. Appendix E reports the dictionaries and matching rules.

Table 4 reports the raw feature prevalence by entrant type and the full first-post benchmark. The displayed type rows exclude users with not enough prior history and those with enough

[6] If $\hat{g}$ is a noisy label satisfying $\hat{g} \perp k \mid g$, then $\hat{g}$, g and k form a Markov chain and the data processing inequality gives $I(\hat{g};k) \leq I(g;k)$. The same logic applies to the total variation distances, since misclassification mixes two types' security distributions toward one another, which can only reduce the distance between them.

history but no finance-related activity. The clearest differences align with the pre-entry type labels: *Politics and news* and *Economics* users use substantially more political language, *Personal finance* users use more portfolio language and are more likely to seek advice or ask questions, and *Crypto* users use far more crypto language. *Gambling* users have the highest prevalence of trading language, while *Economics* users are most likely to use economic terms, showing that differences observed before entry remain visible in users' first stock-market contributions.

The semantic features show that pre-entry pathways are associated with systematically different forms of initial engagement with stock-market discussion, revealing what users discuss but providing only limited evidence on why users enter or whether their first post concerns a personally consequential financial decision.

We therefore annotate a sample of 100,000 first posts, drawn proportionally across the nine defined user types, using GPT-5.6 Luna and a prespecified codebook (Appendix B.2), which assigns each post a primary motive and separately identifies whether the user expresses a personal financial stake or contemplates a concrete action. These measures allow us to distinguish posts aimed at learning, seeking advice, discussing events, evaluating investments, or participating socially, and to identify whether entry is connected to a stated financial decision or plan. Figure 4 reports the raw distribution of primary motives in users' first stock-market posts, separately for comments and submissions.

Figure 4 shows that entry motives broadly align with users' types. *Personal Finance*, *Law and Legal*, and *Non-Stock Investing* users frequently seek advice or validation and address implementation problems in their first post. *Home and Housing* users are most likely to describe a concrete financial decision in a submission (19%), while *Personal Finance* users also do so frequently (17%). *Politics and News* users are distinctively conversational and reactive, with social reactions accounting for 46% of their comments, whereas *Economics* users are especially likely to enter to learn or interpret external events. *Crypto* users are the most information-oriented group, with view or information motives accounting for 24% of submissions and 33% of comments. These descriptive results are consistent with what one would expect from each type: household-finance, legal, and general-investing interests lead to practical questions; political and economic interests lead to discussion of events and information; and *Crypto* interest leads to information-oriented engagement.

Table 5 then reports the prevalence of stated financial stakes and contemplated actions. *Personal Finance* entrants are most likely to express a personal stake (27.8%), contemplate a concrete action (21.3%), and do both (16.3%). These patterns show that a meaningful share of first contributions concerns concrete financial choices.

Taken together, the first-post evidence shows that the user types capture meaningful

differences in how people arrive at stock-market discussion, with users' stated motives and financial concerns at entry systematically aligning with the interests that characterized their earlier activity on the platform. More broadly, the results in this section suggest that interest in financial markets does not emerge through a single path, since some users arrive through household-finance concerns and others through economic and political events, investing, *Crypto*, or related interests.

4 Conditions and the Composition of Entry

Section 3 showed that future entrants differ substantially in their activity before they enter stock-market discussions and in the intentions with which they enter. We now turn from this cross-sectional heterogeneity to the conditions under which entry occurs, asking whether the mix of users arriving responds to the macroeconomic and market environment, and whether different types respond to different conditions.

4.1 The Question, and a Short Answer

The routes documented in Section 3 need not be equally open at all times. An environment that draws one kind of user into these communities need not draw the others in equally, raising the question of whether the composition of entry is itself a function of conditions.

The answer separates into two margins, and keeping them apart is what makes it legible. On the first, periods of elevated uncertainty raise entry both from users who were already active elsewhere on Reddit and from users with no visible platform history, drawing harder on the former, leaving the arriving population tilted toward people whose prior interests can be read. On the second, among entrants whose past is legible, uncertainty shifts the mix toward users arriving from politics and news communities and away from those arriving from budgeting and other investing communities, with no route associated with speculation gaining ground, while large market moves operate on a distinct and faster margin that affects different groups again. A third result connects the two to the people themselves, since entrants arriving during a dated episode write about that episode rather than about policy in general, in the first thing they ever post about stocks. Magnitudes appear in Sections 4.5 to 4.7.

4.2 Entry Timing Across User Types

The first fact we document is that entry is highly uneven over time, with Figure 5 showing by far the largest increase during the GameStop short squeeze in early 2021. Smaller increases occur around the onset of COVID-19 in early 2020 and the rise in U.S. policy uncertainty in

late 2024 and early 2025. The data suggest that people enter stock-market discussions all the time, but that entry rises sharply when either the platform (GameStop) or the market itself (COVID crash) becomes more salient to the public.

Aggregate entry conceals substantial differences in when different user types enter, a contrast that Figure 6 illustrates for *Gambling* and *Economics* users. Entry among *Gambling* users is concentrated during the speculative market environment of early 2021, whereas entry among *Economics* users rises more visibly during the macroeconomic and policy episodes of 2020 and 2025. While entry therefore seems to be connected to external events, an event that draws one type of user into stock-market discussion need not have the same relationship with other types.

4.3 A Daily Panel of Entry

We construct a daily panel covering 1 May 2015 to 31 July 2026, in which each of the 6,792,874 users who begin posting in a stock-market subreddit during the window appears exactly once, on the calendar day of their first such post. The panel therefore records arrivals rather than activity, and a user who entered in 2017 and has posted every day since contributes to a single day only, making every outcome below a property of the arriving population rather than of the conversation as a whole.[7]

Assigning a pre-entry type requires a legible past, and the requirement divides entrants three ways in a manner that shapes the analysis that follows. Slightly more than half, 3,575,079 users or 52.6 percent, post fewer than ten times in the twelve months before entry and cannot be assigned a type on any evidence, placing them beyond a measurement boundary rather than inside a substantive category. A further 1,318,083 users, 19.4 percent, clear that threshold with a median of forty-two pre-entry posts. Their dominant interest simply falls outside the nine categories the paper treats as routes into finance. They come instead from gaming, which accounts for 21.6 percent of the group, from media and entertainment at 8.5 percent, and from sports at 7.1 percent, making this a destination in its own right rather than an absence of interest. The remaining 1,899,712 users, 28.0 percent, receive one of the nine types. Entrants with a legible past therefore number 3,217,795, and it is within that population, where every member was eligible for every category, that the composition of entry can be compared.

We measure the environment with four daily series. Economic policy uncertainty enters as the log of the daily news-based index of Baker et al. (2016) and its first lag, market-implied

[7]Entry is dated in Coordinated Universal Time, and a first post may be either a submission or a comment. Comments account for 5.27 million of the 6.87 million entries in the full file, most users entering by replying to a thread that someone else began rather than by starting one.

volatility as the log VIX, and realized market movement as the absolute daily return on the S&P 500. Because entry into a discussion forum is not confined to market hours, the panel runs on calendar days including weekends, with market series carried forward across non-trading days and never beyond their own last observation, and with day-of-week fixed effects absorbing the weekly cycle in posting. All four series are available together from 3 May 2015 to 30 December 2025, giving an estimation sample of 3,895 days.

4.4 Specification

We estimate the two margins separately rather than as a single compositional system, because the two groups that lack a type are different objects. Users with too little history sit at a measurement boundary, and could not have been assigned a route on any evidence. Users whose dominant interest lies outside the nine categories are a substantive destination. Combining a boundary with a category inside one system would let variation in what can be measured masquerade as variation in where people come from.

The first stage takes the whole arriving population and asks whether conditions change the balance between entrants with and without a legible past,

$$y_t = \alpha + \gamma' X_t + \lambda_{y(t)} + \delta_{d(t)} + u_t, \tag{3}$$

where X_t collects the four conditions, $\lambda_{y(t)}$ and $\delta_{d(t)}$ are year and day-of-week fixed effects, and y_t is in turn the share of entrants with no usable history and the log entrant count of each group, the latter separating whether the environment recruits from outside the platform or activates users already on it.

The second stage restricts to entrants with a legible past and estimates, for each of ten categories,

$$c_{kt} = \alpha_k + \beta_k' X_t + \lambda_{k,y(t)} + \delta_{k,d(t)} + \varepsilon_{kt}, \tag{4}$$

where c_{kt} is category k's share of that day's entrants with a legible past, the ten categories being the nine types plus the residual destination described above.

Two features of this system are worth stating explicitly, since they determine what the joint test means. The ten shares sum to one and the design matrix is common across equations, leaving the coefficients on any regressor to sum to zero exactly, a restriction the estimates satisfy to eight decimal places. That has a useful consequence. A world in which every type responds identically and a world in which none responds are the same world, making the joint test of $\beta_{1j} = \cdots = \beta_{Kj} = 0$ precisely the compositional claim rather than a proxy for it. Weighting is likewise not a matter of taste. The median daily count of entrants with an

assigned type rises from twenty-three in 2015 to three hundred and fifty-seven in 2025, leaving the early years so thinly populated that their shares are close to noise. Unweighted estimation would give those years the same influence as the years in which the shares are precisely measured. We therefore weight by daily entrant counts, standardize each condition to unit standard deviation, and report Newey-West standard errors with a thirty-day bandwidth throughout.

4.5 Who Arrives: the Participation Margin

Table 6 reports the first stage. Weighting each day by its entrants, as the regressions do, 52.5 percent of entrants in the estimation sample arrive without a usable pre-entry history, and that share falls significantly when uncertainty is elevated, by 0.91 percentage points for a one standard deviation increase in the current log index and by a further 0.81 points on its lag. Periods of high uncertainty therefore deliver an arriving population that is more legible than usual, tilted toward users who were already posting somewhere on Reddit before they said anything about stocks.

The decomposition beneath that share is the more informative result, and it rules out the reading a fall in the share might otherwise invite. Entry rises in both groups when uncertainty is high. The contemporaneous index raises entry by 0.058 log points among users with no usable history and by 0.097 among users with a legible past, and its lag raises the two by 0.138 and 0.174 respectively, with all four coefficients positive and the latter three conventionally significant. It is not that newcomers stop arriving during uncertain periods; users already active on the platform simply arrive in greater proportion, and the compositional shift follows from the gap between two positive responses rather than from any withdrawal. The same pattern holds for market-implied volatility, where both groups again respond positively and the difference between them is small, whereas absolute market returns move neither group's count, giving the first indication that the two conditions operate through different channels.

This stage also disposes of a concern that would otherwise shadow everything in Section 4.6. Only 28.0 percent of entrants can be assigned a route, and if that classification rate itself moved with conditions, shares computed within the classified group could shift without anyone behaving differently. Here the classification rate is not a nuisance parameter but the outcome, and its movement is measured rather than assumed away. The compositional results that follow are then estimated within the population whose past is legible, leaving the boundary itself outside the system it might otherwise contaminate.

4.6 Which Pathway: Composition Among Entrants with a Legible Past

Table 7 reports the second stage, estimated on entrants with a legible past and excluding January and February 2021 for reasons set out in Section 4.8. The joint test rejects for every condition, with $\chi^2(9)$ of 25.66 for the log uncertainty index, 18.46 for its lag, 33.89 for the log VIX and 30.92 for absolute market returns, the smallest of these carrying a p-value of 0.030. The environment does not simply raise or lower entry; it changes who arrives.

Uncertainty moves the composition along a single clear axis. The share of entrants arriving from politics and news communities rises by 0.56 percentage points for a one standard deviation increase in the index, a movement on the order of a sixth of that category's mean share. Every other route that responds moves the other way. The share arriving from budgeting and retirement communities falls by 0.29 points, from non-stock investing by 0.06 and from gambling by 0.04, each significant at conventional levels, with crypto falling by 0.17 at the ten percent level. These estimates are not equally robust, and Section 4.8 separates them. The rise in politics and news survives every episode exclusion and every control for Reddit's own composition that we report, as do the declines among budgeting and non-stock investing communities on the current or the lagged index, whereas the declines in gambling and crypto are small and rest substantially on the tariff period of late 2024 and 2025. What survives throughout is the absence of any gain for those routes. Uncertainty does not trade one speculative route against another, drawing gamblers in as savers retreat; it draws in users whose prior attention already lay with public events, and in no specification we report does the uncertainty index significantly raise the share of a route associated with speculation, which is the opposite of the pattern that a story about uncertainty inviting speculation would predict.

Large market moves operate on a different margin and reach a different group. Absolute market returns leave the politics share statistically unchanged and instead reduce the share arriving from productivity and career communities by 0.06 points, an estimate that is the most stable single coefficient in the panel across every sample cut we report. Market-implied volatility is closer to a regime measure than to an event, and it reproduces the uncertainty pattern rather than the market-return one. It raises the politics share by 1.05 points and lowers the share of entrants whose interests lie outside the nine categories by 1.49 points, indicating that sustained fear keeps away precisely the gamers, sports followers and general posters whom ordinary conditions bring in.

These two channels are not equally well identified, and it would be misleading to present them as though they were. Absolute market returns vary at daily frequency and are measured

from prices rather than from counts of newspaper articles. They also survive the most demanding trend control we impose, leaving little room for a slow-moving omitted factor to generate them. The uncertainty index moves far more slowly, and Section 4.8 shows that its compositional effect resides chiefly in that slow variation, the joint test failing once month-of-sample fixed effects are imposed, which is the part of any time series most exposed to confounding by unmodelled trends. We therefore read the market-return result as evidence about how conditions shift composition and the uncertainty result as an association whose timing is consistent with such a shift, without asserting the two on equal terms. Section 4.7 narrows that gap from a different direction, showing at the level of the individual entrant that arrivals during an episode write about the episode, which is evidence the daily panel of shares cannot supply on its own.

Taken together, the two stages describe an entry margin that responds to the environment in a structured rather than a uniform way. Uncertain periods enlarge entry from every part of the platform, and enlarge it most among users whose past is legible. Within that population they tilt arrivals toward the politically engaged and away from budgeting and other investing communities, without drawing in the routes associated with speculation. Days of large market movement do something narrower, leaving the aggregate mix largely intact and acting instead on a single group. The composition of who is talking about stocks at any moment is therefore a function of conditions and not a fixed property of the platform, which is what makes composition, rather than the level of attention, the object that Section 5 carries into the market.

4.7 What Entrants Write on Arrival

Everything established so far is a property of a day rather than of a person, and a daily association between a published index and the composition of arrivals leaves one reading unaddressed. Nothing in Table 7 shows that any particular entrant arrived on account of the episode the index was registering, since the same correlation would appear if uncertain periods and politically inclined entry were both driven by something we do not observe. We therefore turn to the text of the first post itself, moving the evidence from the composition of a day's arrivals to the content of a single person's arrival, and asking whether people who enter during an episode write about that episode.

A weak version of the question can be put to the daily panel already in hand. We replace the compositional outcome with the share of that day's first posts engaging policy content, keeping the right-hand side, the weighting and the 3,836-day sample of Table 7 unchanged.[8]

[8]Policy content is the union of the four topical vocabularies present throughout the sample, namely tariffs and trade, Federal Reserve policy and rates, inflation and elections. Bank failures and the pandemic are

That share rises by 0.27 percentage points for a one standard deviation increase in the log uncertainty index and by a further 0.27 on its lag, against a mean of 2.65 percent. A day of elevated uncertainty therefore delivers arrivals roughly a tenth more likely than usual to be writing about policy at all. Absolute market returns leave the policy share unmoved, with a t-statistic of 1.5, reproducing on content the same asymmetry that Section 4.6 found on composition. This is suggestive rather than decisive, since a rise in policy talk during uncertain periods could reflect what is in the news rather than what brought anyone to the forum.

We therefore search the text of all 6,792,819 first posts for six topical vocabularies, covering tariffs and trade, Federal Reserve policy and interest rates, inflation, elections, bank failures, and the pandemic, and assemble twenty-two sharply dated episodes distributed across those six topics.[9] Each episode contributes a comparison window of seven to three days before the date and a response window of the date through two days after. We stack the resulting observations at the level of episode by topic by day,

$$y_{ekd} = \alpha_{ek} + \lambda_{ed} + \theta\,(\text{post}_d \times \text{own}_{ek}) + \nu_{ekd}, \tag{5}$$

where y_{ekd} is the log share of that day's first posts mentioning topic k, α_{ek} are episode-by-topic effects and λ_{ed} are episode-by-day effects.

The two sets of fixed effects are what make the design informative. The episode-by-topic effects absorb each vocabulary's own base rate, and the episode-by-day effects absorb anything that lifted every topic at once on a given day, leaving θ identified from the rise in an episode's own topic relative to the other five on the very same days. Generic news volume and generic attention are differenced out by construction, so the coefficient answers a narrow question, which is whether entrants on a tariff date write about tariffs specifically rather than about policy in general. Twenty-one episodes are too few for cluster-robust inference to be credible, so the reported p-value comes instead from a permutation test that reassigns topics to dates at random, exact under the null that the own-topic label carries no information.

Panel A of Table 8 reports the result. Own-topic mentions rise by 0.658 log points on episode days relative to the other five topics on those same days, a proportional increase of 93.0 percent in the mention rate, and no draw out of five thousand permutations of the

excluded here as episode-specific rather than continuously available, and all six are reported separately in Panel A of Table 8.

[9]The episodes are listed in Table 19 and the dictionaries in Appendix F. They are chosen for distinctiveness rather than sampled for coverage, since the identifying comparison here runs across topics rather than across days, and the list therefore carries no claim of completeness. Base mention rates over the whole window run from 0.06 percent of first posts for bank failures to 1.47 percent for elections. Every date was checked against a primary source, the news release for statistical series and the agency announcement for policy actions, taking the date on which the information became public rather than the date of any later write-up of it.

topic-to-date mapping reached that magnitude. All six topics move in the same direction, ranging from a 158 percent rise for bank failures to a 43 percent rise for the pandemic, so the pooled estimate is not the work of one episode family. The largest response belongs to the rarest vocabulary, bank failures appearing in 0.06 percent of first posts across the whole window, and the smallest to the pandemic, which by March 2020 was being discussed under every topic at once. Both are what the design predicts, since the episode-by-day effects absorb whatever lifted all six vocabularies together and leave an own-topic contrast only where the episode was distinctive. Three choices about the episode list move the estimate very little, the pooled figure running from 93 percent on the full list to 94 percent excluding both 2020 election entries and 113 percent excluding the three episodes of March 2020.[10]

The dated design requires us to choose the dates, and a reader entitled to suspect that the choices flatter the result deserves a version in which no one on this project chooses anything. We build one from the categorical decomposition of Baker et al. (2016), who report the policy uncertainty index separately for named policy categories, pairing ten of those categories one-to-one with ten dictionaries written to match the published category definitions rather than any episode we had in mind. Panel B estimates, on a monthly panel of topic by month with topic and month fixed effects, the response of a topic's mention share to its own category's index, identified purely from variation across categories within a month. Own-category uncertainty raises own-topic mentions by 0.381 log points per standard deviation, a proportional increase of 46.3 percent with a t-statistic of 12.9, and the permutation test over the 10! possible mappings returns a p-value of 0.0002.

Two of those ten pairs invite the objection that a news-based index and a text measure of the same subject would move together whatever entrants were doing, and the estimate dropping them is the one we would ask a sceptic to read. Excluding trade policy and health care cuts the coefficient by three fifths, to 0.151 log points or 16.3 percent per standard deviation, where it remains significant at a t-statistic of 4.8 and a permutation p-value of 0.025. The result therefore does not rest on tariff news predicting tariff talk. Three of the ten pairs nonetheless carry the wrong sign, and the explanation is a measurement failure rather than a contrary finding. Taxes, regulation and entitlements have the three flattest mention series in the set, with standard deviations of the log monthly share of 0.40, 0.36 and 0.45 against 0.60 for the next flattest and 1.60 for trade policy, so there is little there for a

[10] A comparison window can contain another episode on the list, which contaminates the control period rather than the response period. This happens twice. The entry for 7 November 2020 has the 4 November entry inside its own comparison window and on the same topic, which attenuates it, and dropping it returns the pooled estimate to 98 percent. The Federal Reserve entry of 15 March 2020 has the pandemic declaration of 11 March inside its comparison window. Dropping the March 2020 episodes raises the pooled estimate, though that change is not a clean measure of the overlap, since two of the three are pandemic episodes and the pandemic is the weakest topic on the list.

moving index to line up with. Counting what those dictionaries actually match, branch by branch, identifies the failure. In the taxes dictionary 78 percent of matches come from the bare word tax, which on a stock-market forum is a user's own liability rather than a policy, against 4 percent from the directional terms such as tax cut, tax hike and tax reform. In the entitlements dictionary 53 percent come from the word pension and a further 24 percent from social security, with under 1 percent from the term entitlement itself, and in the regulation dictionary 41 percent come from the generic root and 28 percent from the Securities and Exchange Commission named as an institution, much of that routine filing talk. The contrast with a pair that works is sharp, since national security is carried by proper nouns tied to particular episodes, Russia supplying 27 percent of its matches, Iran 14 percent and Ukraine 13 percent, which are event words rather than evergreen vocabulary.[11]

Neither design delivers an average treatment effect and neither is offered as one. The episode list is a set of distinctive cases rather than a sample of uncertainty episodes, and the categorical panel measures a monthly association rather than a response to a shock. What the two establish jointly is the individual-level link that a daily panel of shares cannot supply, since they show that people arriving during an episode are writing about that episode, in their own words, in the first thing they ever post about stocks. Taken together with the two margins documented above, this closes the gap between an environment that changes the composition of arrivals and a mechanism running through the attention of the people who arrive.

4.8 Selection and Sensitivity

The meme-stock episode. January and February 2021 are unlike any other period in the sample, and the concern is that a compositional result estimated across them would describe the squeeze rather than the environment. We exclude those fifty-nine days, 1.5 percent of the sample, from the main specification rather than confining the check to a footnote. The exclusion matters, and it matters selectively. The joint test on the log uncertainty index strengthens from 20.26 to 25.66, and the estimates for politics and news and for budgeting communities are essentially unchanged. The response of the crypto share, by contrast, falls from -0.30 to -0.17, and the joint test on the VIX falls from 64.38 to 33.89, remaining significant while losing roughly half its magnitude. The uncertainty and market-return results are properties of the sample as a whole; the crypto and volatility responses draw materially

[11]The matched text also shows a good deal of the pension vocabulary to be non-American, referring to individual savings accounts, superannuation and the British tax authority, so that dictionary is not measuring United States policy discussion at all. We report all ten pairs rather than dropping the three, since selecting pairs on the sign of their estimate is exactly the discretion the categorical design is meant to remove. The branch counts are in `stream_a/output/flat_topic_branches.csv`.

on the episode, and we report them accordingly.

Communities organised around markets outside the United States. The four conditions are American, and 9.5 percent of entrants arrive through communities built around another national market, raising the possibility that the estimates are diluted or distorted by users responding to conditions we do not measure. We exclude twenty-eight such communities, identified from their names before any estimate was inspected and covering 646,109 entrants.[12] Every joint test strengthens, the uncertainty statistic rising from 25.66 to 29.95 and the market-return statistic from 30.92 to 34.69, and no individual estimate weakens. That is the direction the concern predicts, leaving the main specification, which retains these communities, the conservative choice.

The level of uncertainty rather than its innovation. The literature more often measures uncertainty by innovations than by levels, and the natural check is whether the result survives that substitution. It does not. Replacing the log index with its autoregressive innovation leaves the joint test at 13.57 with a p-value of 0.139, and the estimates for budgeting communities and for gambling lose significance, although politics and news survives at a t-statistic of 3.3. The reading we take is that the compositional response is a property of the prevailing level of uncertainty rather than of departures from it, a conclusion that the frequency evidence below reaches independently. We note, however, that an autoregressive innovation removes the persistent component of the series by construction, and we cannot separate the absence of a response to surprises from insufficient variation in the surprise component to detect one.

The frequency at which the association operates. Estimating the same specification under progressively more demanding trend controls locates where each result lives, and the first four columns of Table 20 report the ladder. The joint test on the uncertainty index rejects with no trend control, with year fixed effects and with a quartic polynomial in calendar time, and fails once month-of-sample fixed effects absorb all between-month variation, at 8.0 on the current index and 7.2 on its lag. The politics and news share nonetheless retains about half its main estimate under that control, at a t-statistic of 2.5. The market-return composition survives every control, its joint test standing at 38.2 under month-of-sample effects, and so does the participation margin, the share of entrants without a usable history falling by 0.82 points per standard deviation of the index within the month. The compositional association with uncertainty therefore operates chiefly at monthly and lower frequency, while the association with market movement and the participation response to uncertainty both

[12]The largest are `IndianStockMarket`, `wallstreetbetsGER`, `IndianStreetBets` and `CanadianInvestor`. Communities organised around an individual security are retained even where the name resembles a country, since these are American listings.

operate within the month, which is the basis for the asymmetric reading offered in Section 4.6.

Individual episodes of high uncertainty. Outside the meme-stock months the sample contains two long stretches of elevated uncertainty, the pandemic year and the tariff period running from late 2024 into 2025, and the concern is that the composition result belongs to one of them rather than to uncertainty in general. We drop calendar 2020, the nine months from October 2024 to June 2025, and both together. The joint test on the current index survives all three, at 24.1, 18.0 and 19.6, and the rise in the politics and news share survives at the one percent level in each, ranging from 0.40 to 0.59 percentage points. The decline among budgeting and retirement communities survives each exclusion on the current or the lagged index, and that among non-stock investors on the current index, whereas the decline in gambling does not survive dropping the tariff period and that in crypto does not survive dropping both, which is why Section 4.6 does not rest its reading on either. The tariff period also carries the lagged joint test, which loses significance without it, and part of the participation margin, the share of entrants without a usable history then falling by 0.63 points at a t-statistic of 1.3 on the current index and by 0.69 points at the ten percent level on its lag.

Reddit's own topical composition. The share of all Reddit activity falling in politics and news communities is itself highly cyclical, ranging from 4.7 to 14.6 percent of posts in a month and highest in 2016, 2017 and 2020, and a larger pool of politically active users would raise the politics share of entrants whether or not uncertainty draws anyone in. We add the log share of all posts on the platform falling in each of the nine type categories, measured either in the month of entry or as an average over the preceding twelve months, the latter mirroring the window from which types are assigned. The same nine controls enter every equation, preserving the adding-up restriction and with it the meaning of the joint test. The joint test on the current index rises to 31.8 with the contemporaneous controls and falls to 18.9 with the trailing ones, remaining significant in both, and the politics and news estimate moves from 0.56 to 0.66 and to 0.50, significant at the one percent level each time. The compositional result is therefore not a reflection of swings in the size of the politically active population on the platform, although a monthly share of posts is a coarse proxy for the stock of users at risk of arriving as each type, leaving a rate computed against that stock as the sharper test.

Classification coverage. A reader may ask whether shares computed among entrants with a legible past reflect who arrives or who can be measured. Section 4.5 answers this directly by making the classification rate an outcome rather than a denominator, and the composition estimates are then taken within the population that clears the threshold, leaving no channel through which movement in the boundary can generate movement in the shares.

5 From Entry Types to Market Outcomes

Section 3.5 shows that entry types sort across securities, with different kinds of entrants attending to different companies in ways their pre-entry interests predict. This section asks what that sorting means for markets.

We separate the question into two stages. The first stage asks whether Reddit attention is associated with market outcomes at all, and the second asks whether the composition of that attention matters beyond its level, that is, whether it matters who is talking and not merely how much talk there is. The distinction is important for interpretation, and ex ante it is unclear which way it runs. On the one hand, entry paths may shape behavior, with users who arrive from gambling communities trading more aggressively and herding more than users who arrive from budgeting communities, in which case the composition of attention carries information about price impact beyond the securities involved. On the other hand, entry paths may shape only the allocation of attention, with every group responding to a given security in much the same way and the apparent differences between groups reflecting which securities each one looks at. The evidence favors the second reading.

5.1 Event Construction

We identify attention spikes from the daily count of mentions of each security in stock-market subreddits. A day is a spike day if the security receives at least five mentions and the count exceeds three times the trailing 90-day median of its own daily mentions. The rolling median makes the threshold adapt to how much a security is normally discussed, making a count that is routine for a heavily discussed company a spike for a rarely discussed one. During the first 90 days of a security's series the median is undefined and the five-mention floor applies alone. Spike days separated by five or fewer calendar days are merged into a single event, a gap of six days or more begins a new one, and we date each event at its day of peak mentions. The median event covers a single day, and the most sustained episodes extend to 69 days.

We match events to daily prices and volumes from LSEG, estimating for each event a factor model over trading days -120 to -11 relative to the event day and requiring at least 90 of the 110 possible observations. We then cumulate abnormal returns over days 0 to 5 and report results for the three-factor specification of Fama and French (1993), taking as our return outcome the absolute cumulative abnormal return, since Reddit attention need not predict the direction of a move.

We use three further outcomes. Log abnormal volume is $\log(\bar{V}_{0,5}/\bar{V}_{-120,-11})$, where $\bar{V}$ denotes mean daily share volume over the indicated window. Volume decay is $\log(\bar{V}_{0,5}/\bar{V}_{-120,-11}) - \log(\bar{V}_{6,20}/\bar{V}_{-120,-11})$, where a larger value means volume falls back further

after the event window, and reversal is the cumulative abnormal return over days 6 to 20.

We drop events on the 96 symbols that Appendix J.2 adjudicates as non-securities, which matters here because several of those symbols are themselves listed tickers, so spikes in the token `DRS` or `CDC` generate events whose mentions have nothing to do with the company. These events behave as one would expect of noise, with a mean absolute abnormal return of 0.034 against 0.059 for the retained events and a mean abnormal volume that is slightly negative. The screened sample contains 34,476 events on 2,602 securities.

5.2 Specification

Let e index events, $j(e)$ the security and $t(e)$ the event day. The first stage estimates

$$y_e = \beta \log M_e + \alpha_{t(e)} + \left[\alpha_{j(e)}\right] + \eta_e, \tag{6}$$

where y_e is one of the four outcomes, M_e is the total number of mentions in the event window, α_t is an event-day fixed effect and α_j is a security fixed effect included in the second specification only.

The composition stage adds the mix of authors:

$$y_e = \sum_{g \neq g_0} \gamma_g \, s_{ge} + \delta \, c_e + \phi \log M_e + \alpha_{t(e)} + \left[\alpha_{j(e)}\right] + \left[\mathbf{x}'_{j(e)}\boldsymbol{\lambda}\right] + \eta_e. \tag{7}$$

Here s_{ge} is the share of window mentions written by authors of type g, and g_0 is the omitted reference group. In Panel A of Table 10 the shares run over three groups defined by pre-entry history, with users having fewer than ten pre-entry posts as the reference. In Panel B they run over the nine focal types among classified mentions, with *Politics and News* as the reference, and c_e is the share of all window mentions that are classified. The vector $\mathbf{x}_j$ collects three time-invariant security characteristics and enters only where indicated, and standard errors are two-way clustered by security and event day throughout.

5.3 Attention Spikes and Market Reactions

Table 9 relates the size of an attention spike to the market outcomes around it.[13] The results are what a first stage requires: larger spikes go with larger absolute abnormal returns and higher abnormal volume, that volume falls back further over the following two weeks, and returns partly reverse. All four relationships hold with event-day fixed effects, so they are not

[13] We measure spike size by the total number of mentions in the event window. The peak single-day count is an alternative measure, but it correlates 0.92 with the window total, and including both produces unstable coefficients whose signs flip across specifications. We therefore use the window total throughout.

driven by market-wide days on which both Reddit activity and volatility are high, and all four survive the addition of security fixed effects, so they reflect variation within a security over time rather than a comparison between heavily discussed and lightly discussed companies. Reversal is the weakest of the four and is significant only at the 10 percent level once security fixed effects are included.

The magnitudes are moderate, with a one log point increase in window mentions associated with a 0.0128 increase in the absolute abnormal return, against a sample mean of 0.059. The reversal coefficient is negative and small, with larger spikes followed by lower returns over days 6 to 20, consistent with partial reversal of an attention-driven price move rather than with the incorporation of lasting news.

5.4 Composition Matters, But Only Across Securities

We now ask whether it matters who is talking, computing for each event the composition of mentions in the spike window by the pre-entry type of the author.

Panel A of Table 10 splits mentions by how much pre-entry Reddit history the author has, and events dominated by users with at least ten pre-entry posts show smaller reactions on all three outcomes. The pattern survives security fixed effects for the share of authors who receive a type, though for authors whose history is not finance-related it weakens and is no longer significant for abnormal volume. The finding is about the presence of a history rather than its content, since we cannot reject that the two coefficients are equal on any outcome, with the tests giving $p = 0.171$ for absolute abnormal returns, $p = 0.102$ for abnormal volume and $p = 0.342$ for volume decay. We therefore read this as an account-maturity result rather than a sophistication result, with authors who have fewer than ten pre-entry posts concentrating in the episodes with the largest reactions.[14]

Panel B turns to the nine types, where composition predicts market outcomes strongly among classified mentions. The eight type shares are jointly significant at $p = 2.5 \times 10^{-11}$ for absolute abnormal returns and $p = 4.7 \times 10^{-9}$ for abnormal volume once we include event-day fixed effects. Two types carry the result, since relative to *Politics and News* a higher *Gambling* share predicts larger reactions and a higher *Personal Finance* share predicts smaller ones. The gap between them is economically large, at 0.0271 in absolute abnormal return, roughly half the sample mean, and 0.176 in log abnormal volume, roughly the size of the mean itself.

The result disappears once we add security fixed effects, with the joint tests becoming

[14]The reference group is defined by a threshold on observed posts, not by the absence of any Reddit history. A user with nine pre-entry posts falls in it. We use the group as a proxy for account maturity and note that it is a noisy one.

$p = 0.492$ and $p = 0.311$, leaving no evidence that within a given security the mix of entry types discussing it predicts how the market responds.

This is a null, so it is worth stating what the data can and cannot rule out. The within-security gap between *Gambling* and *Personal Finance* is 0.0035 with a 95 percent confidence interval of $[-0.0040, 0.0110]$ for absolute abnormal returns, an interval that excludes the across-security estimate of 0.0271. The smallest effect the test would detect at 80 percent power is 0.0124, less than half the across-security estimate, so measured against the type shares as we observe them the design is powered to find an effect of the size that exists between securities, and it does not find one. The corresponding figures for abnormal volume are a gap of 0.0424 with an interval of $[-0.0102, 0.0950]$ against an across-security estimate of 0.176, and a detectable effect of 0.0752.

Two qualifications matter. First, the comparison has variation to work with, since for the difference between the *Gambling* and *Personal Finance* shares 82 percent of the total variance is within security rather than between, and 1,519 of the 2,135 securities in this sample contribute more than one event. A high within-security variance share is also what one would observe if much of the event-to-event variation were classification noise, however, leaving this on its own no evidence that the specification retains signal.

Second, and more seriously, the type shares are built from a classifier and are measured with error, and the within transformation removes the variation that is common to a security, which is the part most likely to be signal, raising the noise share of what remains. Classical measurement error of that kind shrinks the within-security coefficients toward zero whether or not a true within-security effect exists. The power calculation does not settle this, since it establishes that the design would reject an effect of a given size in the measured share, not in the true one.

These estimates cannot separate the two readings on their own, and the next subsection separates them directly, using two independent classifications of every author.

5.5 Is the Null Measurement Error?

The concern is specific. If the type shares are noisy measures of the true composition, the within-security coefficients are attenuated whether or not a true within-security effect exists, and the null in Table 10 would be an artifact of the classifier rather than a fact about markets.

The concern is also testable. We obtain two independent classifications of every author. For each user we split their pre-entry posting history at random into two halves and run the identical assignment procedure separately on each, which yields two labels for the same

person.[15]

Aggregating each set of labels over the same event window gives two composition measures, s^1_{ge} and s^2_{ge}. They are computed over an identical set of authors with an identical denominator. Only the labels differ. Writing $s^k_{ge} = s^*_{ge} + \varepsilon^k_{ge}$, the two measurement errors are independent conditional on the true composition, so s^2 satisfies the requirements for an instrument for s^1: it is correlated with the signal and uncorrelated with the other measure's error. Because the labels are built from pre-entry Reddit activity and the split is a seeded random draw, neither error can be related to market outcomes except through the true composition.

Table 11 reports the result. The first stage is very strong, with F statistics on the excluded instruments between 7,164 and 25,846, which is what two readings of the same underlying history should produce. The sample is 25,419 events on 2,001 securities, close to the 28,400 events used in Table 12.

Three things follow. First, correcting for measurement error does not restore the effect. The instrumented within-security gap between *Gambling* and *Personal Finance* is -0.0034 with a 95 percent confidence interval of $[-0.0111, 0.0044]$, which excludes the across-security estimate of 0.0249 from the same sample. The joint test on the eight type shares gives $p = 0.65$.

Second, the correction is small. The ratio of the ordinary least squares estimate to the instrumented estimate is 0.94 with event-day fixed effects and 0.85 with security fixed effects, so classification error was costing between six and fifteen percent of the coefficient. To lift -0.0034 to 0.0249 would require a reliability near 0.18, and attenuation shrinks estimates toward zero in any case rather than reversing their sign. The reason the correction is so modest is that the regressor is a share averaged over many mentions, so errors in individual labels largely cancel before they reach the event level.

Third, the two halves are exchangeable by construction, and reversing the roles of instrument and regressor changes nothing: the within-security gap moves from -0.0034 to -0.0032 for absolute abnormal returns.

One limitation remains. The instrument purges the component of classification error that depends on which posts fell in which half. It cannot purge error common to both halves, which arises when a user's pre-entry history is genuinely ambiguous between two categories. The correction is therefore partial, and the instrumented coefficient is a lower bound in absolute value. Given that the measured correction is at most fifteen percent, the common component would have to be both far larger and sign-reversing to overturn the conclusion.

[15]The split is implemented as an independent Binomial$(n_c, 0.5)$ draw on each pre-entry category count, which is equivalent to randomly assigning the user's individual posts to two halves. Both halves use the same Reddit-wide benchmark and the same ten-post and two-winning-post rules. Appendix C reports the agreement rates.

We read the evidence as ruling out measurement error as the explanation for the within-security null.

5.6 The Composition Effect Operates Through Security Selection

The mechanism can be tested directly, without relying on the within-security comparison at all, by adding three observable security characteristics to the event-day specification in Table 12: return volatility, dollar volume and price. If types differ because they select different kinds of companies, these characteristics should absorb the type coefficients.

They do. Volatility enters with a coefficient of 2.13 and is the dominant channel, with the three characteristics together absorbing 68 percent of the gap between *Gambling* and *Personal Finance* in absolute abnormal returns and 34 percent of the gap in abnormal volume. Security fixed effects, which absorb everything fixed about a company rather than only these three observables, remove 87 percent and 76 percent respectively. Three observable characteristics therefore recover most of what a full set of security fixed effects recovers, which is what a selection account predicts and what a behavioral account does not.

This is our central evidence on the second stage, and it does not share the vulnerability of the preceding subsection. The specification carries event-day fixed effects only, and the composition variable is never transformed within security, leaving the attenuation described above with nothing to operate on. Measurement error in the type shares would shrink every type coefficient toward zero, in all three columns alike, and would not explain why adding return volatility in particular is what closes the gap between *Gambling* and *Personal Finance*.

The reading is direct. *Gambling* entrants do not move prices more than *Personal Finance* entrants when they discuss the same company. They discuss different companies, and DraftKings and MGM Resorts are more volatile than Vanguard's total international stock and total bond funds, while volatile securities respond more to attention. The composition variable is informative about the market reaction because it is informative about the security, not because it identifies a group of users with unusual price impact.

5.7 Attention and Retail Holdings

Prices and volumes respond to every trader in the market, and none of the outcomes so far shows that attention on Reddit coincides with retail investors taking positions. We therefore turn to Robintrack, an archive of the number of Robinhood accounts holding each security, which Robinhood published until August 2020 and which Welch (2022) and Barber et al. (2022) use to study the holdings of its customers. The archive covers 811 securities from May 2018 to August 2020, leaving 1,925 of our attention spikes on 236 securities with complete

estimation and outcome windows. For each spike we measure abnormal holder growth as the change in log holders over days 0 to 5, net of the security's own trend over days −120 to −11.[16]

Table 21 shows that larger spikes go with faster growth in the number of Robinhood holders. With event-day and security fixed effects, a doubling of window mentions is associated with additional holder growth of about 1.1 percent over six days, roughly two fifths of the average abnormal growth around a spike, and the estimate is significant well below the one percent level. Because the comparison holds the security fixed, the association does not reflect popular stocks being both widely discussed and widely held, showing instead that the same security gains holders on the occasions when Reddit discusses it more. Over days 0 to 20 the estimates remain positive but are imprecise, leaving open whether the additional holders persist. The evidence is associational, since the news that raises discussion may also raise buying directly, and Robintrack cannot tell us whether the new holders are the users we observe. It nonetheless places the attention we measure alongside a change in retail positions, and not only alongside prices and volumes.

We also ask whether the composition of a spike predicts holder growth, repeating the specification of Panel B of Table 10 on this outcome. Table 22 finds no such relation, with the eight type shares jointly insignificant in every specification and the within-security gap between *Gambling* and *Personal Finance* close to zero. This null is uninformative, however, and we report it as a bound rather than as evidence. The smallest gap the design could detect is 0.0695, whereas a gap as large, relative to the variability of the outcome, as the one we find between securities for absolute abnormal returns would be 0.0637, leaving a sample that predates the GameStop episode and is less than a tenth the size of the price sample too small to answer the question.

Taken together, the two stages describe a chain rather than a direct effect, in which pre-entry interests are associated with which securities new participants attend to, and the characteristics of those securities are associated with how markets respond to that attention. Our contribution is the first link, which has not previously been observable, and we take the second from an established literature on retail attention and prices (Barber and Odean, 2008; Da et al., 2011; Cookson et al., 2023). Attention and prices are jointly determined and we do not identify a causal effect in either link, so the evidence does not establish that some entry paths carry more price impact than others, and we do not claim it.

[16]Holder counts are a stock of accounts rather than a flow of trades, so their growth measures how many accounts open a position and not how much they buy. Appendix H describes the construction and reports every coefficient.

6 Conclusion

Episodes of heightened retail attention to the stock market tend to be discussed as a single phenomenon. The GameStop episode of early 2021, when attention converged on Reddit's own stock-market communities, made speculation its most familiar form. Such readings rest on the volume of attention, however, and aggregate series cannot tell a new saver from a new gambler, leaving open who actually turns toward markets when they become salient.

This paper follows 6.8 million Reddit users across the platform's topic communities to show who begins discussing stocks, and under which conditions. Users arrive along distinct routes that are legible a year before their first contribution, running through budgeting and retirement, economics and politics, crypto, gambling, and several other finance-adjacent interests. These routes persist in the language and motives of first posts and in the securities users go on to discuss. The mix of arrivals also changes with conditions, though not in the direction a story about speculation would suggest. When policy uncertainty is high, a larger share of entrants arrives from politics and news, no route associated with speculation gains, and users arriving during a dated episode write about that episode when they arrive. Finally, the route a user took relates to market outcomes only through which securities draw attention. Episodes dominated by different types differ in their price reactions across securities but not within the same security, making selection rather than price impact the reading the evidence supports.

Three implications follow. First, a surge in retail attention is not one phenomenon, since the people who begin following markets during a policy crisis differ from those who begin during a speculative episode. Inferring either the motive or the likely price impact of a surge from its size alone therefore risks mistaking who is behind it. Second, among the users we observe, interest in markets forms along several routes rather than one, suggesting that efforts to broaden participation may work better through the concerns that bring people to markets than through a single appeal. For some those concerns are household budgets, and for others public events, housing, or careers, although whether the same holds for households who never discuss markets in public lies beyond what our data can show. Periods in which markets become unusually salient may offer moments for timely financial education, while also being moments when newcomers are most exposed to confusion, poor advice, or speculative narratives. Third, public discussion provides an observable stage between passive attention and financial action, complementing surveys, search data, and brokerage records. Linking this stage to subsequent account opening, portfolio choice, and trading would show when public engagement develops into financial participation, and when it does not.

Several limitations bound these conclusions. We observe public discussion rather than

portfolios or trades, leaving open whether the users we follow go on to invest. Our entry measure records a user's first contribution to a stock-market community, which need not coincide with the moment their interest in markets began. Routes can be read only for users with enough prior activity, and just over half of entrants post too little beforehand to be classified, making our routes a description of entrants with a legible history rather than of all entrants. The subreddit taxonomy and the first-post motive labels are assigned by a language model and checked against a single human annotator on a limited subsample, leaving us without a human-to-human agreement benchmark for either. The compositional results are associations that operate chiefly between months, the frequency most exposed to slow-moving confounders. They are also measured as shares of entrants rather than as entry rates among users at risk of entering, a gap that our controls for Reddit's own topical composition narrow but do not close. Moreover, the twenty-two dated episodes behind the first-post evidence were chosen rather than sampled, allowing that evidence to tie arrivals to the content of those episodes but not to identify an average effect of news on entry. The market evidence is associational, since attention and prices are jointly determined, and we do not claim that Reddit discussion causes the reactions we document. The within-security result is a null on shares measured with error, though instrumenting one classification with a second, independent classification of the same authors moves the coefficient by less than fifteen percent and leaves it far from the across-security estimate. Our market data end in December 2025, so the final months of the Reddit sample inform the analysis of entry but not the analysis of returns. Finally, Reddit users are not a representative sample of households, and the platform's own composition shifted over our sample, most visibly around the GameStop episode.

Understanding how people arrive at financial markets, and not only what they do once they are there, is a step toward explaining who participates and who does not. A natural next step is to ask which personal circumstances, alongside market conditions, time a person's first turn toward the stock market.

References

Anand, A. and Pathak, J. (2022). The role of Reddit in the GameStop short squeeze. *Economics Letters*, 211:110249.

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

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Barber, B. M., Huang, X., Odean, T., and Schwarz, C. (2022). Attention-induced trading and returns: Evidence from Robinhood users. *The Journal of Finance*, 77(6):3141–3190.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2):785–818.

Bartov, E., Faurel, L., and Mohanram, P. S. (2018). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review*, 93(3):25–57.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Brown, J. R., Ivković, Z., Smith, P. A., and Weisbenner, S. (2008). Neighbors matter: Causal community effects and stock market participation. *The Journal of Finance*, 63(3):1509–1531.

Campbell, J. Y. (2006). Household finance. *The Journal of Finance*, 61(4):1553–1604.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Cookson, J. A., Engelberg, J. E., and Mullins, W. (2023). Echo chambers. *The Review of Financial Studies*, 36(2):450–500.

Cookson, J. A., Lu, R., Mullins, W., and Niessner, M. (2024). The social signal. *Journal of Financial Economics*, 158:103870.

Cookson, J. A. and Niessner, M. (2020). Why don't we agree? Evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Duraj, K., Grunow, D., Haliassos, M., Laudenbach, C., and Siegel, S. (2025). Rethinking the stock market participation puzzle: A qualitative approach. Technical report, CESifo Working Paper.

Fagereng, A., Gottlieb, C., and Guiso, L. (2017). Asset market participation and portfolio choice over the life-cycle. *The Journal of Finance*, 72(2):705–750.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Giannetti, M. and Wang, T. Y. (2016). Corporate scandals and household stock market participation. *The Journal of Finance*, 71(6):2591–2636.

Giannini, R., Irvine, P., and Shu, T. (2019). The convergence and divergence of investors' opinions around earnings news: Evidence from a social network. *Journal of Financial Markets*, 42:94–120.

Guiso, L. and Jappelli, T. (2005). Awareness and stock market participation. *Review of Finance*, 9(4):537–567.

Guiso, L., Sapienza, P., and Zingales, L. (2008). Trusting the stock market. *The Journal of Finance*, 63(6):2557–2600.

Haliassos, M. and Bertaut, C. C. (1995). Why do so few hold stocks? *The Economic Journal*, 105(432):1110–1129.

Hamilton, W. L., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., and Leskovec, J. (2017). Loyalty in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Hirshleifer, D. (2020). Presidential address: Social transmission bias in economics and finance. *The Journal of Finance*, 75(4):1779–1831.

Hong, H., Kubik, J. D., and Stein, J. C. (2004). Social interaction and stock-market participation. *The Journal of Finance*, 59(1):137–163.

Kuchler, T. and Stroebel, J. (2021). Social finance. *Annual Review of Financial Economics*, 13(1):37–55.

Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, pages 933–943.

Mankiw, N. G. and Zeldes, S. P. (1991). The consumption of stockholders and nonstockholders. *Journal of Financial Economics*, 29(1):97–112.

Semenova, V. and Winkler, J. (2021). Social contagion and asset prices: Reddit's self-organised bull runs. *arXiv preprint arXiv:2104.01847*.

Shiller, R. J. (2019). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, Princeton, NJ.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Tumarkin, R. and Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51.

van Rooij, M., Lusardi, A., and Alessie, R. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, 101(2):449–472.

Vissing-Jørgensen, A. (2002). Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures. *NBER Working Paper*, (8884).

Waller, I. and Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature*, 600:264–268.

Warkulat, S. and Pelster, M. (2024). Social media attention and retail investor behavior: Evidence from r/wallstreetbets. *International Review of Financial Analysis*, 96:103721.

Watchful1 (2026). Subreddit comments/submissions 2005-06 to 2025-12. Academic Torrents. Watchful1 is the data-collector handle; based on the Pushshift Reddit dumps; zstandard-compressed ndjson files for the 40,000 largest subreddits.

Welch, I. (2022). The wisdom of the Robinhood crowd. *The Journal of Finance*, 77(3):1489–1527.

Zhang, J., Hamilton, W. L., Danescu-Niculescu-Mizil, C., Jurafsky, D., and Leskovec, J. (2017). Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Tables

Table 1: Relative Entry-Community Deviations by Pre-Entry User Type

User type	*r/WallStreetBets*	*r/Bogleheads*	*r/Options*	*r/StockMarket*	*r/PennyStocks*
Politics and News	+38.5%	-58.2%	-69.0%	+33.3%	-37.4%
Personal Finance	-27.6%	+308.2%	-39.3%	-12.4%	-36.1%
Productivity and Career	-1.4%	+70.9%	-38.1%	+21.9%	-30.3%
Home and Housing	-11.6%	+137.8%	-49.5%	+57.6%	-36.5%
Crypto	+17.5%	-51.7%	-23.3%	+33.2%	-20.4%
Economics	-15.3%	+0.0%	-32.3%	+66.8%	-45.0%
Law and Legal	-14.5%	+31.2%	-53.3%	+23.2%	-38.6%
Non-Stock Investing	-6.6%	+72.8%	-29.1%	+19.0%	-28.2%
Gambling	+22.0%	-7.5%	+7.3%	-4.2%	+33.6%

Notes: Each cell reports $100 \times [(\text{type-specific share}/\text{overall share}) - 1]$, where shares refer to entrants whose first stock-market contribution is in the indicated community. Thus, +50% means that the type enters through the indicated community 50% more often than the average entrant, while -50% means half as often. The table uses the full first-post file for the January 2015–July 2026 window; user types are assigned from pre-entry activity.

Table 2: Most Over-Represented Securities by Pre-Entry User Type

User type	Securities, ordered by standardized enrichment					
Politics and News	TSLA (79)	PLTR (62)	RKT (54)	ICLN (50)	CRSR (50)	AAPL (42)
Personal Finance	VXUS (205)	BND (114)	QQQM (62)	QQQ (55)	ATO (50)	BLNE (50)
Productivity and Career	VXUS (56)	FTWO (50)	PLTR (38)	ASTS (38)	QQQM (35)	QQQ (35)
Home and Housing	VXUS (69)	PULS (63)	BGM (42)	BND (36)	ICLN (25)	TMC (25)
Crypto	MSTR (189)	TSLA (63)	NVDA (50)	QQQ (43)	IBIT (43)	USO (41)
Economics	AVD (33)	TSLA (33)	EPI (28)	VXUS (27)	BYD (23)	ACHR (23)
Law and Legal	LYV (60)	PK (27)	DXC (27)	PLBY (27)	IFF (26)	ICLN (25)
Non-Stock Investing	FLNT (57)	AEHR (51)	PERI (40)	RCMT (30)	IONQ (29)	HIMX (29)
Gambling	DKNG (89)	MGM (51)	PTON (29)	PLTR (27)	BTU (27)	TSLA (26)

Notes: Each cell reports a ticker and, in parentheses, its standardized enrichment $S_{gk} = (p_{gk} - q_k)/\sqrt{q_k(1-q_k)/N_g}$, where p_{gk} is the share of type g's mentions that concern ticker k, q_k is ticker k's share of all classified mentions, and N_g is type g's total mentions. Higher values indicate that the type discusses the security more than the average classified user does. The sample covers all submissions and comments in the stock-market subreddits from January 2015 to July 2026, restricted to securities with at least 100 classified mentions and at least 250 trading days of price data, giving 16,302,391 mentions of 4,410 securities. Ticker mentions are identified by the rule in Appendix J; mentions from accounts flagged by the screening in Appendix K are excluded, as are the 96 symbols adjudicated not to be securities in this corpus. User types are assigned from pre-entry activity only and use no market data.

Table 3: Separation Between User Types in the Securities They Discuss

Type A	Type B	Observed	Null	Excess
Gambling	Not enough prior history	0.236	0.043	0.192
Gambling	Personal Finance	0.224	0.047	0.177
Gambling	Politics and News	0.161	0.045	0.116
Crypto	Politics and News	0.162	0.029	0.133
Crypto	Personal Finance	0.182	0.032	0.150
Politics and News	Personal Finance	0.174	0.024	0.150
Economics	Politics and News	0.130	0.034	0.095
Productivity and Career	Enough history, not finance	0.096	0.027	0.069
Normalized mutual information $I(g;k)/H(g)$		0.0281	0.00102	27.5×

Notes: Total variation distance between the ticker distributions of the two types, equal to half the sum of absolute differences in mention shares across securities. A value of 0.20 means that 20 percent of one type's mentions would have to move to different securities for the two distributions to coincide. The null column reports the mean over 40 draws in which mention counts are generated under independence between type and security while preserving the expected row and column margins, so each type keeps its total mentions and each security keeps its total mentions. The final row reports the mutual information between type and security, normalized by the entropy of the type distribution, with the ratio of observed to null in the last column. The sample is the same as in Table 2.

Table 4: Raw First-Post Features by Entrant User Type

User type	*N*	Economic	Political	Portfolio	Trading	Crypto	Advice	Question	Words
Overall first-post sample	6,804,360	1.6	2.2	4.7	2.5	2.5	3.7	25.3	25.6
Politics and news	657,490	2.3	5.4	2.1	3.3	1.1	1.9	22.1	23.0
Personal finance	282,970	2.3	2.1	**9.9**	3.1	1.6	**3.9**	**29.9**	**32.8**
Productivity and career	278,910	2.1	3.3	5.7	3.2	1.6	3.6	27.2	28.1
Home and housing	165,586	2.1	4.6	5.6	2.6	1.3	2.8	26.0	27.3
Crypto	157,229	1.8	1.5	3.5	4.5	**9.1**	2.6	24.2	23.9
Economics	156,934	**2.9**	**6.1**	3.8	3.0	1.6	2.9	25.3	28.7
Law and legal	107,380	1.9	4.5	4.7	2.8	1.4	3.3	27.6	29.1
Non-stock investing	56,795	2.8	2.9	5.4	3.8	1.8	3.0	25.5	28.2
Gambling	40,761	1.3	2.0	4.4	**4.7**	2.1	3.5	27.1	23.8

Notes: Economic, Political, Portfolio, Trading, Crypto, Advice, and Question are percentages of first posts containing at least one term or phrase from the respective dictionary. Words is the mean alphabetic word count. The Overall row uses non-deleted first observed stock-market contributions in the January 2015–July 2026 main analysis window; type rows exclude Not enough prior history and Enough history, not finance-related.

Table 5: Personal Financial Engagement at Stock-Market Entry

User type	N	Personal stake	Contemplated action	Stake + action
Politics and News	34,531	14.4	11.4	8.2
Personal Finance	14,862	27.8	21.3	16.3
Productivity and Career	14,648	22.1	17.1	12.9
Home and Housing	8,696	19.7	14.8	11.2
Crypto	8,257	18.5	16.1	10.6
Economics	8,242	15.2	12.3	8.5
Law and Legal	5,640	21.6	16.9	13.0
Non-Stock Investing	2,983	20.9	16.4	11.7
Gambling	2,141	22.2	19.0	14.3

Notes: Cells are raw within-type percentages. "Stake + action" equals one only when the author both discloses a personal financial stake and contemplates a concrete financial action. The annotation uses only the first contribution's text and is not used to construct types.

Table 6: Entry and the Environment: the Participation Margin

	log EPU_t	log EPU_{t-1}	log VIX	$\lvert r_t^m \rvert$
Share with no usable history (pp)	−0.906**	−0.808**	−0.797	−0.532
	(−2.0)	(−2.1)	(−1.1)	(−1.6)
log(1 + entrants), no usable history	0.058	0.138***	0.615**	−0.025
	(1.2)	(2.7)	(2.3)	(−0.8)
log(1 + entrants), legible history	0.097**	0.174***	0.645***	−0.002
	(2.1)	(3.0)	(2.6)	(−0.0)
log(1 + entrants), all entrants	0.076	0.155***	0.632**	−0.009
	(1.6)	(2.9)	(2.5)	(−0.3)
Days	3,895			

Notes: Each cell is a separate weighted least-squares regression of the row outcome on the four conditions jointly, estimated on a daily panel of entry into stock-market subreddits covering 3 May 2015 to 30 December 2025, the window over which all four conditioning series are available. A user enters the panel once, on the calendar day in UTC of their first post in a stock-market subreddit, and never again, making these arrival counts rather than posting counts. An entrant has a legible history if at least ten of their posts fall in the twelve months before entry, which is the requirement for assigning a pre-entry type; entrants below that threshold have no usable history. All specifications include year and day-of-week fixed effects and are weighted by daily entrant counts. Conditions are the daily news-based Economic Policy Uncertainty index of Baker et al. (2016) in logs, its first lag, the log VIX, and the absolute daily return on the S&P 500, each standardized to unit standard deviation, so coefficients report the effect of a one standard deviation move. Market series are carried forward across non-trading days and never beyond their own last observation. Newey-West t-statistics with a thirty-day bandwidth are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7: Entry and the Environment: Composition Among Entrants with a Legible Past

Pre-entry category	log EPU_t	log EPU_{t-1}	log VIX	$\lvert r_t^m \rvert$
Politics and News	0.559***	0.213	1.047***	0.119
	(3.7)	(1.3)	(3.7)	(1.3)
Personal Finance	−0.285***	−0.257***	−0.036	0.021
	(−3.0)	(−3.7)	(−0.3)	(0.5)
Crypto	−0.165*	−0.107	−0.048	0.013
	(−1.8)	(−1.2)	(−0.3)	(0.4)
Non-Stock Investing	−0.057**	−0.005	−0.010	−0.005
	(−2.5)	(−0.3)	(−0.4)	(−0.5)
Gambling	−0.040**	−0.029*	0.029	0.009
	(−2.1)	(−1.7)	(1.0)	(0.8)
Economics	0.011	−0.013	0.255	−0.051
	(0.1)	(−0.1)	(1.4)	(−1.2)
Home and Housing	0.060	−0.000	0.132	−0.013
	(1.2)	(−0.0)	(1.6)	(−0.4)
Law and Legal	0.009	0.025	0.038	−0.013
	(0.4)	(0.9)	(1.1)	(−1.5)
Productivity and Career	−0.012	−0.065	0.078	−0.063***
	(−0.2)	(−1.5)	(1.0)	(−3.4)
Outside the nine categories	−0.081	0.237	−1.485***	−0.017
	(−0.3)	(1.0)	(−3.7)	(−0.1)
Joint test, $\chi^2(9)$	25.66***	18.46**	33.89***	30.92***
p-value	0.0023	0.0302	0.0001	0.0003
Days	3,836			

Notes: The sample is entrants with a legible pre-entry history, defined as at least ten posts in the twelve months before entry, covering 3 May 2015 to 30 December 2025 and excluding January and February 2021, the months of the GameStop short squeeze. The dependent variable in each row is that category's share of daily entrants with a legible history, in percentage points; the ten categories are the nine pre-entry types plus a residual category holding entrants whose dominant pre-entry interest falls outside the nine, which is itself substantive and is dominated by gaming, media and sports rather than being an absence of interest. Because the ten shares sum to one and the design matrix is common across equations, the coefficients in each column sum to zero by construction, making the joint test of equal response and the joint test of no response the same hypothesis. All specifications include year and day-of-week fixed effects and are weighted by the daily count of entrants with a legible history. Conditions are defined as in Table 6 and are standardized to unit standard deviation. Newey-West t-statistics with a thirty-day bandwidth are in parentheses, and the joint test uses the analytic cross-equation covariance evaluated at rank nine. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8: The Content of the First Post

Panel A. Dated episodes: topic-by-date difference in differences

Topic	Episodes	$\hat{\theta}$ (log points)	Implied change
Bank failures	3	0.946	+158%
Federal Reserve and rates	4	0.834	+130%
Inflation	3	0.726	+107%
Elections	4	0.602	+83%
Tariffs and trade	5	0.526	+69%
Pandemic	3	0.360	+43%
All episodes pooled	22	0.658	+93%
Permutation p-value		< 0.0002	
Sensitivity to the episode list			
Excluding 7 November 2020	21	0.684	+98%
Excluding both 2020 elections	20	0.663	+94%
Excluding March 2020	19	0.755	+113%

Panel B. Categorical policy uncertainty: topic-by-month

Uncertainty category	Topic dictionary	Own slope	Implied change
Trade policy	Tariffs and trade	1.352	+286%
National security	Security and conflict	0.332	+39%
Health care	Health and pandemic	0.226	+25%
Government spending	Fiscal and stimulus	0.174	+19%
Monetary policy	Federal Reserve	0.079	+8%
Financial regulation	Banks and supervision	0.033	+3%
Sovereign debt	Debt and downgrades	0.025	+2%
Regulation	Regulation	−0.108	−10%
Entitlement programs	Entitlements	−0.117	−11%
Taxes	Taxes	−0.135	−13%
All ten pairs pooled		0.381***	+46%
		(12.9)	
Permutation p-value		0.0002	
Dropping trade policy and health care		0.151***	+16%
		(4.8)	
Permutation p-value		0.0248	

Notes: The sample in both panels is the text of the first stock-market post of each of the 6,792,819 entrants with non-empty text, covering 1 May 2015 to 31 July 2026. A post counts as mentioning a topic if its text matches that topic's regular-expression dictionary; dictionaries are listed in Appendix F. Panel A stacks observations at the episode-by-topic-by-day level over the twenty-two episodes listed in Table 19, using a comparison window of seven to three days before each date and a response window of the date through two days after. Every episode date was checked against a primary source in September 2026, taking the date on which the information became public, which for the Consumer Price Index is the 8:30am Eastern news release rather than the date of any summary article published afterwards. The sensitivity rows re-estimate the pooled coefficient on the stated subsets; the first two address the 7 November 2020 entry, whose comparison window contains the 4 November entry on the same topic, and the third addresses the March 2020 episodes, where the Federal Reserve entry's comparison window contains the pandemic declaration of 11 March. The dependent variable is $\log((\text{hits} + 0.5)/\text{posts with text})$, and the specification includes episode-by-topic and episode-by-day fixed effects, so $\hat{\theta}$ is the rise in an episode's own topic relative to the other five topics on the same days. The episode list is chosen for topical distinctiveness rather than sampled for coverage, and no coefficient in Panel A should be read as an average treatment effect. Twenty-two episodes are too few for cluster-robust inference, so the p-value comes from a permutation test reassigning topics to dates over 5,000 draws, with a null distribution of mean +0.005 and standard deviation 0.161. Panel B is a monthly panel of topic by month running from May

Table 9: Attention Spikes and Market Outcomes

Outcome	Event-day FE	Event-day + security FE
|CAR|, three-factor, days 0–5	0.0128***	0.0083***
	(0.0014)	(0.0010)
Log abnormal volume, days 0–5	0.1586***	0.1464***
	(0.0113)	(0.0094)
Volume decay	0.0632***	0.0595***
	(0.0080)	(0.0068)
CAR, days 6–20 (reversal)	−0.0042**	−0.0034*
	(0.0017)	(0.0017)
Events	34,476	34,476
Securities	2,602	2,602
Event days	3,541	3,541

Notes: Each cell is from a separate regression of the stated outcome on the log number of mentions in the event window, following equation (6). Cluster-robust standard errors are in parentheses, two-way clustered by security and event day. Abnormal returns are cumulated over trading days 0 to 5 relative to the event day, using the three-factor model of Fama and French (1993) estimated over days −120 to −11 with at least 90 of 110 observations. |CAR| is the absolute value of that cumulative abnormal return. Log abnormal volume is $\log(\bar{V}_{0,5}/\bar{V}_{-120,-11})$, where $\bar{V}$ is mean daily share volume over the window. Volume decay is $\log(\bar{V}_{0,5}/\bar{V}_{-120,-11}) - \log(\bar{V}_{6,20}/\bar{V}_{-120,-11})$, so a larger value means volume falls back further after the event. Reversal is the cumulative abnormal return over days 6 to 20. The sample is all spike events on securities with LSEG price coverage, excluding events on the 96 symbols adjudicated as non-securities in Appendix J.2. Each specification also absorbs the log peak single-day mention count. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 10: Composition of Attention and Market Outcomes

	No FE	Event-day FE	Event-day + security FE
Panel A. Pre-entry history. Reference: fewer than ten pre-entry posts. $N = 34{,}407$.			
Outcome: |CAR|, three-factor			
Share with an assigned type	−0.0318***	−0.0307***	−0.0080***
	(0.0032)	(0.0027)	(0.0021)
Share with history, not finance	−0.0216***	−0.0197***	−0.0043*
	(0.0043)	(0.0035)	(0.0025)
Difference			−0.0037
95% CI			[−0.0090, 0.0016]
Test of equality (p)			0.171
Outcome: log abnormal volume			
Share with an assigned type	−0.1568***	−0.1764***	−0.0716***
	(0.0234)	(0.0209)	(0.0192)
Share with history, not finance	−0.0960***	−0.1150***	−0.0305
	(0.0289)	(0.0244)	(0.0226)
Difference			−0.0411
95% CI			[−0.0904, 0.0082]
Test of equality (p)			0.102
Outcome: volume decay			
Share with an assigned type	−0.1694***	−0.1798***	−0.0909***
	(0.0179)	(0.0154)	(0.0150)
Share with history, not finance	−0.1617***	−0.1568***	−0.0716***
	(0.0246)	(0.0186)	(0.0178)
Difference			−0.0194
95% CI			[−0.0593, 0.0205]
Test of equality (p)			0.342
Panel B. Nine entry types. Reference: *Politics and News*. $N = 28{,}413$.			
Outcome: |CAR|, three-factor			
Gambling share	0.0169***	0.0131***	0.0026
	(0.0045)	(0.0043)	(0.0038)
Personal Finance share	−0.0122***	−0.0140***	−0.0009
	(0.0024)	(0.0021)	(0.0015)
Gambling minus Personal Finance	0.0290	0.0271	0.0035
95% CI	[0.0203, 0.0377]	[0.0184, 0.0358]	[−0.0041, 0.0110]
Detectable at 80% power	0.0124	0.0124	0.0107
Joint test, eight shares (p)	3.0e−11	2.5e−11	0.492
Outcome: log abnormal volume			
Gambling share	0.1092***	0.0951***	0.0217
	(0.0325)	(0.0317)	(0.0265)
Personal Finance share	−0.0674***	−0.0809***	−0.0207
	(0.0173)	(0.0158)	(0.0137)
Gambling minus Personal Finance	0.1766	0.1760	0.0424
95% CI	[0.1138, 0.2393]	[0.1129, 0.2391]	[−0.0102, 0.0949]
Detectable at 80% power	0.0897	0.0902	0.0752
Joint test, eight shares (p)	7.2e−07	4.7e−09	0.311

Notes: Estimates of equation (7), with cluster-robust standard errors two-way clustered by security and event day in parentheses. Composition is the share of mentions in the event window written by authors of each pre-entry type. Panel A omits the share written by authors with fewer than ten pre-entry posts, so coefficients are relative to that group and the equality test is a Wald test on the two reported coefficients. Panel B restricts to windows with at least one classified mention, uses shares of classified mentions, omits *Politics and News*, and adds the classified share as a control; the joint test covers the eight included shares, whose full vector appears in Table 12. All specifications control for the log number of window mentions and the log peak single-day count. Intervals are 95 percent, and the detectable-effect row reports the smallest difference rejectable at 80 percent power against a two-sided 5 percent alternative. Sample restrictions as in Table 9. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 11: Split-Sample Instrumental Variables Estimates of the Composition Effect

	Event-day FE	Event-day + security FE
Outcome: $\lvert CAR \rvert$, three-factor		
OLS, half-one labels	0.0234	−0.0029
	(0.0042)	(0.0037)
OLS, half-two labels	0.0229	−0.0031
	(0.0041)	(0.0036)
2SLS	0.0249	−0.0034
	(0.0044)	(0.0039)
95% CI	[0.0163, 0.0336]	[−0.0111, 0.0044]
2SLS, roles reversed	0.0253	−0.0032
	(0.0045)	(0.0039)
Joint test, eight shares (p)	3.0e−08	0.655
First-stage F, min–max	7,164–25,846	7,271–24,856
OLS / 2SLS	0.94	0.85
Outcome: log abnormal volume		
OLS, half-one labels	0.1455	0.0045
	(0.0299)	(0.0257)
OLS, half-two labels	0.1403	−0.0023
	(0.0300)	(0.0257)
2SLS	0.1526	−0.0027
	(0.0324)	(0.0279)
95% CI	[0.0890, 0.2161]	[−0.0574, 0.0520]
2SLS, roles reversed	0.1578	0.0045
	(0.0318)	(0.0276)
Joint test, eight shares (p)	1.6e−05	0.961
Events	25,419	25,419
Securities	2,001	2,001

Notes: Each entry is the difference between the *Gambling* and *Personal Finance* coefficients in equation (7), estimated on composition shares built from two independent classifications of the same authors. Cluster-robust standard errors are in parentheses, two-way clustered by security and event day. Each author's pre-entry history is split at random into two halves and the identical assignment procedure is run on each, so the two composition measures cover the same authors over the same event window and differ only in the labels. The 2SLS rows instrument the half-one shares with the half-two shares; the reversed rows exchange their roles, which is legitimate because the halves are exchangeable by construction. The first-stage F range covers the eight endogenous type shares. The sample is restricted to events containing at least one mention by an author who receives a substantive label in both halves, which requires enough pre-entry history to survive being divided, giving 25,419 of the 28,400 events in Table 12. The composition measured here is therefore the mix among authors with well-documented pre-entry histories. All specifications control for the log number of window mentions, the log peak single-day count, and the share of mentions contributed by two-label authors. Sample screening as in Table 9.

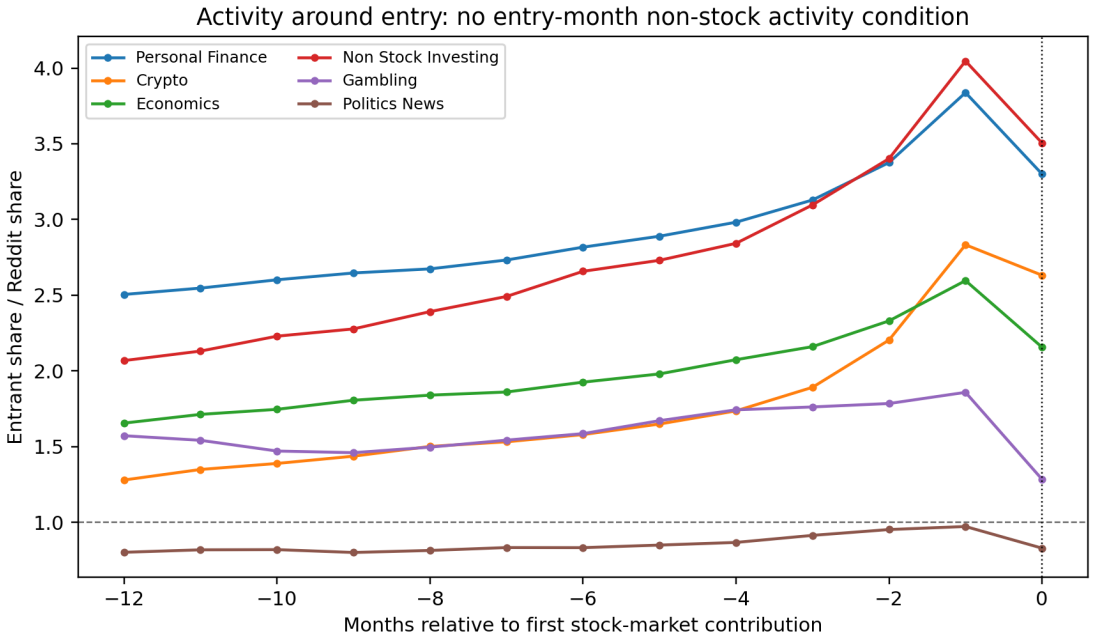
Table 12: The Composition Effect Runs Through Security Characteristics

	Event-day FE	+ characteristics	+ security FE
Outcome: |CAR|, three-factor			
Personal Finance	−0.0141***	−0.0061***	−0.0009
	(0.0021)	(0.0016)	(0.0015)
Productivity and Career	−0.0060**	−0.0047**	−0.0025
	(0.0025)	(0.0021)	(0.0019)
Home and Housing	−0.0091***	−0.0088***	−0.0059**
	(0.0033)	(0.0028)	(0.0027)
Crypto	−0.0012	−0.0042*	0.0005
	(0.0030)	(0.0023)	(0.0022)
Economics	−0.0065**	−0.0042	−0.0006
	(0.0030)	(0.0026)	(0.0023)
Law and Legal	−0.0021	−0.0043	−0.0012
	(0.0038)	(0.0032)	(0.0031)
Non-Stock Investing	0.0013	−0.0032	0.0007
	(0.0043)	(0.0038)	(0.0038)
Gambling	0.0131***	0.0025	0.0026
	(0.0043)	(0.0037)	(0.0038)
Return volatility		2.1321***	
		(0.0714)	
Log dollar volume		−0.0022***	
		(0.0004)	
Gambling minus Personal Finance	0.0272	0.0086	0.0035
95% CI	[0.0185, 0.0358]	[0.0012, 0.0160]	[−0.0040, 0.0110]
Share of the gap absorbed		68%	87%
Joint test, eight shares (p)	2.3e−11	0.0031	0.513
Outcome: log abnormal volume			
Personal Finance	−0.0809***	−0.0546***	−0.0208
	(0.0158)	(0.0165)	(0.0137)
Productivity and Career	0.0161	0.0219	0.0093
	(0.0201)	(0.0197)	(0.0174)
Home and Housing	0.0177	0.0190	0.0201
	(0.0282)	(0.0279)	(0.0250)
Crypto	−0.0134	−0.0242	−0.0239
	(0.0213)	(0.0208)	(0.0184)
Economics	−0.0401*	−0.0332	−0.0273
	(0.0239)	(0.0232)	(0.0210)
Law and Legal	−0.0077	−0.0162	−0.0078
	(0.0310)	(0.0298)	(0.0278)
Non-Stock Investing	0.0101	−0.0050	0.0080
	(0.0353)	(0.0355)	(0.0326)
Gambling	0.0953***	0.0622**	0.0216
	(0.0317)	(0.0302)	(0.0265)
Return volatility		7.8258***	
		(0.6332)	
Log dollar volume		−0.0171***	
		(0.0044)	
Gambling minus Personal Finance	0.1761	0.1168	0.0424
95% CI	[0.1130, 0.2393]	[0.0557, 0.1780]	[−0.0102, 0.0950]
Share of the gap absorbed		34%	76%
Joint test, eight shares (p)	6.5e−09	0.0009	0.310

Notes: Estimates of equation (7) on the $N = 28{,}400$ events for which all three security characteristics are available, with cluster-robust standard errors two-way clustered by security and event day in parentheses. Each coefficient is on the stated type's share of classified mentions in the event window, relative to the omitted *Politics and News* share. The middle column adds three time-invariant characteristics computed over the full LSEG panel for securities with at least 250 trading days: the standard deviation of daily returns, the log of median daily dollar volume, and the log of median closing price, the last included but not reported. All specifications control for the log number of window mentions, the log peak single-day count, and the share of all mentions that are classified. The gap row is the difference between the *Gambling* and *Personal Finance* coefficients, the widest among the nine types, and the absorbed row is the share of the event-day-FE gap that each column removes. Joint tests differ slightly from Panel B of Table 10, which retains 13 further events that lack characteristics. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Figures

Figure 1: Activity Around Stock-Market Entry



Notes: The figure plots the year before a user's first observed stock-market contribution, plus the entry month itself (months −12 through 0), as the average ratio of entrants' category shares to the corresponding platform-wide category shares. The sample comprises 3,224,789 users whose first observed stock-market contribution is to any stock-market destination and who made at least 10 non-stock posts during months −12 to −1. We impose no requirement of activity outside stock-market subreddits in the entry month, because that activity is realized at the entry date. Appendix C reports the earlier version of the figure, which imposed that condition and covered 2,883,824 users. Stock market category activity is excluded from both user-level and platform-wide denominators.

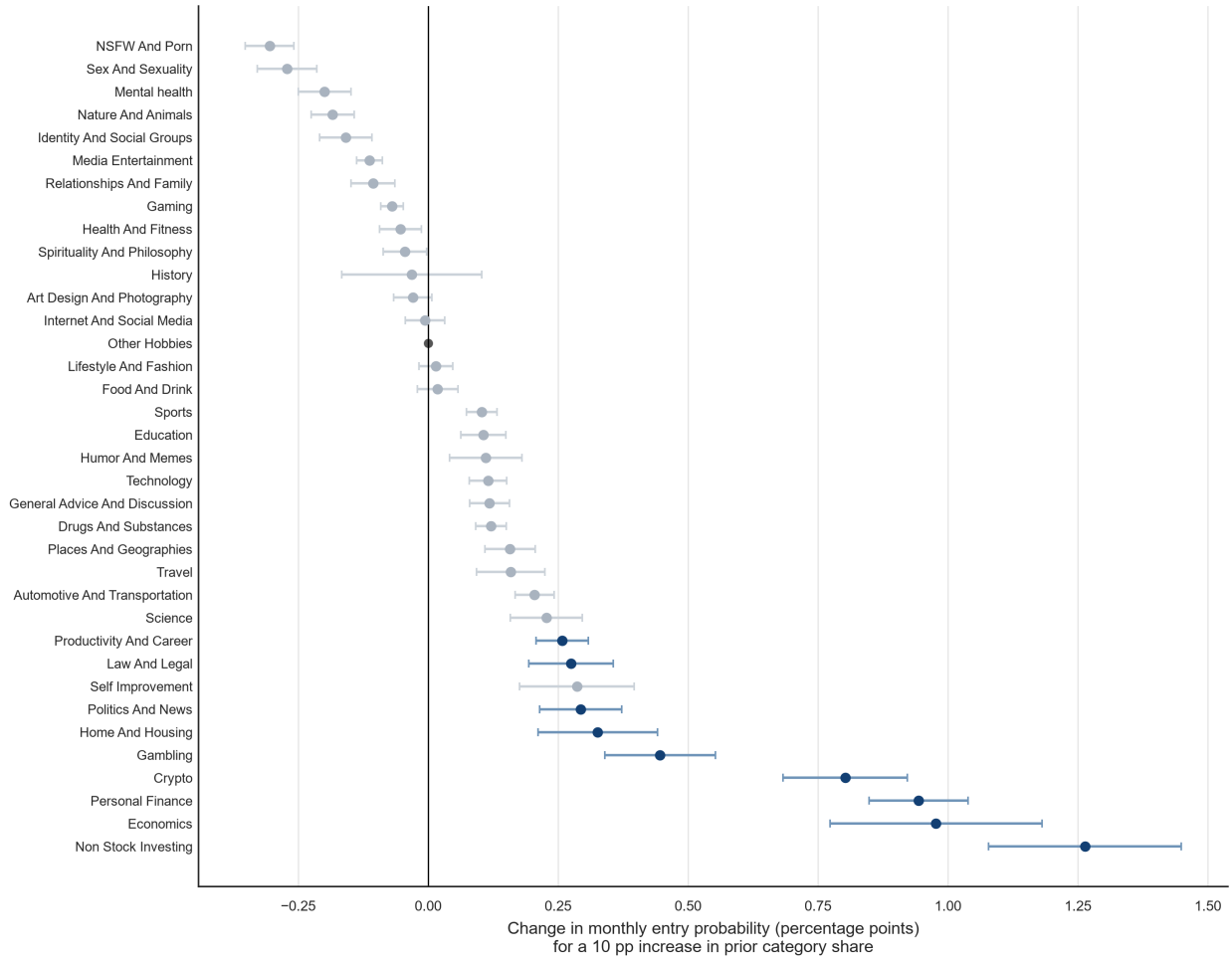


Figure 2: Risk-Set Model Estimates

Notes: The figure reports estimated associations between a user's category shares over the preceding twelve months and entry in the following month from the monthly risk-set model in Equation 1. Each point shows the change in monthly entry probability, in percentage points, associated with a ten-percentage-point increase in the prior share of activity in the indicated category, relative to Other Hobbies. Estimates are inverse-probability weighted for the choice-based sampling design, with weights of 9.61 for entrants and 91.62 for control user-months. Whiskers show 95% confidence intervals based on two-way clustering by user and calendar month.

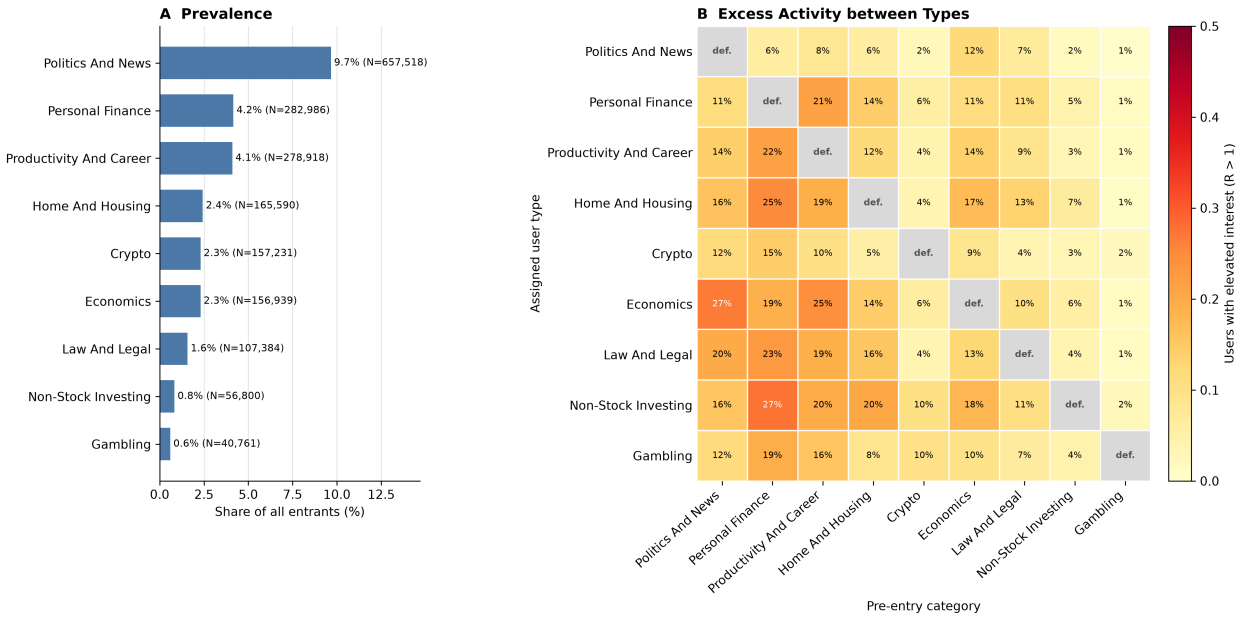


Figure 3: User-Type Prevalence and Pre-Entry Co-Interest Profiles

Notes: Panel A reports the number and share of all 6,804,621 entrants assigned to each substantive type. Panel B reports, for each row type, the percentage of users whose activity in the column category exceeds the contemporaneous Reddit-wide benchmark ($R_{ic} > 1$). Gray "def." cells indicate the category used to define the type and are therefore not independent validation outcomes. The two residual groups, Not enough prior history ($N = 3{,}579{,}832$) and Enough history, not finance-related ($N = 1{,}320{,}662$), are omitted.

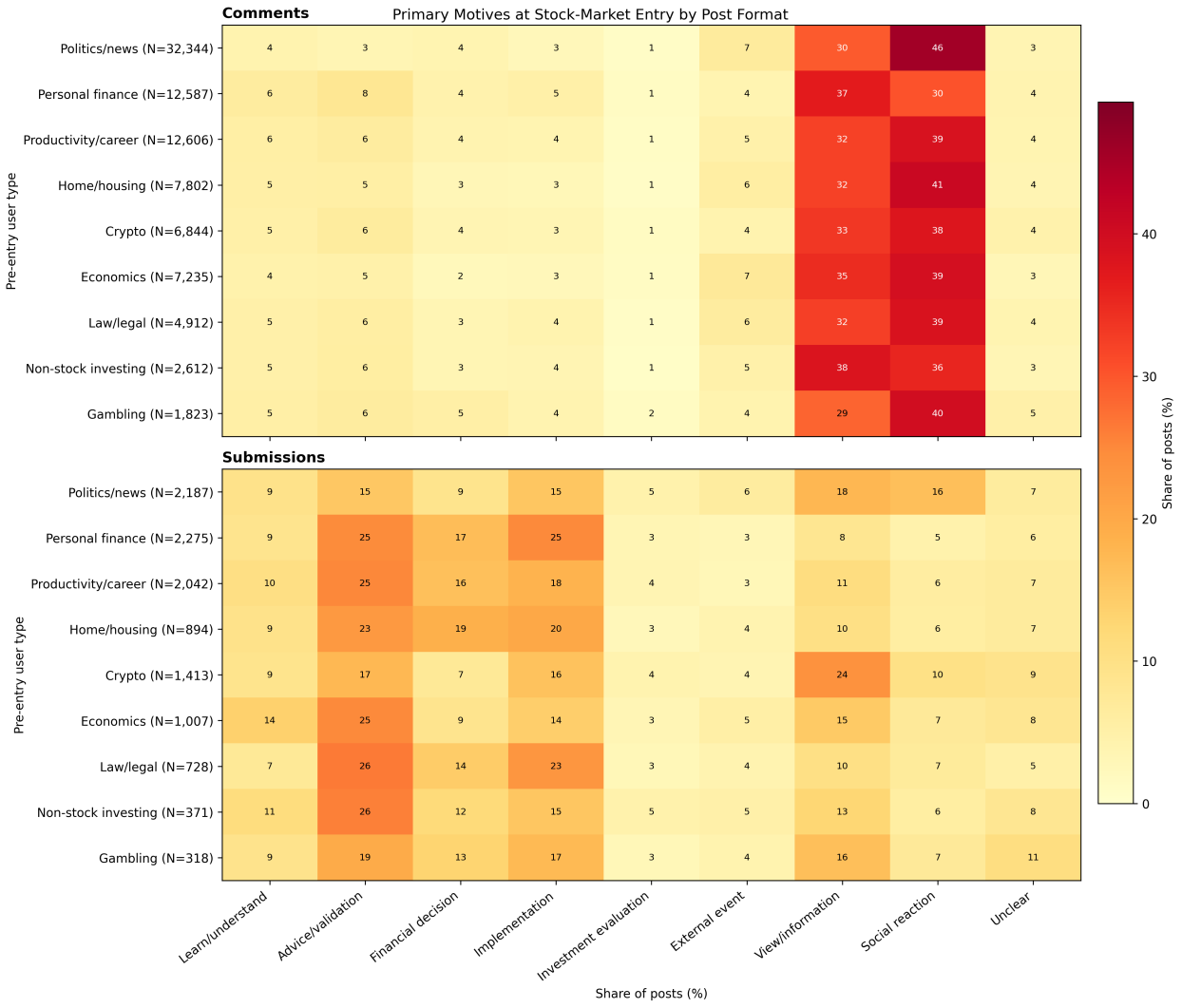
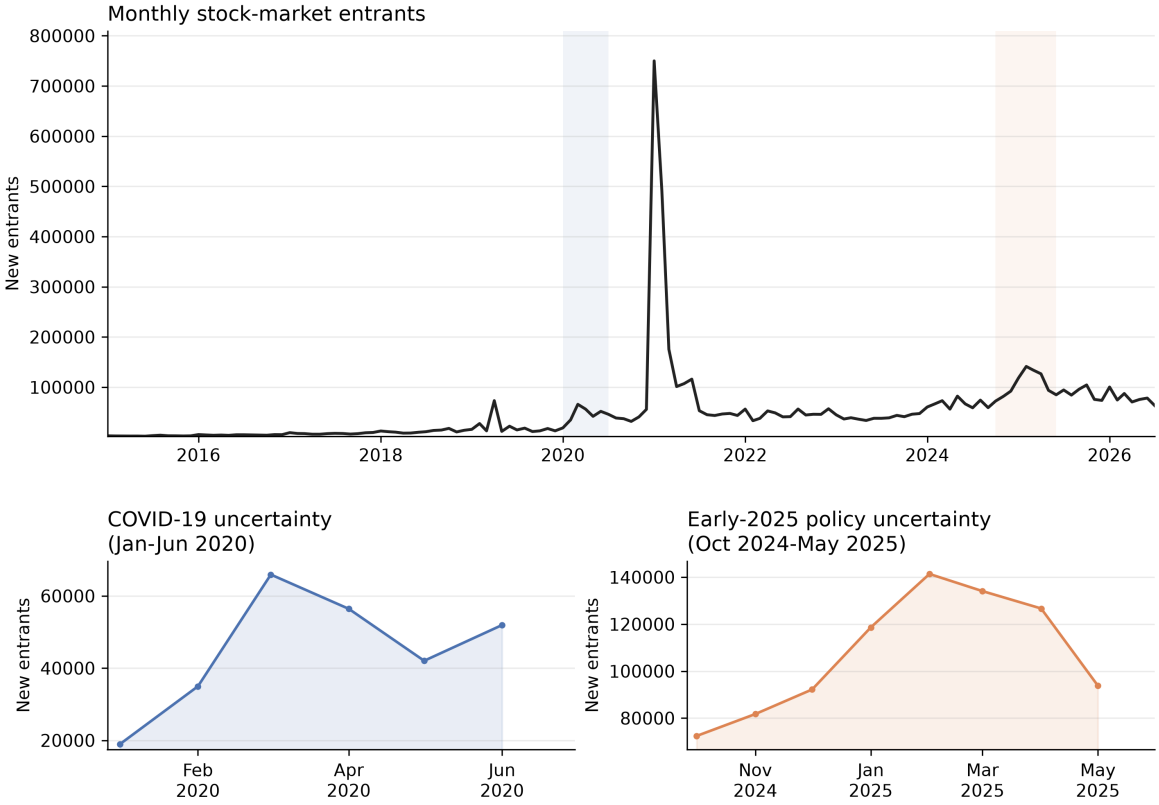


Figure 4: Primary Motives at Stock-Market Entry, by User Type and Post Format

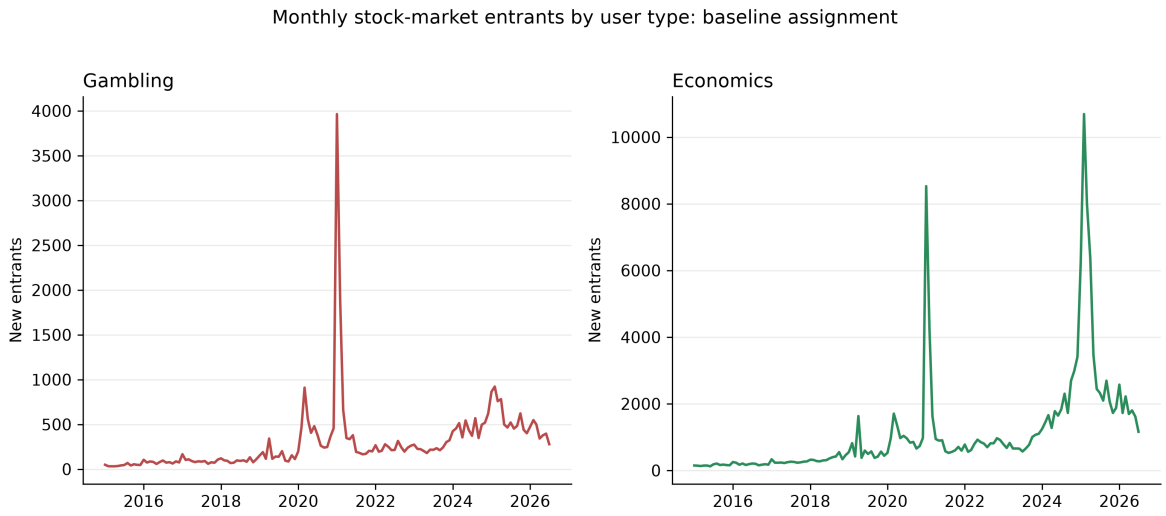
Notes: The figure reports raw within-type percentages for 100,000 first stock-market posts, drawn proportionally across the nine defined user types. Panels separately report comments and submissions. Each post receives one primary motive from the prespecified LLM codebook. The model observes at most the first 1,000 characters of the available title, self-text, or comment body.

Figure 5: Aggregate Monthly Entry into Stock-Market Discussion



Notes: The upper panel plots monthly first-time entry into stock-market discussion from January 2015 through July 2026. The lower panels zoom in on the COVID-19 episode (January–June 2020) and the rise in U.S. policy uncertainty surrounding tariff announcements (October 2024–May 2025); each uses its own vertical scale. Entry is a user's first observed contribution to any stock-market subreddit.

Figure 6: Monthly Entry for *Gambling* and *Economics* User Types

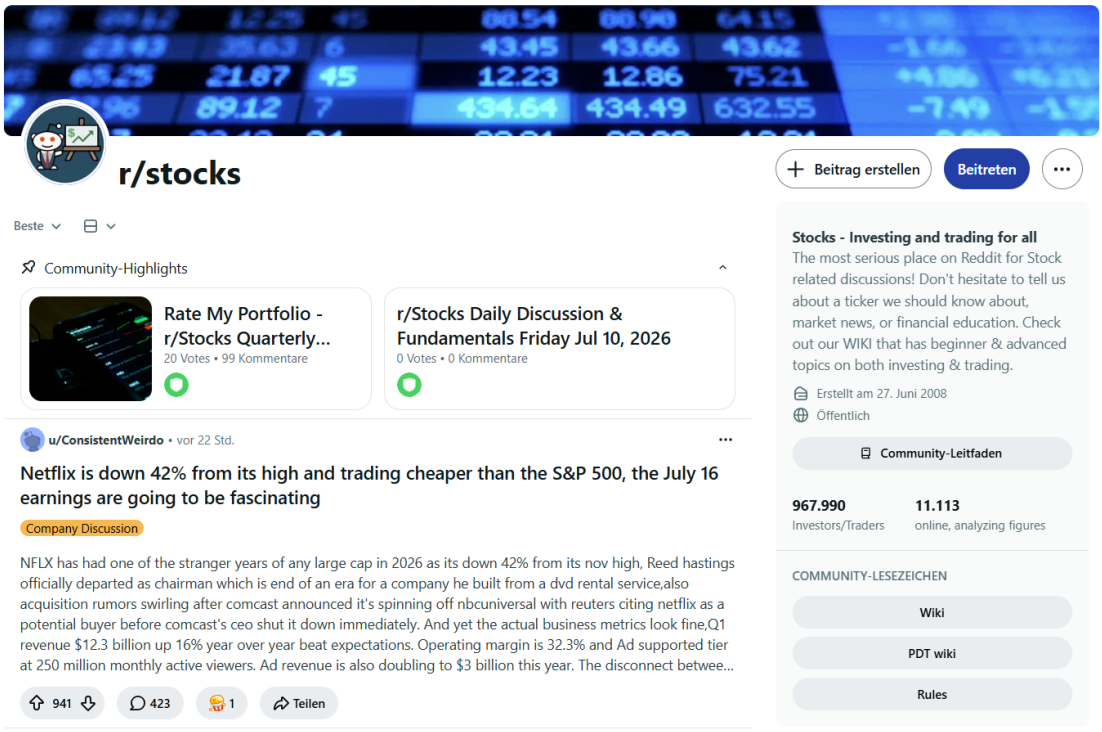


Notes: Each panel plots monthly entry for the indicated pre-entry user type: *Gambling* ($N = 40{,}761$) and *Economics* ($N = 156{,}939$). Vertical scales differ across panels.

A Structure of Reddit

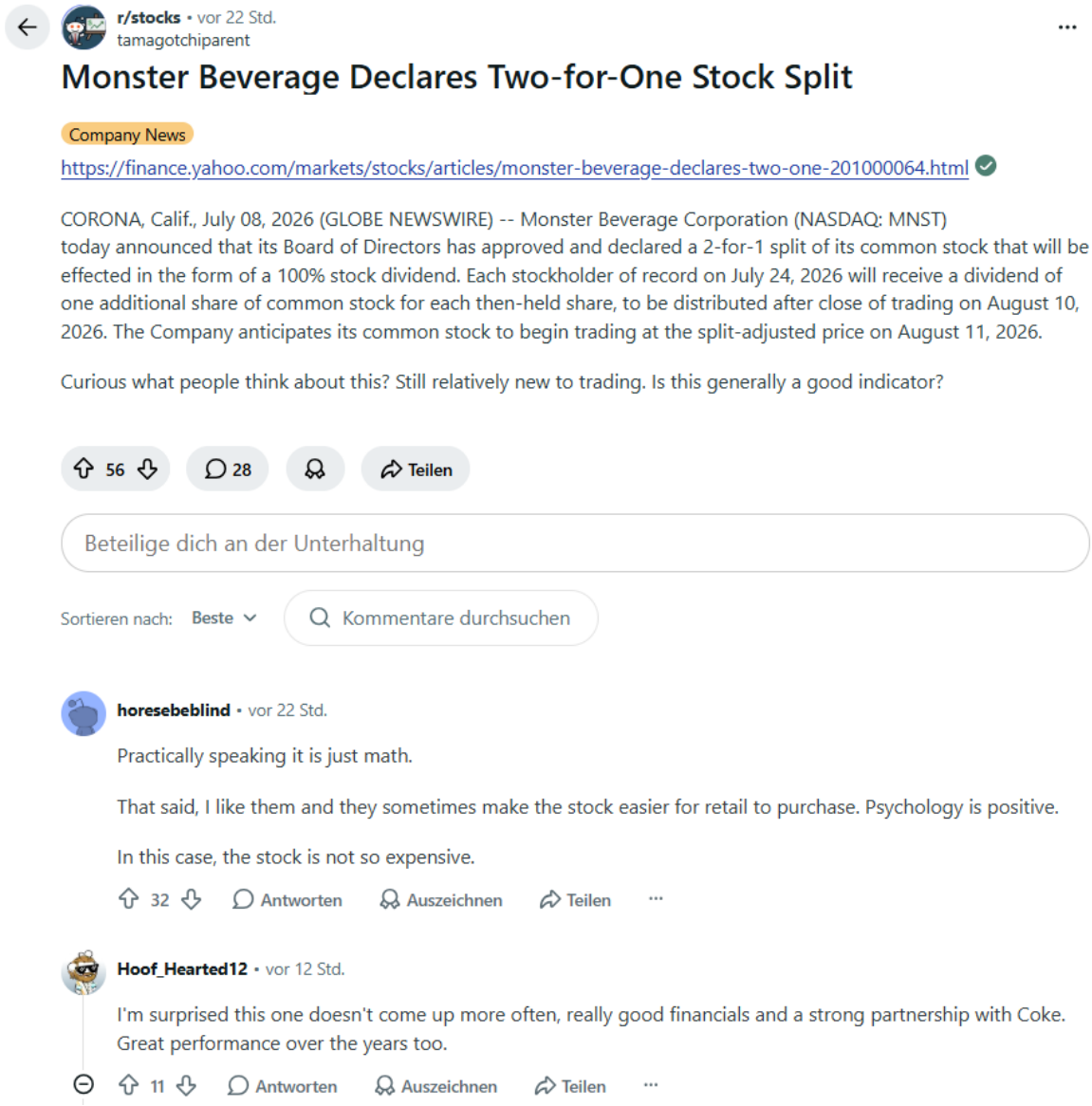
Reddit is organized into topic-specific communities called subreddits. Each subreddit has a front page that collects submissions and comments around the community's topic, while individual submissions can open into threaded discussions with nested replies. The two examples below illustrate these two levels of organization: a subreddit front page and a discussion tree.

Figure 7: Example Subreddit Front Page



Notes: Illustrative screenshot of a subreddit front page; it is not part of the analytical sample. Interface captured in July 2026; the displayed interface language is German.

Figure 8: Example Discussion Tree



Notes: Illustrative screenshot of a threaded discussion; it is not part of the analytical sample. Interface captured in July 2026; the displayed interface language is German.

B LLM Prompts

B.1 Taxonomy Prompt

Subreddits are assigned to taxonomy categories using GPT-5.6 Luna. For each subreddit, the model receives the subreddit name, its self-description when available, and eight representative submissions with their titles. It is instructed to assign one primary category to each subreddit. The operative instruction is:

> Assign each subreddit to exactly one category. For each subreddit, you receive its name, its self-description when available, and eight representative submissions with their titles. Use all of this information to identify the subreddit's overall topic; prefer consistent evidence across the description and submissions over incidental words in a single post. If evidence is weak, use `other_unknown` and a low confidence. Return JSON only.
>
> **Categories:**
>
> - `stock_market`: public equity investing, stock trading, ETFs, options, stock analysis, company valuation, portfolio management, and WallStreetBets-style investing.
> - `non_stock_investing`: investing outside public equities or cryptocurrencies, including bonds, fixed income, commodities, precious metals, private equity, venture capital, collectibles, alternative investments, real estate investing, rental properties, house flipping, and REITs.
> - `crypto`: cryptocurrencies, blockchain, NFTs, DeFi, tokens, mining, wallets, and crypto investing or trading.
> - `personal_finance`: budgeting, saving, debt, taxes, retirement planning, FIRE, banking, insurance, credit cards, mortgages, home buying, rent-versus-buy decisions, and household financial decisions.
> - `economics`: economics, macroeconomics, economic policy, business strategy, entrepreneurship, industries, firms, labor markets, and business or economic news.
> - `politics_news`: politics, government, elections, public policy, political ideology, current events, and news.
> - `technology`: programming, software, hardware, artificial intelligence, cybersecurity, consumer technology, operating systems, engineering, and software tools.
> - `science`: academic research, natural sciences, medicine as a scientific discipline, mathematics, physics, chemistry, biology, and scientific discussion.
> - `gaming`: video games, tabletop games, game-specific communities, esports, game development, and gaming culture.
> - `sports`: sports teams, athletes, leagues, competitions, sporting events, and sports discussion.

- `health_and_fitness`: exercise, fitness, nutrition, physical health, healthy lifestyles, sports training, recovery from physical injury, and general wellness.
- `mental_health_support`: mental health support, emotional wellbeing, therapy, anxiety, depression, ADHD, trauma, addiction recovery, crisis support, peer support, and communities focused on coping with psychological distress.
- `drugs_and_substances`: drug use, harm reduction, recovery, sobriety, psychoactive substances, cannabis, and communities focused on substance use or its consequences.
- `automotive_and_transportation`: cars, motorcycles, trucks, aviation, trains, bicycles, public transport, vehicle ownership, and transportation.
- `other_hobbies`: recreational hobbies and enthusiast communities including DIY, woodworking, crafts, model building, collecting, outdoor recreation, musical instruments, firearms, guns, shooting sports, weapons collecting, and other leisure activities.
- `art_design_and_photography`: visual arts, illustration, graphic design, photography, architecture, drawing, painting, and creative design.
- `food_and_drink`: cooking, recipes, baking, restaurants, beverages, coffee, tea, food culture, and culinary discussion.
- `lifestyle_and_fashion`: fashion, beauty, skincare, grooming, interior lifestyle, personal style, and aesthetics.
- `travel`: travel, tourism, destinations, trip planning, digital nomads, and travel advice.
- `productivity_and_career`: jobs, careers, resumes, interviews, workplace issues, studying for careers, professional development, certifications, productivity, and time management.
- `relationships_and_family`: dating, marriage, parenting, friendships, family relationships, and interpersonal relationships.
- `self_improvement`: habits, motivation, self-discipline, personal growth, life improvement, and non-clinical self-development.
- `sex_and_sexuality`: sexuality, sexual health, sexual relationships, sexual identity, and non-explicit sexual discussion outside adult-content communities.
- `spirituality_and_philosophy`: religion, spirituality, philosophy, ethics, belief systems, and existential discussion.
- `gambling`: sports betting, casinos, poker, lotteries, betting strategies, and gambling communities.
- `media_entertainment`: movies, television, music, anime, books, comics, podcasts, celebrities, and entertainment media.
- `humor_and_memes`: memes, jokes, satire, comedy, humorous images, shitposting,

and parody communities.

- `education`: schools, universities, learning resources, studying, teaching, academic life, and educational discussion.
- `history`: historical events, periods, people, archives, historiography, and discussion of the past.
- `identity_and_social_groups`: communities organized around identity, culture, ethnicity, nationality, gender, sexual orientation, disability, or shared social experience.
- `internet_and_social_media`: online communities, websites, social-media platforms, internet culture, content creators, and digital social interaction.
- `nature_and_animals`: pets, animals, wildlife, gardening, plants, environmental topics, and nature.
- `home_and_housing`: housing and home-related communities, including renting, apartments, roommates, home improvement, DIY renovation, appliances, interior design, and everyday housing issues not primarily focused on financial decisions or investment returns.
- `law_and_legal`: law, legal advice, courts, legal professions, law school, and legal discussion.
- `nsfw_and_porn`: adult content, pornography, explicit sexual content, and NSFW communities.
- `places_and_geographies`: countries, cities, regions, local communities, languages, and location-based subreddits.
- `general_advice_and_discussion`: general discussion, question-and-answer communities, AskReddit-style forums, advice, AMAs, support, and broad conversation.
- `other_unknown`: cannot be confidently assigned to any substantive category.

Output for each subreddit: `subreddit`, `primary`, `llm_confidence` from 0 to 1, and `llm_reason` as a short phrase.

B.2 First-Post Annotation Prompt

We separately annotate a fixed sample of users' first contributions to stock-market discussion. The model sees the post identifier and the first 1,000 characters formed from the title, self-text, and comment body. It is asked to classify the purpose of the contribution rather than its topic. The operative instruction is:

> You are classifying users' first contributions to a stock-market discussion forum.
>
> The input contains multiple posts. Return one annotation for every post, preserving input order. The id is only for alignment and must be copied exactly.

The goal is to identify why the user appears to be participating: what they want to achieve, learn, decide, or express. Do not classify the post merely by topic or keywords. Use only information explicitly stated or strongly implied by the text. Do not infer motives that are not supported.

Assign exactly one `primary_motive`, optionally one different `secondary_motive` (otherwise the string `none`), `personal_stake`, and `contemplated_action`. The two Boolean fields must be JSON `true` or `false`, not strings.

Primary-motive codebook:

`learn_understand`
: Wants to understand an investing concept, financial mechanism, market phenomenon, or how something works.

`seek_advice_validation`
: Wants others to recommend, evaluate, or validate a plan, idea, strategy, or possible choice.

`make_financial_decision`
: Is actively deciding what to do with their own money or investments, such as buying, selling, holding, allocating, investing, or changing a position. Prefer this over advice or validation when a concrete personal decision is central.

`solve_implementation_problem`
: Wants practical help executing an action, such as using a brokerage account, transferring funds, handling taxes, placing an order, or resolving platform or account mechanics.

`evaluate_investment`
: Wants to assess the attractiveness, value, prospects, or risk of a particular stock, asset, company, or investment opportunity.

`interpret_external_event`
: Wants to understand or discuss the financial implications of a current market, macroeconomic, political, geopolitical, or company-specific event.

`share_view_information`
: Primarily contributes an opinion, analysis, thesis, information, or advice rather than asking for something.

`social_reaction`
: Primarily jokes, agrees, disagrees, reacts, or participates conversationally without a clear informational or financial objective.

`other_unclear`
: No motive can be identified confidently.

Decision rules: classify purpose, not subject matter; classify a question about the author's own concrete financial choice as `make_financial_decision`; classify a request for feedback without a clear concrete action as `seek_advice_validation`; classify a conceptual question as `learn_understand`; and classify a post prompted by a recent event as `interpret_external_event`. Use `secondary_motive` only when a second motive is clearly present. `personal_stake` and `contemplated_action` are independent of the motive labels.

Boolean codebook: Set `personal_stake=true` only when the text states or strongly implies that the author or the author's household has money, income, debt, holdings, gains, losses, an account, or another direct financial exposure at issue. Generic discussion of an amount, ticker, or hypothetical investor is not sufficient. Set `contemplated_action=true` only when the author states or strongly implies a current or planned action involving money or an investment, such as buying, selling, holding, allocating, opening or changing an account, or paying a financial obligation. A general question about what investors might do, without an author-specific plan or decision, is false. If the text is too incomplete to support either flag, use false rather than infer the field.

Return JSON only as an object with an `items` array. Each item must contain the copied id and exactly the four annotation keys `primary_motive`, `secondary_motive`, `personal_stake`, and `contemplated_action`. `secondary_motive` must be `none` or a different value from the primary-motive codebook. Do not include explanations, confidence scores, or any other keys. Return items in the same order as the input posts.

C Robustness of Finance-Related User-Type Assignments

The user-type classification rests on two thresholds: the minimum pre-entry history required for a user to be classified at all, and the minimum number of posts the winning category must contain for its label to be assigned. Both are somewhat arbitrary, so this appendix asks whether a different reasonable choice would materially change the resulting types. We use the same nine finance-adjacent categories as the main text (politics/news, personal finance, crypto, economics, non-stock investing, gambling, law/legal, productivity/career, and home/housing) and the same standardized-enrichment procedure: each user is compared against the contemporaneous Reddit-wide benchmark over the twelve months before entry; users below the history threshold are labeled *Not enough prior history*, and users above it whose top category fails to clear the winning-post threshold are labeled *Enough history, not finance-related*.

The baseline rule requires at least ten pre-entry posts and at least two posts in the winning category. Under this rule, just under 28% of entrants receive a focal-category label, a little over half lack sufficient history, and the remainder have history but no qualifying category. Politics/news is the largest type, followed by personal finance and productivity/career, with the remaining six categories forming a longer tail (exact shares in the main text).

Table 13 varies both thresholds, and optionally adds a minimum raw enrichment ratio, and reports the resulting typed share alongside label agreement with the baseline. The classification is stable under sensible variation: raising the history requirement to fifteen or twenty posts, or the winning-post requirement to three or five, leaves most baseline labels unchanged among users typed at baseline. Adding a raw-enrichment floor shrinks the typed population but does not disturb the labels of users who remain typed. Tightening any threshold mainly moves marginal users into the two unclassified categories rather than reshuffling them across finance-adjacent types, which is the pattern expected if the labels track real, if noisy, differences in pre-entry behavior rather than artifacts of where the cutoffs happen to sit.

These checks support the ten-post/two-winning-post rule as a transparent compromise between coverage and conservatism: stricter raw-ratio cutoffs trade one for the other without changing the underlying story. As throughout the main text, every assignment uses only pre-entry activity, so classification is fixed before, and cannot be influenced by, entry timing, first-post content, or post-entry behavior.

Table 13: Robustness to User-Type Assignment Cutoffs

Min. history	Min. winner	Min. raw ratio	Typed share	Not enough history	Agreement[†]
5	2	–	29.43%	46.37%	–
10	1	–	34.80%	52.61%	86.9%
10	2	–	27.98%	52.61%	100.0%
10	3	–	24.32%	52.61%	86.9%
10	5	–	19.92%	52.61%	71.2%
15	2	–	26.78%	56.65%	95.7%
20	2	–	25.73%	59.70%	91.9%
30	2	–	23.96%	64.18%	85.6%
10	2	1.25	23.79%	52.61%	85.0%
10	2	1.50	23.01%	52.61%	82.2%
10	2	2.00	21.60%	52.61%	77.2%

[†]Share of baseline-typed users retaining their label.

Assignment Stability

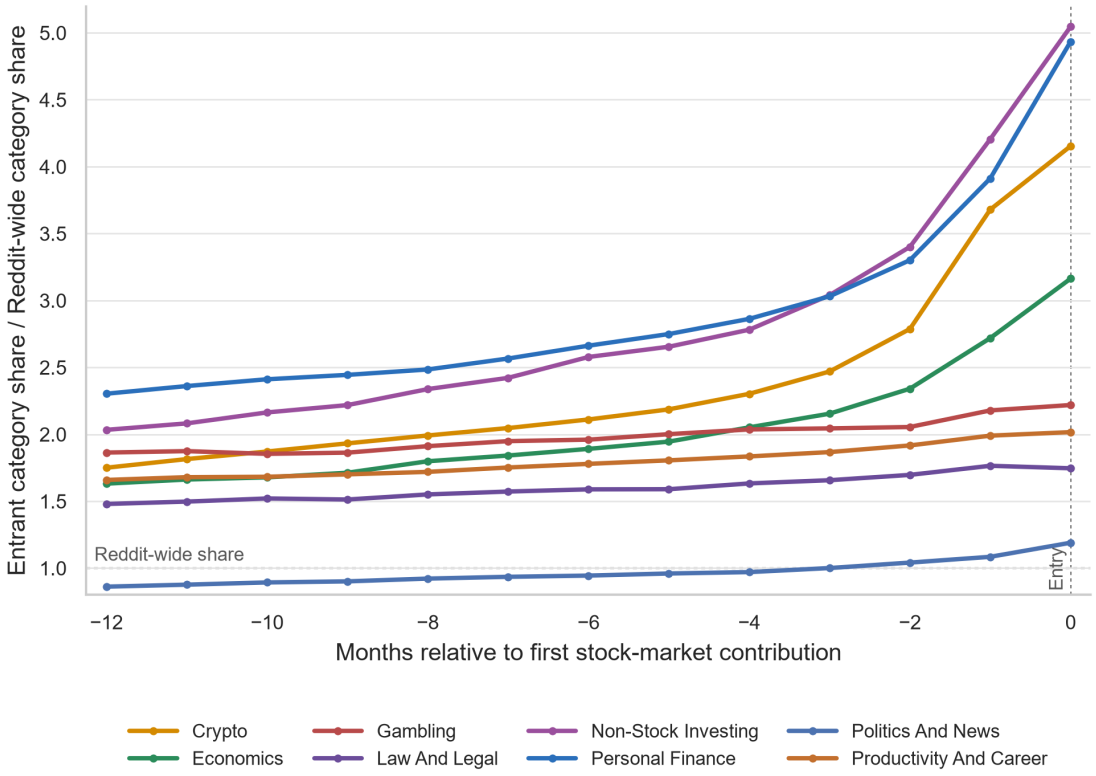
Two diagnostics bear on the hard-assignment rule. The first is the gap between a user's top two standardized category scores, which measures how close the argmax came to selecting a different label. The second splits each user's pre-entry history at random into two halves, runs the identical assignment procedure on each half, and reports how often the two labels agree.

Table 14: Assignment Stability by Pre-Entry Type

Type	Users	Median score gap	Near-tie share	Split-half agreement
Crypto	157,237	16.78	2.6%	93.3%
Economics	167,142	5.88	5.8%	80.4%
Gambling	41,470	21.77	1.1%	92.0%
Law and Legal	115,001	5.92	5.5%	80.0%
Non-Stock Investing	60,550	16.91	1.7%	86.6%
Personal Finance	301,398	7.29	4.6%	87.1%
Politics and News	525,201	5.26	7.6%	93.1%
Productivity and Career	285,568	5.87	6.2%	86.7%

Notes: The table covers eight of the nine substantive types, together 1,653,567 entrants under the baseline ten-post and two-winning-post rule. Home and Housing is absent from the diagnostic file and is not shown. The median score gap is the difference between a user's largest and second-largest standardized excess-activity score S_{ic}. The near-tie share is the fraction of users for whom that gap falls below 0.50. Split-half agreement divides each user's pre-entry category counts into two independent halves by a Binomial$(n_c, 0.5)$ draw with a fixed seed, which is equivalent to randomly assigning the user's individual posts to two halves, then runs the identical assignment on each half using the same Reddit-wide benchmark. The column reports the share of users whose two half-history labels coincide, conditional on both halves yielding a substantive label. Because each half uses half the history, these rates understate the reliability of the full-history labels used throughout the paper. Agreement across all users including the two residual groups is 89.3 percent.

Figure 9: Activity Around Stock-Market Entry, Conditional Sample



Notes: This is the earlier version of Figure 1, which additionally required users to be active outside stock-market subreddits in the entry month. That condition selects on behavior realized at the entry date, so the main text uses the unconditioned version. The conditional sample covers 2,883,824 users against 3,224,789 in the main figure. The pre-entry gradients are positive and significant in both.

D Taxonomy Categories: Coverage, Pre-Entry Trends, and Reliability

D.1 Taxonomy Counts and Confidence

Table 15 reports the number of subreddits, post prevalence, and average classifier confidence for each taxonomy category.

Table 15: Subreddit taxonomy counts, post prevalence, and user activity

Category	N	Posts (%)	Users ever (%)	Posts/active month	Posts/subreddit	Mean conf.
NSFW and porn	7,428	8.51%	24.2%	14.2	274,953	0.988
Gaming	5,134	16.36%	30.2%	12.0	764,650	0.987
Media entertainment	4,565	10.33%	26.1%	6.5	543,048	0.972
Humor and memes	1,834	7.57%	20.0%	7.2	991,017	0.949
Technology	1,785	3.35%	17.0%	5.0	450,005	0.973
Places and geographies	1,505	5.42%	16.2%	6.3	864,864	0.970
Other hobbies	1,412	2.83%	13.7%	7.0	481,145	0.952
Sex and sexuality	1,381	1.64%	11.6%	7.2	284,947	0.966
Internet and social media	1,190	2.01%	12.2%	6.3	405,724	0.922
Politics and news	1,182	4.82%	10.1%	9.1	979,155	0.929
Sports	1,148	6.53%	8.8%	14.6	1,366,012	0.981
Health and fitness	903	1.47%	9.2%	4.8	391,253	0.965
Automotive and transportation	879	1.56%	7.3%	6.2	424,772	0.989
Nature and animals	862	1.59%	12.5%	5.1	442,808	0.982
Lifestyle and fashion	817	1.35%	10.4%	6.2	396,338	0.951
Education	775	1.45%	11.2%	4.3	448,147	0.971
Art, design, and photography	687	1.61%	10.9%	4.0	562,314	0.945
General advice and discussion	683	7.82%	20.0%	7.6	2,746,454	0.830
Identity and social groups	599	2.12%	9.4%	7.8	849,482	0.937
Productivity and career	550	1.01%	9.5%	5.5	439,102	0.952
Food and drink	501	0.82%	7.0%	4.3	395,076	0.981
Drugs and substances	472	1.01%	4.9%	7.1	512,728	0.979
Crypto	462	0.75%	3.8%	10.5	389,004	0.989
Spirituality and philosophy	402	0.96%	4.9%	6.8	573,653	0.961
Relationships and family	397	1.69%	11.8%	5.1	1,023,978	0.955
Mental health support	326	0.81%	6.4%	4.3	597,725	0.967
Stock market	320	1.01%	2.4%	12.1	753,906	0.973
Personal finance	286	0.66%	5.1%	4.6	552,238	0.948
Other unknown	270	0.43%	3.9%	5.4	378,145	0.552
Science	265	0.46%	4.0%	4.2	417,819	0.928
Home and housing	171	0.35%	4.1%	4.6	490,212	0.958
History	168	0.35%	3.2%	3.7	495,786	0.942
Economics	154	0.33%	3.8%	5.3	518,474	0.915
Law and legal	150	0.29%	3.3%	5.4	464,084	0.964
Travel	149	0.24%	4.3%	4.1	389,651	0.974
Self improvement	76	0.22%	4.0%	3.4	687,861	0.907
Gambling	68	0.14%	0.8%	9.0	499,131	0.968
Non-stock investing	44	0.13%	0.9%	4.5	696,630	0.920

Notes: The table covers 40,000 uniquely classified subreddits. Posts (%) uses all observed posts as its denominator. Users ever (%) is the percentage of all observed Reddit users with at least one post in the category. Posts/active month is total category posts divided by the estimated number of active user–month observations in the category, where a user–month is active if the user makes at least one post in that category during that calendar month. Thus, it measures the average number of posts per active user per month, not posts per active calendar month.

D.2 Pre-Entry Trends by Category

This appendix reports the full set of category-specific pre-entry trends summarized in the main text. The sample is the same cohort used in Figure 1: users whose first contribution is to any stock-market destination, who made at least ten non-stock posts in the twelve months before entry, and who were active outside stock-market subreddits in the entry month. This yields 2,883,824 users. Because users appear only in months with observed non-stock activity, the resulting panel is unbalanced.

For each non-stock category c, we estimate a within-user linear trend in excess attention over the twelve months before entry. Let $\tau \in \{-12, \ldots, -1\}$ index event time, with $\tau = -1$ denoting the month immediately preceding entry, and let $\text{Proximity}_{i\tau} = \tau + 12$ increase linearly as entry approaches. The dependent variable, $y_{i\tau c}$, is user i's activity share in category c in event month τ, net of the platform-wide share of category c in the same calendar month; stock-market activity is excluded from both shares. We estimate

$$y_{i\tau c} = \alpha_i + \lambda_{m(i,\tau)} + \beta_c \, \text{Proximity}_{i\tau} + \varepsilon_{i\tau c},$$

where α_i are user fixed effects and $\lambda_{m(i,\tau)}$ are calendar-month fixed effects, so β_c is identified from within-user changes in category c activity as entry approaches, net of aggregate seasonal patterns. We interpret β_c as the average monthly increase in excess attention to category c over the year before entry.

Table 16 reports β_c for every category, with standard errors clustered by user. To make magnitudes comparable across categories of very different overall size, coefficients are scaled by each category's mean platform-wide prevalence.

Table 16: Relative Pre-Entry Pathway Slopes by Category

Category	Monthly slope	Category	Monthly slope
Personal Finance	12.709^{***} (0.277)	Other Hobbies	-0.187^{***} (0.070)
Non-Stock Investing	7.132^{***} (0.575)	Media and Entertainment	-0.329^{***} (0.027)
Economics	4.020^{***} (0.262)	Spirituality and Philosophy	-0.411^{***} (0.100)
Crypto	3.148^{***} (0.282)	Self-Improvement	-0.431 (0.277)
Politics and News	2.310^{***} (0.044)	Technology	-0.490^{***} (0.084)
Home and Housing	1.828^{***} (0.238)	Lifestyle and Fashion	-0.499^{***} (0.098)
Law and Legal	1.704^{***} (0.240)	Gaming	-0.516^{***} (0.021)
Travel	1.630^{***} (0.308)	Internet and Social Media	-0.521^{***} (0.058)
Places and Geographies	1.513^{***} (0.052)	Art, Design, and Photography	-0.675^{***} (0.091)
Automotive and Transportation	1.333^{***} (0.114)	Education	-0.900^{***} (0.117)
Sports	1.137^{***} (0.040)	Nature and Animals	-0.912^{***} (0.086)
Productivity and Career	1.005^{***} (0.150)	Health and Fitness	-1.005^{***} (0.113)
Gambling	0.547 (0.444)	Drugs and Substances	-1.066^{***} (0.129)
Science	0.323^{*} (0.193)	Relationships and Family	-1.145^{***} (0.076)
History	0.295^{*} (0.163)	NSFW and Porn	-1.726^{***} (0.014)
Other/Unknown	0.234 (0.157)	Sex and Sexuality	-1.879^{***} (0.047)
Food and Drink	0.145 (0.139)	Mental Health and Support	-2.159^{***} (0.121)
Humor and Memes	0.103^{***} (0.032)		
General Advice and Discussion	-0.052 (0.038)		
Identity and Social Groups	-0.101^{**} (0.050)		

Notes: Each entry reports the all-users monthly slope in excess attention, scaled by the category's typical platform prevalence; standard errors are in parentheses. A positive coefficient indicates that attention to the category rises as stock-market entry approaches. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D.3 Human and LLM Annotation Agreement

We assess the reliability of the main subreddit taxonomy in a two-stage validation exercise. In Stage 1, a human annotator assigns a fixed random sample of 250 subreddits to the mutually exclusive categories of the full taxonomy. For each subreddit, the annotator observes its name, the category definitions, and eight randomly sampled submissions, including titles and self-text.

We repeat this task independently using GPT-5.6 Luna, Gemini, OpenAI GPT-5 mini, and DeepSeek. Each model receives the same subreddit-level information, taxonomy definitions, and sampled submissions as the human annotator. This common information set ensures that differences in labels reflect differences across annotators rather than differences in the material available to them.

Table 17 reports pairwise Cohen's kappa coefficients across the 250 sampled subreddits. Agreement is high: the unweighted mean kappa across the ten distinct annotator pairs is 0.795 for the full taxonomy.[17]

Table 17: Pairwise Cohen's kappa: full taxonomy

	Human	GPT-5.6 Luna	Gemini	OpenAI GPT-5 mini	DeepSeek
Human	1.000	0.734	0.733	0.707	0.736
GPT-5.6 Luna	0.734	1.000	0.870	0.810	0.839
Gemini	0.733	0.870	1.000	0.831	0.843
OpenAI GPT-5 mini	0.707	0.810	0.831	1.000	0.848
DeepSeek	0.736	0.839	0.843	0.848	1.000

Notes: Each cell reports Cohen's kappa for the corresponding pair of annotators across 250 subreddits. Cohen's kappa adjusts observed agreement for the agreement expected from the annotators' marginal label distributions. Diagonal entries equal one by construction.

[17] Cohen's kappa is defined for a pair of annotators. We therefore summarize overall agreement using the unweighted mean across the ten distinct pairs formed by the five annotators.

E Text Feature Construction

This appendix documents the transparent first-post features used in Section 3.6. The goal is to test whether user types defined exclusively from pre-entry Reddit activity differ in the structure and content of their first stock-market contribution. These features are not used in the type assignment.

For submissions, the document is the title plus selftext; for comments, it is the body. We normalize whitespace and exclude documents that are empty, `[deleted]`, or `[removed]`. Word counts use alphabetic tokens. Text is lowercased for dictionary matching. We use whole-word and complete phrase matching, with longer phrases placed first within each dictionary. Thus, for example, a match to *interest rates* is not separately counted as *interest rate*. The principal dictionary outcomes are binary indicators equal to one if at least one term or phrase is present. This avoids mechanically inflating dictionary scores for short comments. Matches-per-100-words results are retained in the replication output as a robustness check.

Table 18: Dictionaries Used in the First-Post Semantic Validation

Economic terms.

economy; economic; macro; macroeconomic; inflation; deflation; stagflation; cpi; core cpi; ppi; pce; federal reserve; the fed; fomc; Jerome Powell; interest rates; rate hikes; rate cuts; quantitative easing; quantitative tightening; unemployment; labor market; jobs report; nonfarm payrolls; gdp; recession; soft landing; hard landing; economic growth; monetary policy; fiscal policy; yield curve; treasury yields; business cycle; consumer confidence; housing market; mortgage rates; debt ceiling; stimulus; national debt.

Political terms.

election; president; presidential; congress; congressional; senate; democrat; republican; gop; maga; trump; biden; administration; white house; supreme court; scotus; midterms; primary; government shutdown; executive order; tariff; trade war; sanctions; immigration policy.

Portfolio and household-investing terms.

asset allocation; portfolio allocation; diversification; diversified; rebalance; rebalancing; index fund; index funds; etf; etfs; mutual fund; mutual funds; expense ratio; target date fund; risk tolerance; time horizon; dollar cost averaging; dca; buy and hold; roth ira; traditional ira; 401k; brokerage account; taxable account; retirement account; tax advantaged; tax loss harvesting; cost basis; net worth; emergency fund; compound interest; hsa; 529; fiduciary; financial advisor; fee-only; asset class; S&P 500; spy; vti; voo; bond fund; bonds; total market fund.

Trading and speculation terms.

call option; call options; put option; put options; options trading; strike price; implied volatility; IV

crush; 0dte; leverage; short selling; short squeeze; margin call; day trading; swing trading; stop loss; technical analysis; yolo; gamble; stonks; tendies; diamond hands; paper hands; to the moon; wsb; meme stock; penny stock; theta; delta; gamma; pump and dump; rug pull; loss porn; gain porn; ape.

Crypto terms.

crypto; cryptocurrency; bitcoin; btc; ethereum; eth; blockchain; altcoin; altcoins; stablecoin; defi; nft; coinbase; binance; hodl; satoshi; sats; web3; dao; airdrop; wallet; cold storage; hot wallet; dogecoin; shitcoin; memecoin; solana; sol; xrp; staking; gas fee.

Advice/help phrases.

help; advice; any advice; looking for advice; seeking advice; any feedback; looking for feedback; any suggestions; any recommendations; what do you think; what are your thoughts; should i buy; should i sell; should i hold; should i invest; what should i do; where should i start; am i missing something; does this make sense; is this a good idea; roast my portfolio; sanity check; am I doing this right; is this normal; wwyd.

Question and length.

Question is an indicator equal to one if a document contains at least one `?`. Post length is $\log(1 + \text{alphabetic word count})$.

Dictionary audit.

Before the full analysis, we drew a fixed-seed one-percent Bernoulli sample of first posts and inspected 20 positive matches for each dictionary. The matches were overwhelmingly on-concept. The dictionaries use context-specific phrases where a standalone term would be ambiguous (for example, *margin call* rather than *margin*). The reproducible audit excerpts are retained with the analysis output.

F First-Post Topic Dictionaries and Episode Dates

This appendix records the two inputs to Section 4.7 that rest on a judgement of ours, namely the dictionaries applied to first-post text and the dated episodes used in Panel A of Table 8. Both are fixed before estimation and neither is revised in response to a result. A first post counts as mentioning a topic if the case-insensitive regular expression matches anywhere in its text, where text is the body for a comment and the title concatenated with the self-text for a submission.

Dictionaries are deliberately narrow rather than broad. First-post text is short, the median comment running to fifty-nine characters and the median submission title to thirty-four, so recall is necessarily low and precision matters more than coverage, a false positive on a common word being enough to swamp a real signal.

F.1 Dictionaries for the Dated Design

Panel A uses the six vocabularies below, chosen so that the episodes on the list can be distinguished from one another rather than so that policy discussion is covered exhaustively. The continuous specification reported in Section 4.7 uses the union of the four that are present throughout the sample, namely Tariffs and trade, Federal Reserve and rates, Inflation and Elections, excluding bank failures and the pandemic as episode-specific rather than continuously available.

Tariffs and trade.

```
\btariffs?\b|\btrade war\b|\btrade deal\b|\bimport dut(y|ies)\b|
\bprotectionis|\btrade policy\b
```

Federal Reserve and rates.

```
\bfomc\b|\bfederal reserve\b|\bthe fed\b|
\brate (hike|cut|rise|increase|decision)\b|\bbasis points?\b|\bbps\b|
\bpowell\b|\byield curve\b|\bhawkish\b|\bdovish\b
```

Inflation.

```
\binflation(ary)?\b|\bcpi\b|\bconsumer price\b|\bdeflation\b|
\bstagflation\b|\bcost of living\b
```

Elections.

```
\belections?\b|\btrump\b|\bbiden\b|\bharris\b|\bkamala\b|
\bwhite house\b|\bcongress\b|\bsenate\b|\bpresidential\b
```

Bank failures.

```
\bbank run\b|\bbank failure\b|\bsvb\b|\bsilicon valley bank\b|
\bfirst republic\b|\bcredit suisse\b|\bsignature bank\b|\bfdic\b|
\bbailout\b
```

Pandemic.

```
\bcovid\b|\bcoronavirus\b|\bpandemic\b|\blockdown\b|\bquarantine\b
```

F.2 Dictionaries for the Categorical Design

Panel B uses a separate set of ten dictionaries, written to match the ten named categories of the Baker et al. (2016) uncertainty decomposition rather than any episode we had in mind, which is what makes the pairing a matter of the published decomposition rather than of our judgement. Where a category has a counterpart among the six above, the two dictionaries are related but not identical, the categorical version being written to the scope of the published category and the dated version to the distinctiveness of the episodes.

Three overlaps are recorded here because they were known before estimation and are left unrepaired. Medicare and Medicaid appear in both health care and entitlements, since they appear in both categories as the decomposition defines them. Bank failures sit in the financial regulation dictionary, a failure being what generates regulatory uncertainty. Sanctions appear in national security although they are also a trade instrument. Overlap of this kind attenuates the contrast across topics and therefore works against the estimates reported rather than for them.

Federal Reserve and rates, paired with "2. Monetary policy".

```
\bfomc\b|\bfederal reserve\b|\bthe fed\b|
\bfed\b(?=\s+(chair|meeting|rate|cut|hike|decision|mtg))|
\brate (hike|cut|rise|increase|decision)s?\b|\bbasis points?\b|
\bbps\b|\bpowell\b|\byield curve\b|\bhawkish\b|\bdovish\b|
\bmonetary policy\b|\bquantitative easing\b|\btaper(ing)?\b|
\bqe\d?\b
```

Taxes, paired with "3. Taxes".

```
\btax(es|ation|payer)?\b|\bcapital gains\b|
\btax (cut|hike|rate|bill|reform|break|loss)\b|\birs\b|
\bincome tax\b|\bcorporate tax\b|\bwash sale\b|\bwrite[- ]off\b|
\btax bracket\b
```

Fiscal and stimulus, paired with "4. Government spending".

```
\bgovernment spending\b|\bstimulus\b|\binfrastructure bill\b|
\bspending bill\b|\bappropriations?\b|
\bfiscal (stimulus|policy|spending)\b|
\brelief (package|bill|check)\b|\bgovernment shutdown\b|
\bfederal budget\b|\bomnibus\b
```

Health and pandemic, paired with "5. Health care".

```
\bhealth ?care\b|\bobamacare\b|\baffordable care act\b|\bmedicare\b|
\bmedicaid\b|\bdrug pric(e|ing)\b|\bfda\b|\bcovid\b|\bcoronavirus\b|
\bpandemic\b|\bvaccin(e|es|ation)\b|\blockdown\b|\bquarantine\b
```

Security and conflict, paired with "6. National security".

```
\bmilitary\b|\bdefen[cs]e spending\b|\bukrain(e|ian)\b|
\brussia(n)?\b|\bisrael(i)?\b|\bgaza\b|\biran(ian)?\b|
\bnorth korea\b|\bterroris(m|t)\b|\bsanctions?\b|
\bgeopolitic(al|s)\b|\bnato\b|\bwarfare\b|\binvasion\b
```

Entitlements, paired with "7. Entitlement programs".

```
\bsocial security\b|\bentitlement(s| program)\b|\bpension(s)?\b|
\bwelfare\b|\bunemployment (benefit|insurance)\b|\bmedicare\b|
\bretirement benefit(s)?\b|\bfood stamps\b|\bsnap benefits\b
```

Regulation, paired with "8. Regulation".

```
\bregulat(ion|ory|e|ors|ing)\b|\bderegulation\b|\bantitrust\b|
\bthe sec\b|\bsec (filing|investigation|rule|charge|probe)\b|\bftc\b|
\bcftc\b|\bcompliance\b|\brule change\b|\blobby(ing|ist)\b
```

Banks and supervision, paired with "Financial Regulation".

```
\bdodd[- ]frank\b|\bbasel\b|\bcapital requirement(s)?\b|
\bstress test\b|\bfdic\b|\bbank(ing)? regulation\b|\bbailout\b|
\btoo big to fail\b|\bbank run\b|\bbank failure\b|\bsvb\b|
\bsilicon valley bank\b|\bfirst republic\b|\bcredit suisse\b|
\bsignature bank\b
```

Tariffs and trade, paired with "9. Trade policy".

```
\btariffs?\b|\btrade war\b|\btrade deal\b|\bimport dut(y|ies)\b|
\bprotectionis|\btrade policy\b|\bwto\b|\bnafta\b|\busmca\b|
\bexport control(s)?\b|\btrade barrier(s)?\b
```

Debt and downgrades, paired with "10. Sovereign debt, currency crises".

```
\bdebt ceiling\b|\bsovereign debt\b|\bdefault(ing|s)? on\b|
\bcredit rating\b|\bdowngrade(d)?\b|\beurozone\b|\bcurrency crisis\b|
\bdevaluation\b|\bimf\b|\bdebt crisis\b|\bausterity\b
```

F.3 Episode Dates

Table 19 lists the twenty-two episodes used in Panel A. They are selected for topical distinctiveness rather than sampled for coverage, because the identifying comparison in that design runs across topics on a given day rather than across days, making each topic the control for the others. The list therefore carries no claim of completeness and supports one claim only, which is that first-post content moves with the specific content of a dated episode.

Every date was checked against a primary source in September 2026, the news release for statistical series, the Federal Register or the agency press release for policy actions, and the race call for elections. One relabelling followed. The entry at 4 November 2020 was described as the election result, when it is the day the polls closed with the count still running, the race not being called until the Saturday, and it is now labelled accordingly with 7 November added as a separate entry. In 2016 and 2024 the polls closed in the evening Eastern time and the race was called in the early hours, both of which fall on the same date in Coordinated Universal Time, so those years need only one entry each, leaving 2020 the only election on the list that separates into two.

Three limitations are worth stating. Topic vocabularies are not perfectly disjoint, so a tariff announcement made by a president may move the election dictionary through a name, and the bank failures of March 2023 moved rate expectations sharply, with both contaminations pushing the own-topic estimate toward zero. A comparison window can also contain another entry on the list, which contaminates the control period rather than the response period, and this happens twice: the 7 November 2020 entry has the 4 November entry inside its own comparison window, and the Federal Reserve entry of 15 March 2020 has the pandemic declaration of 11 March inside its own, the first attenuating the estimate and the second inflating it. The 2016 and 2018 dates also fall in the thin part of the panel, where daily entrants number in the low hundreds, leaving the early topics far less precisely estimated than the 2022 to 2025 ones.

Table 19: Dated Episodes Used in Panel A of Table 8

Date	Topic	Episode
2018-03-08	Tariffs and trade	Section 232 steel and aluminium proclamation
2018-06-15	Tariffs and trade	Twenty-five percent duty on Chinese goods announced
2019-08-01	Tariffs and trade	Ten percent duty on the remaining Chinese imports announced
2025-04-02	Tariffs and trade	Reciprocal tariff announcement
2025-04-09	Tariffs and trade	Ninety-day pause announced
2020-03-15	Federal Reserve and rates	Unscheduled emergency cut to the zero lower bound
2022-06-15	Federal Reserve and rates	First seventy-five basis point increase of the cycle
2022-09-21	Federal Reserve and rates	Third consecutive seventy-five basis point increase
2024-09-18	Federal Reserve and rates	First cut of the easing cycle, fifty basis points
2021-05-12	Inflation	April CPI at 4.2 percent, above consensus
2022-06-10	Inflation	May CPI at 8.6 percent, above consensus
2022-07-13	Inflation	June CPI at 9.1 percent, the cycle peak
2023-03-10	Bank failures	Silicon Valley Bank closed by regulators
2023-03-19	Bank failures	UBS agrees to acquire Credit Suisse
2023-05-01	Bank failures	First Republic seized and sold to JPMorgan
2016-11-09	Elections	Presidential election, polls closed and race called
2020-11-04	Elections	Presidential election, polls closed, count still underway
2020-11-07	Elections	Presidential election called for Biden
2024-11-06	Elections	Presidential election, polls closed and race called
2020-02-24	Pandemic	Global selloff begins
2020-03-11	Pandemic	World Health Organization declares a pandemic
2020-03-16	Pandemic	Largest single-day decline of the episode

Notes: Twenty-two episodes across six topics. Each contributes a comparison window of seven to three days before the date and a response window of the date through two days after. Dates are in Coordinated Universal Time, matching the dating of entry throughout the paper, and each was checked against a primary source in September 2026. For statistical series the date is the date of the news release itself, at 8:30am Eastern for the Consumer Price Index, rather than the date of any summary article published afterwards.

G Sensitivity of the Entry Estimates

Table 20 collects the checks discussed in Section 4.8 for the two stages of the entry analysis. Each column changes one feature of the specifications in Tables 6 and 7, and the first column reproduces those tables, so every estimate in the table is directly comparable with the main results. The columns fall into three groups, the first varying the trend control from none to month-of-sample fixed effects, the second removing the pandemic year, the tariff period of late 2024 and 2025, or both, and the third adding Reddit-wide shares of activity by category as controls for the platform's own topical composition.

Table 20: Sensitivity of the Entry Estimates to Trend Controls, Episodes and Reddit's Own Composition

	Trend control				Episode excluded			Platform shares	
	Main	No trend	Quartic trend	Month FE	Drop 2020	Drop 24Q4 to 25Q2	Drop both	Platform month	Platform $t-12$
Panel A. Share among entrants with a legible past, on log EPU$_t$									
Politics and News	0.558***	0.176	1.167***	0.276**	0.594***	0.403***	0.427***	0.655***	0.502***
	(3.7)	(0.7)	(5.2)	(2.5)	(3.5)	(3.2)	(3.0)	(4.4)	(3.8)
Personal Finance	−0.285***	−0.186	−0.420***	−0.080	−0.278***	−0.199*	−0.153	−0.276***	−0.199**
	(−3.0)	(−1.6)	(−2.9)	(−1.2)	(−2.8)	(−1.9)	(−1.4)	(−3.0)	(−2.3)
Non-Stock Investing	−0.057**	−0.002	−0.092***	−0.032	−0.065**	−0.053**	−0.057**	−0.059***	−0.041*
	(−2.5)	(−0.1)	(−3.3)	(−1.5)	(−2.6)	(−2.2)	(−2.1)	(−2.6)	(−1.9)
Gambling	−0.040**	−0.043*	−0.012	−0.004	−0.034*	−0.017	−0.006	−0.042**	−0.030*
	(−2.1)	(−1.9)	(−0.5)	(−0.3)	(−1.8)	(−0.8)	(−0.3)	(−2.5)	(−1.7)
Crypto	−0.165*	−1.054***	−0.632***	−0.039	−0.175*	−0.195*	−0.220	−0.168**	−0.105
	(−1.8)	(−7.6)	(−3.9)	(−0.9)	(−1.7)	(−1.8)	(−1.5)	(−2.3)	(−1.6)
Joint $\chi^2(9)$, log EPU$_t$	25.7***	112.9***	77.3***	8.0	24.1***	18.0**	19.6**	31.8***	18.9**
	[0.002]	[0.000]	[0.000]	[0.536]	[0.004]	[0.035]	[0.021]	[0.000]	[0.026]
Joint $\chi^2(9)$, log EPU$_{t-1}$	18.5**	119.9***	65.8***	7.2	16.8*	12.9	12.7	18.9**	18.5**
	[0.030]	[0.000]	[0.000]	[0.612]	[0.053]	[0.167]	[0.175]	[0.026]	[0.030]
Panel B. Share among entrants with a legible past, on $\lvert r_t^m \rvert$									
Productivity and Career	−0.063***	−0.088	−0.104***	−0.061***	−0.079***	−0.073***	−0.080*	−0.043**	−0.051***
	(−3.4)	(−0.8)	(−2.8)	(−3.8)	(−3.1)	(−2.9)	(−1.7)	(−2.5)	(−3.3)
Joint $\chi^2(9)$	30.9***	54.6***	36.7***	38.2***	34.8***	36.9***	17.4**	53.4***	39.9***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.042]	[0.000]	[0.000]
Panel C. Share of all entrants with no usable history									
On log EPU$_t$	−0.906**	−1.776**	−1.856***	−0.824**	−1.217***	−0.629	−1.002*		
	(−2.0)	(−2.5)	(−3.4)	(−2.1)	(−2.6)	(−1.3)	(−1.9)		
On log EPU$_{t-1}$	−0.808**	−1.412**	−1.473***	−0.579**	−1.114***	−0.693*	−1.027***		
	(−2.1)	(−2.2)	(−3.6)	(−2.5)	(−3.0)	(−1.7)	(−2.7)		
Days, Panels A and B	3,836	3,836	3,836	3,836	3,470	3,563	3,197	3,836	3,836
Days, Panel C	3,895	3,895	3,895	3,895	3,529	3,622	3,256		

Notes: Each column re-estimates the specifications of Table 6 (Panel C) and Table 7 (Panels A and B) under one change, and the Main column reproduces those tables. The daily panel of first posts in stock-market subreddits covers 3 May 2015 to 30 December 2025; Panels A and B exclude January and February 2021 in every column, as in Table 7. Outcomes in Panels A and B are a category's share, in percentage points, of that day's entrants with a legible pre-entry history, defined as at least ten posts in the twelve months before entry, over the nine pre-entry types and a tenth category for legible entrants whose dominant interest lies outside them; only selected categories are shown, and the joint tests cover all ten, whose coefficients sum to zero in every column. The Panel C outcome is the share of all entrants with no usable history. All four conditions (log EPU and its lag, log VIX and the absolute S&P 500 return) enter every regression, standardized to unit standard deviation, with day-of-week fixed effects and weights equal to the relevant daily entrant count. *Trend control:* the Main column uses year fixed effects; No trend uses none; Quartic trend replaces year effects with a fourth-order polynomial in calendar time; Month FE replaces them with month-of-sample fixed effects. *Episode excluded:* drops calendar 2020, 1 October 2024 to 30 June 2025, or both, with year fixed effects. *Platform shares:* adds, with year fixed effects, the log share of all posts on the platform falling in each of the nine type categories, measured in the calendar month (Platform month) or averaged over the twelve preceding months (Platform $t-12$), the same nine controls entering every equation; these are not estimated for Panel C. Newey-West t-statistics with a thirty-day bandwidth are in parentheses; joint tests use the analytic cross-equation covariance at rank nine, with p-values in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

H Robinhood Holdings Around Attention Spikes

Robintrack collected the popularity figures that Robinhood published for each security, the number of Robinhood accounts holding it, several times a day until Robinhood withdrew them in August 2020. We take the last reading of each calendar day and carry it forward over days without one, giving a daily panel of 811 securities from May 2018 to August 2020, and we keep the attention spikes of Section 5 whose estimation window, days −120 to −11, and outcome window both fall inside that period. Holder counts are a stock rather than a flow, measuring how many accounts hold a security rather than how much they trade, which makes their growth a direct measure of retail position taking though not of its size.

Table 21 reports the first stage of Section 5.7, relating log window mentions to abnormal holder growth over days 0 to 5 and 0 to 20. Table 22 reports the composition stage for days 0 to 5, with every coefficient, the joint test, the gap between *Gambling* and *Personal Finance*, and the minimum detectable effect. Three of the twenty-four type coefficients are significant at the 5 percent level, two of them in the column without fixed effects, and the joint test does not reject in any of the three columns.

Table 21: Attention Spikes and Robinhood Holder Growth

	Days 0 to 5		Days 0 to 20	
	(1)	(2)	(3)	(4)
Log mentions	0.0085***	0.0165***	0.0131	0.0062
	(0.0026)	(0.0034)	(0.0084)	(0.0069)
Event-day fixed effects	Yes	Yes	Yes	Yes
Security fixed effects	No	Yes	No	Yes
Events	1,925	1,925	1,925	1,925
Securities	236	236	236	236

Notes: The sample is the attention spikes of Section 5 on securities covered by Robintrack, which recorded the number of Robinhood accounts holding each of 811 securities between 2 May 2018 and 13 August 2020. Events are kept if their estimation window, days −120 to −11, and their outcome window fall inside that period, and if the security is not one of the 96 non-security symbols of Appendix J.2. The daily holder count is the last reading of each calendar day, carried forward over days without a reading. The outcome is abnormal holder growth, the change in log holders from day −1 to day 5 (columns 1 and 2) or day 20 (columns 3 and 4), less the security's average daily log growth over the estimation window times the length of the outcome window, all in calendar days. Mean abnormal growth over days 0 to 5 is 0.0284, and the median event starts from 15,126 holders. Log mentions is the log of total mentions in the event window. Standard errors in parentheses are two-way clustered by security and event day. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 22: Entry-Type Composition and Robinhood Holder Growth

	(1)	(2)	(3)
Personal Finance share	−0.0347**	−0.0266	−0.0115
	(0.0165)	(0.0168)	(0.0109)
Productivity and Career share	0.0004	−0.0321	−0.0475***
	(0.0254)	(0.0315)	(0.0170)
Home and Housing share	−0.0242	−0.0110	0.0063
	(0.0358)	(0.0381)	(0.0234)
Crypto share	−0.0348**	−0.0354	−0.0120
	(0.0176)	(0.0313)	(0.0134)
Economics share	−0.0029	0.0023	0.0261*
	(0.0161)	(0.0182)	(0.0137)
Law and Legal share	0.0257	−0.0142	−0.0182
	(0.0340)	(0.0348)	(0.0258)
Non-Stock Investing share	−0.0541	−0.0201	−0.0150
	(0.0381)	(0.0266)	(0.0263)
Gambling share	0.0186	−0.0172	−0.0141
	(0.0252)	(0.0290)	(0.0226)
Classified share	0.0435	0.0109	0.0073
	(0.0303)	(0.0302)	(0.0226)
Log mentions	0.0124***	0.0096***	0.0140***
	(0.0040)	(0.0029)	(0.0028)
Joint test, eight type shares (p)	0.445	0.871	0.156
Gambling − *Personal Finance*	0.0533	0.0094	−0.0027
95% confidence interval	[−0.003, 0.109]	[−0.047, 0.065]	[−0.051, 0.046]
Minimum detectable effect	0.0798	0.0799	0.0695
Event-day fixed effects	No	Yes	Yes
Security fixed effects	No	No	Yes
Events	1,708	1,708	1,708
Securities	212	212	212

Notes: The sample is the events of Table 21 with at least one classified mention in the event window. The outcome is abnormal holder growth over days 0 to 5, defined as in Table 21, with a mean of 0.0279 and a standard deviation of 0.1820. Type shares are the shares of classified mentions in the event window written by authors of each type, with *Politics and News* as the omitted reference, and the classified share is the share of all window mentions whose author receives a type, as in Panel B of Table 10. The joint test is a Wald test that all eight type coefficients are zero. The minimum detectable effect is the smallest difference between the *Gambling* and *Personal Finance* coefficients that a two-sided test at the 5 percent level would detect with 80 percent power. An effect of the same size, in standard deviations of the outcome, as the across-security gap in absolute abnormal returns in Section 5 would be 0.0637. Standard errors in parentheses are two-way clustered by security and event day. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

I The Entry Series and Retail Brokerage Activity

Our entry measure counts first posts in stock-market communities on a single platform whose users are not a representative sample of households, raising the question of whether the months in which entry is unusually high are unusual for retail investors generally or only for Reddit. We answer it with two series that are measured outside Reddit and are unaffected by anything happening on it, comparing each with monthly entry aggregated from the daily panel of Section 4.

Interactive Brokers discloses daily average revenue trades each month, giving a measure of trading intensity among active traders, and Charles Schwab discloses new brokerage accounts, giving a measure of account formation among a broader retail population.[18] The two answer different questions, since placing a trade is a decision an existing customer can take within the month while opening an account is slower, and the second is the closer analogue of entry into a new community.

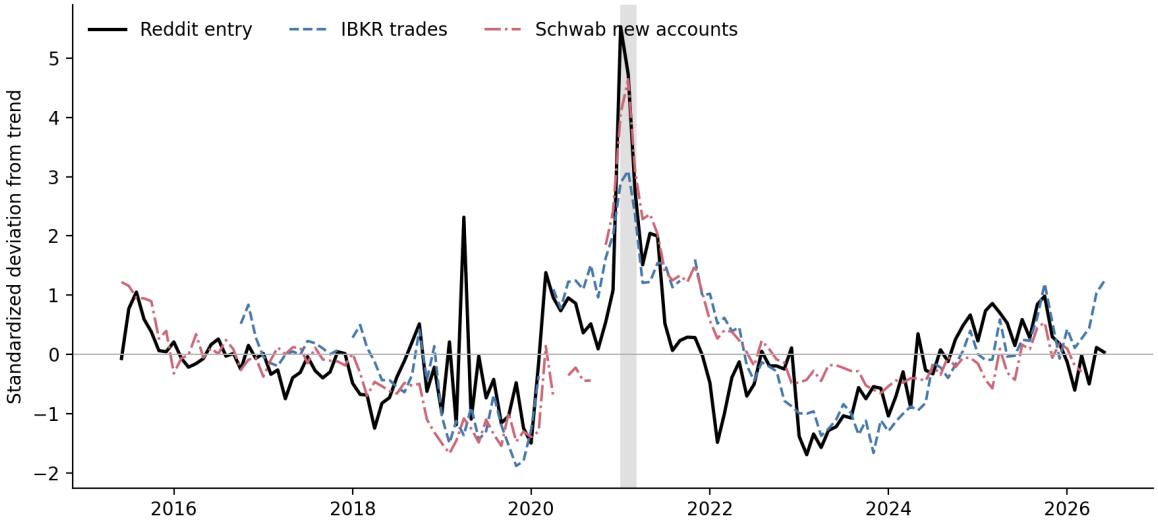
Figure 10 plots all three series as deviations from a quadratic trend and calendar-month fixed effects, standardized so their amplitudes are comparable, and the common movement is visible without any estimation. All three rise around the onset of the pandemic in early 2020, peak together in the GameStop episode of early 2021, fall through 2022 and 2023, and turn up again from 2024. The correspondence is closest in the episodes and looser in ordinary months, which is what one would expect of series measured on different populations by different means.

Table 23 puts numbers to it. The correlation of the detrended series with Reddit entry is 0.69 for trading and 0.70 for account openings, and 0.62 and 0.55 with the two GameStop months excluded, so the comovement is not the artifact of a single episode. Twelve-month log differences, which remove any slow-moving common drift the trend terms leave behind, give correlations of 0.64 and 0.76. Regressing each external series on log entrants with the same trend and seasonal controls gives elasticities of 0.32 and 0.45, both significant at well below the one percent level and both surviving the exclusion of the GameStop months.

We read this as evidence that the platform's timing is not idiosyncratic, since the months when unusually many users start talking about stocks on Reddit are the months when unusually many people trade and open brokerage accounts elsewhere. It does not show that Reddit discussion causes brokerage activity, and nothing here identifies a direction, because the most natural reading is that both respond to the same conditions. Nor does it repair

[18] Both come from company disclosures, the first from monthly press releases and the second from the trailing tables embedded in quarterly filings. Schwab books accounts acquired with a business as new brokerage accounts, so the USAA and TD Ameritrade acquisitions register as 1.2 million and 14.5 million account openings in May and October 2020. We treat those two months as missing.

Figure 10: Reddit Entry, Trading and Account Openings



Notes: Each series is monthly and in logs, shown as the residual from a regression on a quadratic trend and calendar-month fixed effects, then standardized by its own standard deviation so that the three amplitudes are comparable. Reddit entry is the monthly count of first posts in stock-market communities, aggregated from the daily panel of Section 4, June 2015 to June 2026. IBKR trades are Interactive Brokers' daily average revenue trades, October 2016 to June 2026. Schwab new accounts are Charles Schwab's new brokerage accounts, June 2015 to March 2026, with May and October 2020 omitted as acquisition months. The shaded band marks January and February 2021, the GameStop episode. Table 23 reports the corresponding correlations and elasticities.

the representativeness limitation, since these are aggregate national series and cannot speak to whether the individuals in our sample resemble retail investors in general. What it does establish is that the variation our entry measure records is variation in retail engagement at large, and not a feature of one online community.

Table 23: Reddit Entry and Retail Brokerage Activity

	IBKR trades (1)	Schwab new accounts (2)
Correlation with Reddit entry		
Deviations from trend	0.692	0.701
excluding January and February 2021	0.621	0.549
Twelve-month log differences	0.637	0.761
Elasticity with respect to Reddit entry		
Log entrants	0.319***	0.453***
	(0.038)	(0.083)
excluding January and February 2021	0.348***	0.373***
	(0.071)	(0.129)
Months	114	128
Sample window	2016-10 to 2026-06	2015-06 to 2026-03

Notes: Reddit entry is the monthly count of first posts in stock-market communities, aggregated from the daily panel of Section 4, covering 2015-06 to 2026-06. IBKR trades are Interactive Brokers' daily average revenue trades and Schwab new accounts are Charles Schwab's new brokerage accounts, both from company disclosures and both monthly. Each series enters in logs. Deviations from trend are residuals from a regression on a quadratic trend and calendar-month fixed effects, estimated separately for each series. The elasticity is the coefficient on log entrants in a regression of the external series on the entry series with the same trend and calendar-month controls, with Newey-West standard errors allowing for three months of autocorrelation in parentheses. The two GameStop months are excluded where indicated, since they are the largest observation in every series. Sample windows differ because each source has its own coverage. Schwab reports accounts acquired with a business as new brokerage accounts, so May 2020, the USAA acquisition, and October 2020, the TD Ameritrade acquisition, are treated as missing rather than as account openings. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

J Ticker Mention Extraction

This appendix describes how we identify mentions of individual securities in stock-market discussion.

J.1 Source Text and Matching Rule

We process every submission and comment posted in the 320 stock-market subreddits between January 2015 and July 2026. For each post we concatenate the title, the self-text, and the comment body into a single string. We then apply the following rule.

1. We find cashtags matching `$[A-Z]{1,5}`, bounded so that the symbol is not part of a longer alphanumeric token. We retain a cashtag only when its symbol appears in the ticker inventory.
2. If a post contains at least one valid cashtag, we retain those symbols and do not search that post for bare tickers.
3. Otherwise we find bare uppercase tokens matching `[A-Z]{1,5}`. We retain a token only when it appears in the ticker inventory and is absent from the list of the 10,000 most common English words.
4. We deduplicate symbols within a post.

The cashtag priority in step two matters. Users who write $TSLA in a post are signaling the security explicitly, and searching the same post for bare tokens would add noise without adding information. The common-word exclusion in step three removes the largest source of false positives, since many English words are also valid ticker symbols.

The procedure yields 44.7 million post-ticker mentions. Of these, 3.8 million come from cashtags and 40.8 million from bare uppercase tokens. We attach to each mention the author and the posting date, which allows us to merge mentions to the pre-entry user types defined in Section 3. We exclude mentions from accounts flagged by the screening procedure in Appendix K.

J.2 Screening Non-Securities

Bare uppercase tokens can match acronyms that are not securities. This is a serious concern in our setting. Financial discussion is dense with abbreviations, and many of them are also valid ticker symbols. The token `ETF` is a ticker, as are `ATH`, `EOD`, `DCA`, `RSI` and `MACD`. So are `CDC`, `OPEC` and `ESG`. The problem is not merely that these add noise. Different user types use different vocabulary, so uncorrected collisions could produce apparent differences in the securities that types discuss when the underlying difference is only in how they write.

We therefore screen symbols on evidence rather than on judgment about the string. We rank symbols by the number of bare-token mentions they receive and take the 200 largest, which together account for 64 percent of all classified mentions. For each symbol we draw usage snippets from the corpus, retaining a snippet only when our matching rule would have recorded that mention. We then read the snippets and record whether the token is used as a security. We repeat the exercise for every symbol that appears in the enrichment tables, so that each reported cell is validated directly.

The screen removes 96 symbols and 22.7 percent of classified mentions. The largest are `DRS` (1.29 million mentions, direct registration), `ETF` (0.84 million), `WSB` (0.79 million, the subreddit), `ATH` (0.31 million, all-time high) and `BTC` (0.30 million, which is an asset but not an equity). The remainder are trading jargon, broker and exchange names, and political abbreviations.

Two checks indicate that the screen removes noise rather than signal. First, the removed symbols are close to proportionally distributed across types, with type shares between 0.87 and 1.29 times their corpus shares. Second, the results in Section 3.5 are almost unchanged by the screen. Normalized mutual information between type and security falls from 32.4 times chance on the unscreened panel to 27.5 times on the screened panel, even though the screen removes more than a fifth of the data.

J.3 Limitations

Residual collisions. The screen validates the 200 highest-volume symbols and every symbol appearing in our tables, with one exception. PK is too infrequent for our sampling procedure to return usage snippets, so we retain it on the basis that it is a listed company with no common non-financial reading. Symbols outside these groups are not individually validated. Any remaining collisions are small in volume and, by the proportionality check above, unlikely to be concentrated in particular types.

Inventory coverage. We match against a contemporaneous ticker dictionary rather than a date-aware inventory that records which symbols were live in each month. A symbol that was delisted early in the sample and reassigned later is therefore matched throughout. This inflates mention counts, and it inflates them unevenly across symbols. For this reason we base our ticker-level results on standardized enrichment rather than on mention levels. Enrichment compares a type's share of mentions of a given security with that security's share of all mentions, so a per-symbol over-match scales the numerator and the denominator together and largely cancels.

Cashtags are not a cleaner alternative. A natural robustness check restricts attention

to cashtag mentions, which cannot collide with ordinary words. We do not use it as our main specification. Cashtag usage is concentrated in speculative communities and carries its own false positives in the form of joke symbols, so the restriction changes which users are represented rather than isolating a cleaner measurement of the same population.

K Author and Archive Screening

Reddit data carry three risks that could generate spurious structure. Automated accounts post at machine scale and could dominate mention counts. Duplicate accounts operated by one person, or coordinated posting rings, could make a handful of actors look like many users. Gaps in the public archive could mis-date entry and deflate mention counts on affected days. This appendix documents the screens we apply for each.

K.1 Automated and Duplicate Accounts

We run the screen on the full first-post ledger of 6,871,602 accounts, of which 6,804,621 have a first stock-market contribution in the January 2015 to July 2026 window the paper analyzes. The remaining 66,981 accounts entered before 2015 and are screened but not analyzed. Percentages below are shares of the 6,871,602 screened accounts.

Name-based. We flag accounts whose names follow common bot conventions, together with a list of widely deployed Reddit bots such as *AutoModerator*, *VisualMod* and *RemindMeBot*. This flags 9,705 accounts, 0.14 percent of the ledger.

Duplicate text. We hash the normalized text of each account's first contribution and flag texts of at least twenty characters shared by five or more distinct accounts. Identical first posts across many accounts indicate copy-paste farms, karma rings or a single operator. This flags 89,593 accounts, 1.30 percent.

Hyperactivity. We flag accounts posting more than 11,000 times in the twelve months before entry, a volume far beyond plausible human activity. This flags 1,359 accounts, 0.02 percent.

In total 100,539 accounts are flagged, 1.46 percent of the ledger. We exclude their contributions from every mention-based measure in the paper. The screen is deliberately blunt. It is designed to remove the accounts that could mechanically drive results, not to identify every automated account.

Duplicate text is the binding screen. This is what we would expect if the main risk is coordinated posting rather than conventional bots, since bots rarely make a first contribution to a stock-market community in the way our entry definition requires.

K.2 Archive Coverage

We check the daily mention series rebuilt from the archive for gaps that would distort either entry timing or spike detection.

There are no days with zero mentions anywhere in the January 2015 to July 2026 window.

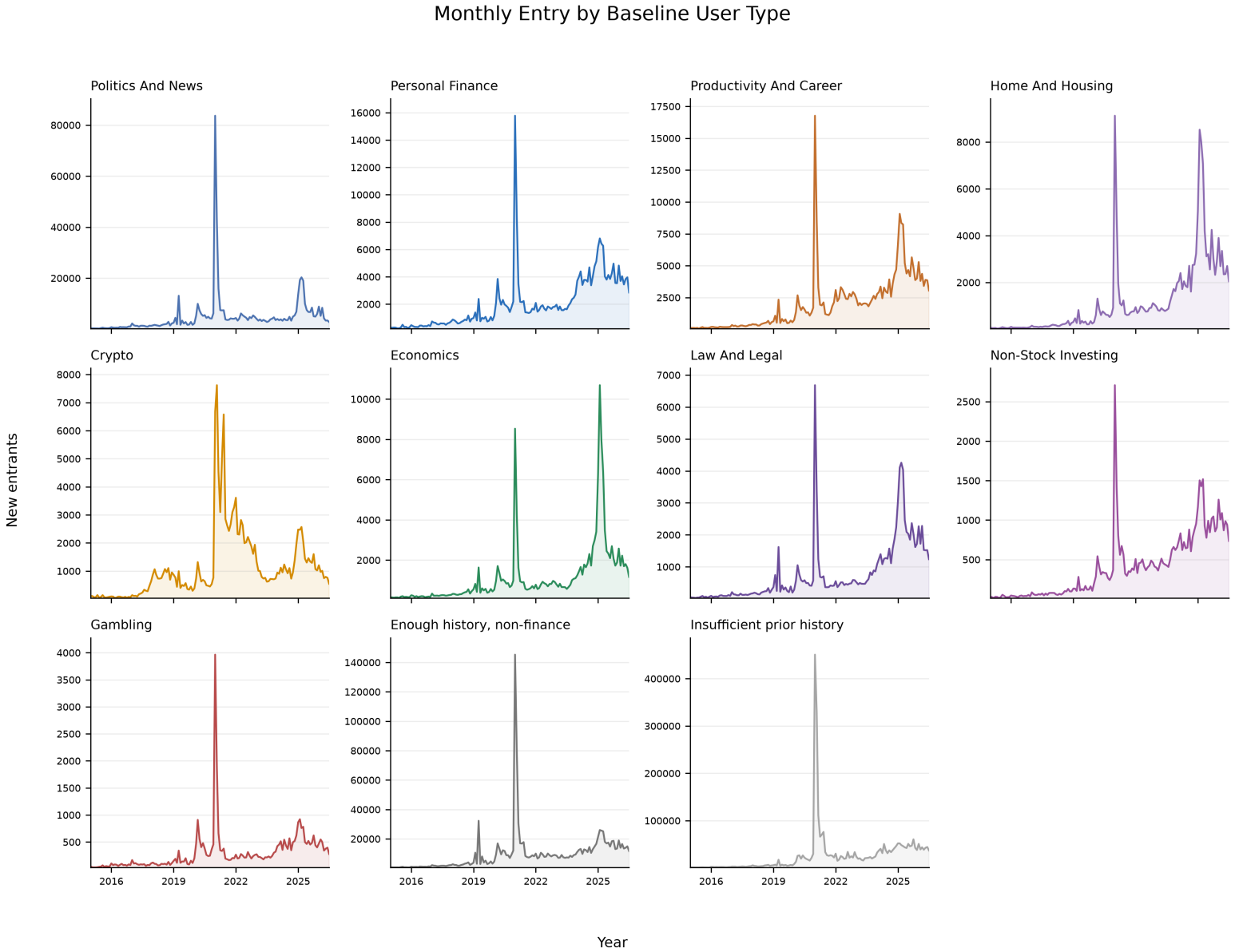
The series shows one anomaly, a depression in activity across March and April 2018. This coincides with a known period of incomplete Pushshift collection and affects both entry counts and mention counts in those two months. It is short relative to the sample and does not overlap the episodes that drive our time-series results.

We take two precautions. First, we re-estimate the uncertainty regressions excluding March, April and May 2018 outright. The aggregate estimate moves from 17.1 percent to 16.7 percent and remains significant at $p = 0.024$, and the test of equal type-specific responses remains significant at $p = 0.006$. The uncertainty results are therefore not driven by the archive discontinuity. Second, our event-study measures are constructed from rolling windows, so a two-month depression affects the estimation window and the event window symmetrically.

L Additional Figures

Figure 11 extends Figure 6 to every user type. The figure uses raw monthly entrant counts and gives each panel its own vertical scale.

Figure 11: Monthly Entry by Baseline User Type



Notes: Each panel plots monthly counts of first-time stock-market entrants assigned to the indicated baseline group from January 2015 through July 2026. The nine substantive types are assigned using pre-entry finance-adjacent activity; the two gray panels contain entrants who do not satisfy the prior-history or finance-concentration requirements. Vertical scales differ across panels. The entry month is excluded from user-type construction.