Quality of Household Finance Advice: The Case of Social Media

This is a preliminary version of the paper

March 2026

Abstract

Household financial decisions are complex, and low financial literacy increases reliance on external guidance, including social media. Despite its growing role—especially for younger and financially constrained users—little is known about the quality and reliability of socially sourced household-finance advice beyond portfolio choice, or how it varies across domains and socioeconomic groups. We conduct a field experiment evaluating tailored social-media advice across six domains (planning, budgeting, saving, investing, retirement, borrowing) using assessments by licensed CFP® professionals, appropriate given that financial advice is a credence good without a unique gold standard. We draw 48 stratified Reddit threads representing realistic retail-consumer situations. Average advice quality is typically moderate, but dispersion varies sharply by domain and subreddit socioeconomic profile. Community endorsement (upvotes) is a weak signal of quality and an unreliable substitute for expert judgment. In contrast, commenter reputation, informational richness, and the advice-seeker's financial sophistication predict higher-quality advice. Counterintuitively, higher-quality and more consistent advice is more prevalent in financially constrained communities, whereas high-net-worth forums exhibit greater noise and downside risk—suggesting relatively better guidance where professional advice is least accessible.

Keywords: Household finance advice, social-media financial advice, advice quality, advice reliability and heterogeneity.

JEL code: D10, D15, G40, G51, G53.

A user's post on Reddit seeking advice: *"I've messed up... Accumulated 22k in credit card debt! Okay so I'll start with the bad news. I made some poor choices and financed my lifestyle during senior year of college + the 3 months after I graduated with credit cards. Accumulated a total of 22k. To be honest, I've been freaking out the past few months and in hindsight can't believe how foolish I was to do this. Regardless, life goes on and I need to deal with this. The good news: I got a job paying 83k a few months ago and have paid down the debt to 16k. Can make additional payments of 1-2k above minimum payments. Even better news: I made the good decision during the prior 4 years to save a lot in my Roth IRA. Meaning in total I have, roughly, 13k in Roth IRA contributions. Should I withdrawal my Roth IRA contributions to pay off a large chunk of this debt? How can I pay this off faster? I'm willing to cut back on my expenses. No interest yet, but in May it will be 20%+."* **January 30, 2023.**

Commenter 1: "*22k on 83k worth of income isn't life ruining luckily.*" **3.3K net upvotes.**

Commenter 2: "*You are able to pay down another 8k before May. So you have 8k left. So it will take another 5 months. But absolutely put 2k per month, not 1k per month towards the debt. Don't touch your prior Roth contributions.*" **1.2K net upvotes.**

Net upvotes: represents the overall community consensus on content quality. A score of 0 or less typically indicates more disapproval than approval.

1 Introduction

Household financial decisions—saving, borrowing, retirement planning, portfolio allocation, and broader financial planning—are complex, heterogeneous, and consequential over the life cycle. Yet financial literacy remains persistently low, sustaining demand for guidance (TIAA Institute-GFLEC, 2025; TIAA Institute-MIT AgeLab, 2025). Consistent with this, U.S. households increasingly obtain advice not only from professionals but also from social media: as of 2025, roughly 60% consult professionals and about 20% rely on social platforms, while among adults aged 18–29 reliance on social media (42%) substantially exceeds consultation with professional advisers (27%) (FINRA, 2022, FINRA, 2024; Gallup, 2025). While the household-finance literature emphasizes information frictions, limited literacy, inattention, search costs, and intermediation as determinants of the information and guidance households bring to financial decisions (Campbell, 2006; Lusardi and Mitchell, 2014; Andersen et al., 2020; Gomes, Haliassos, and Ramadorai, 2021; Campbell and Ramadorai, 2026), the quality of tailored household-finance advice produced on social media—especially beyond investing—remains largely unmeasured.

This paper contributes to household finance by providing direct evidence on an increasingly important but previously unmeasured input into household choice: the quality of tailored financial guidance obtained from online communities across six domains: financial planning, budgeting, saving, investing (portfolio allocation), retirement saving, and borrowing (debt management). Our evidence complements the literature by isolating a distinct friction — the noisiness and downside

risk of socially sourced advice — and by identifying where it is most severe across decision domains and socioeconomic contexts. In doing so, the paper speaks directly to the measurement agenda in household finance and to when digital information environments complement, rather than substitute for, professional intermediation (Campbell and Ramadorai, 2026).

Specifically, we collect all threads posted between January 2022 and December 2023 in the ten largest Reddit personal-finance forums (subreddits), which together had more than 24 million members as of 2024, and draw a representative stratified random sample of 48 cases from 156,944 threads.[1] Cases span the six advisory domains and three subreddit groups defined by the socioeconomic orientation of members. Stratification uses LLM-based classification (ChatGPT-4.0.1) of advisory domains—defined by the decision areas addressed in posts—while subreddit groups are constructed from the forums' stated focus: Group A represents forums oriented toward high earners and high-net-worth individuals; Group B comprises the three largest broad personal-finance forums open to users across the socioeconomic spectrum; and Group C consists of forums oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals. Each case is a Reddit thread consisting of an advice-seeking post, in which a user self-discloses concrete financial circumstances and seeks tailored advice, together with the three 'Best' community comments, as ranked by the subreddit (for example, by net upvotes), reflecting responses generated in real time.[2]

With assistance from the Certified Financial Planner Board of Standards (CFP®), we recruit eight CFP® professionals to evaluate the *practical quality* of the advice. Each case is independently rated by three advisers on a five-point scale, and advisers provide written justifications for their ratings. Advisers are selected to ensure diversity in expertise, geography, and professional experience; the panel includes senior financial planners and wealth advisers, two of whom also serve as adjunct professors of finance (Table 2).

Our objective is to assess the *practical quality* of online financial advice under real-world advisory standards rather than against an abstract theoretical optimum. This distinction is important because financial advice is a credence good: its quality is difficult to assess even ex post (Egan, Matvos, and Seru, 2024). Moreover, neither regulators nor the profession prescribe a unique "gold standard" recommendation for a given client situation; instead, advice quality is evaluated against case-specific suitability and best-interest standards that can support multiple defensible recommendations depending on the client's circumstances (SEC: IA-5248, 2019; FINRA, 2026; CFP Board, 2022; CFP Board, 2026). In this respect, personal finance differs from many medical settings, where clinical practice guidelines more often map patient profiles into relatively

[1] Our study includes the following subreddits: Personal Finance, Financial Planning, Bogleheads, FIRE, HENRY Finance, FatFIRE, FIREyFemmes, RichPeople Personal Finance, Middle Class Finance, and Poverty Finance.

[2] See the example post and responses preceding the Introduction and an illustrative case in Section A.1 of the Internet Appendix.

determinate diagnostic and treatment pathways. Yet even in medicine—despite more codified guidance—studies evaluating online advice typically rely on licensed clinicians or physicians as evaluators (Cole, Watkins, and Kleine, 2016; Ayers et al., 2023; Drozd, Couvillon, and Suarez, 2018; Sterz et al., 2025). We adopt and extend this approach in finance by leveraging LLMs to refine domain stratification and to structure the analysis of advisers' written rationales—brief explanations of why a given comment merits a particular quality score. Accordingly, the relevant question is not whether online advice coincides with a single theoretically optimal action, but whether it would be regarded as substantively useful, appropriately tailored, and professionally defensible in advisory practice.

The paper addresses two broad questions. The first is *whether social-media financial advice is reliable at all*. Average quality alone is not sufficient to answer this question. A platform may produce moderately useful advice on average while still exposing households to substantial downside risk, large instability across user situations, or strong disagreement within the same thread. We therefore study the quality of online advice along three margins: its average level, its *uniformity (heterogeneity)* across comments and cases, and its *downside risk*, defined as the odds of receiving bad rather than good advice. To capture heterogeneity in quality, we distinguish between *macro* dispersion—instability in the community's ability to scale quality guidance across heterogeneous cases and contexts—and *micro* dispersion—disagreement among commenters responding to the same underlying financial situation. Macro dispersion addresses whether Reddit reliably delivers quality advice across settings, whereas micro dispersion captures how conflicting the answers are to a single question. High within-case dispersion implies contradictory recommendations and higher cognitive costs for advice seekers.

The second question is the flip side of quality: *how professionals distinguish higher- from lower-quality online advice*. We study both observable and latent correlates of quality. On the observable side, we examine whether social endorsement, commenter reputation, the informational richness of the advice-seeking post, and the advice seeker's financial sophistication predict professional assessments. On the latent side, we combine an interpretable supervised text classifier with a large language model to analyze advisers' written justifications and recover the evaluative criteria embedded in professional reasoning. This moves beyond documenting heterogeneity in quality to identifying practical signals of good versus bad advice—information that is directly useful for retail consumers navigating social-media platforms.

Our main findings are twofold. First, the average quality of social-media household-finance advice is not uniformly poor; it is typically moderate across advisory domains and forums. **Advisory domain:** quality is nonetheless uneven. More technical or normatively ambiguous domains—budgeting, portfolio allocation, and retirement—exhibit substantially greater noise at both the across-case ("macro") and within-case ("micro") levels; by contrast, borrowing and

planning are relatively stable across cases with low within-thread disagreement, and saving displays flatter, sometimes multi-modal distributions consistent with heterogeneity across and within cases. While dispersion differs sharply by domain, downside risk does not differ meaningfully across domains. **Subreddit groups:** differences are sharper across socioeconomic contexts. Group A (high earners/high net worth) exhibits the greatest inconsistency, with high macro dispersion, pronounced downside risk, and substantial within-thread disagreement; Group B (general personal finance) shows intermediate dispersion with more symmetric distributions; and Group C (financially constrained users) displays the lowest macro and micro dispersion, more uniform quality across cases, meaningful upper-tail mass, and relatively coherent advice within threads. The most striking finding is that social media delivers its most reliable and highest-quality guidance in financially constrained communities, whereas high-net-worth forums exhibit greater variability and lower average professional-assessed quality—a particularly important implication for household finance because professional advice (and the delegation of financial decisions) is often accessible mainly to households able to pay substantial fees (Gomes, Haliassos, and Ramadorai, 2021).

Second, we identify clear signals of good versus bad advice. Observable crowd signals perform poorly—net upvotes are a weak substitute for quality—while commenter reputation (e.g., top contributors), the informational richness of the advice-seeking post, and the advice seeker's financial sophistication robustly predict higher-quality advice. In particular, advice-seeker financial literacy exhibits a large, monotonic, and highly significant association with assessed quality, underscoring the importance of well-articulated problems for eliciting useful guidance. This pattern is consistent with evidence that financially literate individuals are better able to articulate constraints, evaluate recommendations, and elicit higher-quality advice (Lusardi and Mitchell, 2023; Schoar and Sun, 2024). These relationships emerge in three-way fixed-effects regressions that relate professional assessments of comment quality to comment-, post-, and user-level characteristics while absorbing adviser, domain, and year effects. Complementary textual analysis of advisers' written justifications—combining an interpretable supervised text classifier with a large language model—yields the same distinction and clarifies the underlying evaluative criteria: professionals reward advice that is specific, context-aware, and well-justified (often supported by external sources) and penalize advice that is generic, underspecified, or heuristic-driven. Specifically, adviser-rated *Good* advice is characterized by specific, context-rich language (e.g., "time horizon," "risk tolerance," "retirement accounts," "credit score," "emergency fund"), whereas *Bad* advice is driven by generic or heuristic phrasing (e.g., "rule of thumb," "simply states," "generic nature," "offers little").

Finally, we evaluate the robustness of our measurement framework by quantifying how consistently professional advisers rate the same social-media advice. Inter-adviser agreement

is highest and most robust for Saving, moderate for Investing and (and somewhat weaker for Budgeting), and low for Borrowing, Planning, and Retirement once corrected for chance. This pattern is consistent with the interpretation that professional norms and "accepted practice" are more sharply defined in Saving and portfolio allocation evaluations than in Debt and Planning evaluations, where adviser judgments appear more dispersed.

These findings contribute to several strands of literature. First, the paper contributes to household finance by providing direct evidence on an increasingly important but previously unmeasured input into household choice: the quality of tailored financial guidance obtained from online communities. Canonical work emphasizes that household decisions are shaped by information frictions, low financial literacy, inattention, and search costs (Campbell, 2006; Lusardi and Mitchell, 2014; Andersen et al., 2020; Gomes, Haliassos, and Ramadorai, 2021; Lusardi and Mitchell, 2023). Our evidence complements this agenda by isolating a distinct friction—the noisiness and downside risk of socially sourced advice—and by showing where it is most severe across decision domains and socioeconomic contexts. In this sense, the paper speaks directly to the measurement agenda in household finance and to the question of when digital information environments complement, rather than substitute for, professional intermediation (Campbell and Ramadorai, 2026). Additionally, the paper contributes methodologically by combining a field experiment based on naturally occurring social-media cases with professional ratings and textual analysis of advisers' written justifications. This design allows us to evaluate online advice under practical advisory standards, identify the features associated with higher- and lower-quality guidance, and do so in a setting where a single objective benchmark is often unavailable by construction.

Second, the paper contributes to the literature on financial advice and intermediation. Prior work shows that advisers can add value in complex environments but also documents conflicts of interest, heterogeneity in conduct, and adviser fixed effects in portfolio recommendations (Gennaioli, Shleifer, and Vishny, 2015; Bergstresser, J. M. Chalmers, and Tufano, 2008; Christoffersen, Evans, and Musto, 2013; Foerster et al., 2017; J. Chalmers and Reuter, 2020; Linnainmaa, Melzer, and Previtero, 2021; Egan, Matvos, and Seru, 2024; Andries, Bonelli, and Sraer, 2025; Greig et al., 2025). Existing studies, however, focus overwhelmingly on formal advice relationships and on investment settings. By contrast, we study a decentralized advice market in which recommendations are socially produced, publicly observable, and consumed outside a formal advisory contract, and we do so across the broader set of household-finance decisions that occupy much of households' financial lives.

Third, the paper contributes to the growing literature on social media in finance. Existing work focuses primarily on trading behavior, social signals, and investment-oriented platforms, often documenting that retail investors—especially younger cohorts—rely heavily on social media for

investment decisions, sometimes with adverse consequences (Bradley et al., 2024; Cookson et al., 2024; Chen, Peng, and Zhou, 2025; Hirshleifer, Peng, and Wang, 2025; Babolmorad, Bossaerts, and Massoud, 2025). We move beyond stock picking and portfolio allocation by studying socially sourced advice across budgeting, saving, borrowing, retirement, and planning. The contribution is therefore not only to show that social media matters for financial behavior, but to measure whether the advice itself is any good when households seek situation-specific guidance on the broader set of decisions that define household finance.

The remainder of the paper proceeds as follows. Section 2 describes the data and research design. Section 3 presents the main empirical evidence on the quality of social-media financial advice. Section 4 examines how professionals distinguish good versus bad advice. Section 5 reports robustness analyses, including evidence on inter-adviser agreement. Section 6 concludes.

2 Data and Methodology

Our data have two components: (i) Reddit threads and associated comments, and (ii) CFP® professionals' ratings and written evaluations. We first describe the Reddit data and sample-construction procedure, and then present the adviser panel and the evaluation data.

2.1 Reddit Data

Reddit, a discussion-oriented social media platform that hosts large personal-finance forums (i.e., "*subreddits*") with a combined membership exceeding 24 million users as of 2024. Each case corresponds to a Reddit thread and includes an advice-seeking post describing the financial circumstances of an advice seeker, together with the associated community responses—"**comments**"—that provide socially sourced, tailored personal-finance advice (see the example post and responses preceding the Introduction). Reddit advice-seeking threads offer a rich representation of individuals' self-disclosed demand for guidance in personal financial decision-making, spanning a wide socioeconomic spectrum—from high-net-worth individuals on the RichPeoplePF subreddit to financially constrained users on PovertyFinance. Across these forums, users voluntarily articulate concrete financial circumstances and seek tailored advice in real time. The largest subreddit devoted to household finance is PersonalFinance, which alone records approximately 2.4 million weekly visitors and 27,000 weekly contributions as of 2025. The 48 cases form a representative stratified random sample drawn from 159,944 posts and associated comment threads in the ten largest Reddit personal-finance forums between January 2022 and December 2023: Personal Finance, Financial Planning, Bogleheads, FIRE, HENRY Finance, FatFIRE, FIREyFemmes, RichPeople Personal Finance, Middle Class Finance, and

Poverty Finance.

- **r/PersonalFinance:** A general-purpose personal-finance forum covering budgeting, saving, debt, credit, investing, and retirement planning, with extensive community resources and FAQs.
- **r/FinancialPlanning:** A broad forum focused on practical financial-planning topics, including budgeting, retirement planning, insurance, and investing.
- **r/Bogleheads:** A community centered on the Bogleheads philosophy of long-horizon, low-cost, diversified index investing and disciplined portfolio management.
- **r/FIRE:** A forum focused on Financial Independence and Retiring Early, emphasizing savings rates, wealth accumulation, withdrawal rules, and lifestyle design.
- **r/HENRYfinance:** A community for "High Earners, Not Rich Yet," discussing high-income households' wealth-building challenges, taxes, high-cost living, and saving/investing tradeoffs.
- **r/fatFIRE:** A forum focused on achieving financial independence with relatively high post-retirement consumption, often oriented toward higher-income and higher-net-worth households.
- **r/FIREyFemmes:** A FIRE-oriented community with a focus on women and gender-specific perspectives on financial independence, careers, and wealth building.
- **r/RichPeoplePF:** A personal-finance forum oriented toward high-income and high-net-worth individuals, often discussing issues that exceed the typical scope of general personal-finance communities.
- **r/MiddleClassFinance:** A forum discussing financial issues salient to middle-class households, including budgeting, saving, retirement planning, and major household expenditures.
- **r/povertyfinance:** A community oriented toward financially constrained households, emphasizing frugality, coping strategies, and practical guidance under tight budget constraints.

Table 1 reports advice seekers attributes for 156,944 posts drawn from the ten largest personal-finance forums on Reddit between January 2022 and December 2023, with demographic and financial characteristics extracted from the post text using ChatGPT-4.0.1. Two features are

immediate. First, explicit self-disclosure is limited across most dimensions: 73.1% of posts do not mention age, 92.2% do not mention gender, 79.9% do not mention net income, 92.1% do not mention net worth, and 97.4% do not mention risk tolerance. This pattern is consistent with the platform's semi-anonymous environment.

Among posts with extractable characteristics, the disclosed distributions are economically intuitive. Age disclosures concentrate in early adulthood: 21–30 years accounts for 14.3% of all posts and 31–40 years for 5.9%, whereas older age brackets are each below 1.3%. Gender is rarely stated, and when stated, male (5.0%) and female (2.9%) dominate, with non-binary disclosures negligible. Topical composition is broadly diversified across core household-finance themes: Investing (20.2%) and Debt (22.0%) represent large shares, Retirement is particularly prominent (25.7%), and Planning and Saving each comprise 11.3%, while Budgeting accounts for 9.1%. In contrast, "Other" is small (0.4%), indicating that the taxonomy captures the bulk of posts' primary financial focus.

Finally, the extracted net-income and net-worth categories, as well as risk tolerance, are sparsely mentioned but informative when present. Net-income disclosures tilt toward middle and upper-middle categories (5.9% and 4.3%, respectively), with additional mass in higher-income groups, while net-worth disclosures are concentrated in lower-middle and middle categories (3.6% and 2.4%), with fewer posts in upper categories. Risk tolerance is almost never stated explicitly; when it is, aggressive (0.9%) and conservative (1.3%) appear more frequently than moderate (0.4%).

2.2 Methodology

As shown in Figure 3, we draw 48 Reddit cases as a stratified random sample from 156,944 posts and associated comment threads in the ten largest personal-finance forums between January 2022 and December 2023. Stratification relies on LLM-based classification (ChatGPT-4.0.1) into six advisory domains and three subreddit groups defined by users' socioeconomic profiles. We sample 12 cases from Group A (high earners and high-net-worth individuals; 2 cases per domain), 24 cases from Group B (the three largest general personal-finance subreddits spanning the socioeconomic distribution; 4 cases per domain), and 12 cases from Group C (subreddits oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals; 2 cases per domain).

Using 48 social-media cases, we conduct a field experiment in which each case represents a plausible individual situation and information set, together with a set of tailored responses (socially sourced advice) spanning six personal-finance domains. With assistance from the Certified Financial Planner Board of Standards (CFP®), we recruit eight CFP® professionals who

independently evaluate the advice in each case; each case is rated by three advisers. Each case is a Reddit thread initiated by an advice-seeking post. Community members reply to the post and to one another, generating a hierarchical comment tree in which discussions branch through replies to specific comments. For each case, we retain the top three comments ranked by net up-vote score (interpreted as a proxy for community endorsement); when a case contains fewer than three comments, we retain all available comments.

For each case, advisers evaluate the original post and the selected comments on a discrete 1–5 quality scale ($1 =$ "Very Poor," $5 =$ "Excellent") and provide brief written justifications for their ratings. Evaluations are submitted via a structured Google Form. Refer to the Internet Appendix for the experimental instructions and the evaluation Google Form. Table 3 reports descriptive statistics for advisers' evaluations.

Advisers are selected to ensure diversity in expertise, geographic location, and professional experience. Our panel includes senior financial planners and wealth advisers, two of whom also serve as adjunct professors of finance (see Table 2 and Figure 2 for advisers' backgrounds and case assignments).

As of 2024, there are 723,731 registered financial advisers, of whom 43% are broker-dealer–only (without a fiduciary duty), 12% are investment adviser representatives (with a fiduciary duty), and 45% are dual registrants (FINRA, 2025); see Figure 1.[3] Under fiduciary standards, advisers must act in clients' best interests by providing advice tailored to factors such as age, risk tolerance, and financial goals. Investment Adviser Representatives (hereafter, IARs) are the focus of this paper and typically operate within registered advisory firms, where they are actively involved in wealth management.

Investment Adviser Representatives (IARs) typically must pass the Series 65 (Uniform Investment Adviser Law Examination), which is governed by the North American Securities Administrators Association (NASAA). The Series 65 covers investment law and regulation, ethics and fiduciary duties, economic factors, investment vehicles, portfolio strategies, and professional practice standards relevant to advisory services. In addition to licensing requirements, advisers may pursue voluntary professional certifications intended to signal higher competence, most notably the Certified Financial Planner Board of Standards (CFP®) and the Chartered Financial Analyst Institute (CFA®). As of 2024, there are 103,093 CFP® professionals in the United States (international-adviser, 2025). The CFP® certification exam requires in-depth knowledge and skills that integrate financial planning across investments, tax planning, retirement and estate planning, risk management, ethics, applied judgment, client-centric decision-making, and

[3]Broker-dealer advisers are registered with the Financial Industry Regulatory Authority (FINRA), while Investment Adviser Representatives (IARs) are employed by investment advisory firms and regulated by either the Securities and Exchange Commission (SEC) or state authorities. Dual registrants operate as both brokers and IARs and are often registered with both FINRA and the SEC.

behavioral considerations.

As such, NASAA serves as the primary regulatory and competency-setting body for state-level advisory registration, developing examination programs (e.g., the Series 65) and related standards that govern minimum adviser competence. NASAA relies heavily on industry participation and continues to involve practitioners and compliance personnel in the development and maintenance of its exams. By contrast, private certifying bodies such as CFP® do not license practice; rather, they certify individuals who meet education, examination, experience, and ethics requirements, thereby raising the skill and ethical standards for IARs. Our adviser panel consists of investment adviser representatives (IARs) who hold the CFP® certification.

3 Quality of Personal Finance Advice on Social Media

This section examines *whether social-media financial advice is reliable at all*. Average quality alone is not sufficient to answer this question. A platform may produce moderately useful advice on average while still exposing households to substantial downside risk, large instability across user situations, or strong disagreement within the same thread. We therefore study the quality of online advice along three margins: its average level, its *uniformity (heterogeneity)* across comments and cases, and its *downside risk*, defined as the odds of receiving bad rather than good advice. To capture heterogeneity in quality, we distinguish between *macro* dispersion—instability in the community's ability to scale quality guidance across heterogeneous cases and contexts—and *micro* dispersion—disagreement among commenters responding to the same underlying financial situation. Macro dispersion addresses whether Reddit reliably delivers quality advice across settings, whereas micro dispersion captures how conflicting the answers are to a single question. High within-case dispersion implies contradictory recommendations and higher cognitive costs for advice seekers.

Very preliminary evidence from summary statistics (Panel C of Table 3) and the domain-specific histograms (Figure 4) indicates that socially sourced comment quality, relative to professional industry standards, is generally moderate and similar across advisory domains. In most domains, ratings cluster at score 3 ("moderate"), whereas retirement and saving peak at score 4 ("good"). The distributions are unimodal and centered in the midrange; remaining cross-domain differences are concentrated in dispersion and downside tail risk.

In this section, we investigate three main questions about the quality of personal finance advice on social media. (1) ***Is social-media financial advice reliable at all?*** We study the uniformity of advice quality by examining dispersion in professional assessments across comments and cases, providing a novel diagnostic of whether the Reddit community delivers coherent, tailored, and reproducible personal finance guidance. (2) ***When advice differs in quality, why?*** This

question is central to household finance: many retail consumers turn to online communities because professional advice is costly and some consumers do not fully trust financial professionals. Identifying the features associated with higher-quality advice can help households navigate these platforms more effectively by clarifying which characteristics—such as the financial literacy of advice seekers, and the content, structure, context, engagement, and forum type—are associated with better or worse online advice. (3) ***Who benefits more or less from social-media financial advice?*** We examine how the distribution of advice quality varies across online communities serving different socioeconomic groups, highlighting how community context and socioeconomic status shape households' access to reliable financial guidance.

3.1 Uniformity in Advice Quality: Evidence from Comment Fixed Effects

In this section, we assess the uniformity of socially sourced financial advice by examining the dispersion in adviser-assessed comment quality, as summarized by estimated comment fixed effects. Dispersion in these fixed effects reflects heterogeneity in the quality of advice provided and offers a direct diagnostic of whether the Reddit community delivers coherent, tailored, and reproducible guidance.

The central object of interest is heterogeneity in assessed comment quality, measured through the distribution of comment fixed effects both across and within cases. These two dimensions correspond to distinct but complementary forms of inconsistency in advice provision. Across-case dispersion reflects instability in the community's ability to scale high-quality guidance across heterogeneous users and problem types, while within-case dispersion reflects disagreement among commenters responding to the same underlying financial situation.

Econometric Framework

We begin by estimating the following two-way fixed-effects model:

$$y_{ikj} = \mu + \alpha_i + \gamma_j + \varepsilon_{ikj}, \tag{1}$$

Where i indexes comments, k indexes cases (threads), and j indexes professional advisers. The dependent variable y_{ikj} is adviser j's quality rating of comment i in case k, measured on a discrete 1–5 scale ($1 =$ "Very Poor," $5 =$ "Excellent"). The specification includes a constant μ, comment fixed effects α_i, and adviser fixed effects γ_j.

We estimate Equation 1 on the pooled sample and, in addition, separately (i) within each of the three subreddit groups that proxy advice-seeker socioeconomic status (A, B, and C) and (ii) within each of the six advisory domains: financial planning, budgeting, saving, investing (portfolio

allocation), retirement saving, and borrowing (debt management). Group A represents high earners and high-net-worth individuals; Group B comprises the three largest general personal-finance subreddits open to users across the socioeconomic spectrum; and Group C consists of subreddits oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals.

Adviser fixed effects γ_j absorb systematic differences in scoring behavior across advisers (leniency versus stringency), ensuring that variation in α_i reflects heterogeneity in comment quality rather than rater-specific scale use. For example, two advisers may rank comments similarly yet use different scoring thresholds: a relatively stringent adviser might map the three coarse categories (bad, moderate, good) to scores 1, 3, 4, whereas a more lenient adviser might map the same categories to 2, 3, 5. Adviser fixed effects absorb these level differences in scoring.

Comment fixed effects, α_i, measure relative comment quality: they capture comment i's average assessed quality relative to the mean quality of the estimation subsample (pooled, by subreddit group, or by advisory domain), after netting out persistent cross-adviser differences in scoring tendencies (leniency versus stringency). Because each comment is rated by three independent professional advisers, α_i summarizes the comment's mean adviser score and indicates whether comment i is assessed as above- or below-average within the relevant subsample. Accordingly, α_i equals the deviation of comment i's average assessed quality from the subsample mean: $\alpha_i = 0$ denotes average quality, $\alpha_i > 0$ above-average quality, and $\alpha_i < 0$ below-average quality.

The residual ε_{ikj} captures idiosyncratic disagreement at the adviser–comment level and may also reflect measurement noise arising from subjective judgment, fatigue, or inattention. We examine the properties of ε_{ikj} in subsequent robustness analyses to assess the sensitivity of our methodology to such idiosyncratic variation.

The core idea of our framework in this section is to use the cross-sectional distribution of α_i as a diagnostic of advice quality and reproducibility in social media. If socially sourced advice is consistently reliable relative to professional standards, the estimated α_i should be tightly concentrated around zero, approximately symmetric, and thin-tailed. Conversely, if advice quality is noisy from the perspective of an advice seeker, substantial heterogeneity will arise even within the same case and advisory domain, increasing the cost of identifying actionable guidance. In that setting, the distribution of α_i will exhibit greater dispersion and may display asymmetry or heavy tails, with a pronounced left tail reflecting economically meaningful downside risk.

Before analyzing comment fixed effects, we report estimates of μ and adjusted R^2 in Table 4 and Table 5, for subreddit groups and advisory domains, respectively. As shown in Table 4, the estimated constants, μ, coincide with the group-specific means reported in Panel D of Table 3 (Descriptive Statistics), and thus capture the average quality of socially sourced advice (comments)

within each group. Specifically, the model explains a substantial share of variation in quality scores, with an adjusted R^2 of 0.42 (column 1). The estimated group means indicate modest differences in average comment quality: Group A exhibits the lowest mean quality ($\mu = 2.713$) relative to Groups B (μ = 3.055) and C (μ = 3.188), while overall ratings cluster around the moderate range (≈ 3). Model fit varies by group, with adjusted R^2 values of 0.59 for Group A, 0.37 for Group B, and 0.32 for Group C, indicating more precise modeling of adviser ratings in Group A. This pattern suggests that advisers rate advice in high-earner/high-net-worth subreddits (Group A) as lower quality and do so with greater consistency (higher inter-adviser agreement) than in Groups B and C.

With respect to advisory domains, Table 5 shows that the estimated constants, μ, coincide with the domain-specific means reported in Panel C of Table 3 (Descriptive Statistics), and therefore capture the average quality of socially sourced advice (comments) within each domain. The estimated domain means indicate limited variation in average comment quality across domains: comments are generally rated as moderate (≈ 3) by advisers. Portfolio allocation ($\mu = 2.889$) and borrowing ($\mu = 2.773$) exhibit the lowest mean ratings relative to retirement ($\mu = 3.15$), saving ($\mu = 3.13$), budgeting ($\mu = 3.03$), and planning ($\mu = 3.014$). Model fit varies substantially across domains, with the highest adjusted R^2 values for retirement (0.571), saving (0.651), and budgeting (0.404), relative to planning (0.223), portfolio allocation (0.340), and borrowing (0.239).

3.1.1 Macro Inconsistency: Across-Case Dispersion in Advice Quality

Across-case dispersion (macro) in estimated comment fixed effects, $\widehat{\alpha}_i$, captures heterogeneity in the overall quality of advice delivered across different users and problem types, reflecting instability in the community's ability to provide *uniform guidance with respect to quality* across contexts (i.e., topics and cases) or subreddits. Interpretation: *"Does the Reddit community reliably deliver quality advice across different contexts or subreddits?"* High dispersion here implies uneven competence, topic sensitivity, and unreliable consistency in advice quality across cases and subreddits. This captures tailoring failure at the macro level—inconsistency in matching advice quality to case needs.

To operationalize macro-level variation, we aggregate comment fixed effects within each case. Let n_k denote the number of comments in case k, and let $\mathcal{I}_k$ denote the set of comments in that case. We define the case-level mean comment fixed effect as

$$c_k = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \widehat{\alpha}_i. \tag{2}$$

We interpret c_k as the average assessed quality of advice delivered in case k, relative to the relevant

subsample mean, after netting out adviser-specific leniency or stringency. Macro inconsistency is then captured by the dispersion of c_k across cases within a given subreddit group or advisory domain. Importantly, macro dispersion can be high even if advice within each case is internally coherent, because it reflects instability in advice quality across different users and contexts rather than disagreement within a single thread.

Figure 5 plots kernel density estimates of c_k, the case-level mean of comment fixed effects, separately for Subreddit Groups A, B, and C, as well as for the pooled sample. Because the comment fixed effects $\widehat{\alpha}_i$ are defined relative to the group-specific assessed-mean-quality (captured by the constant μ after netting out adviser fixed effects), c_k measures the extent to which the average quality of advice delivered in case k deviates from the group's mean. Accordingly, the figure is designed to capture *macro inconsistency*—that is, cross-case heterogeneity in advice quality within a given subreddit group—rather than differences in average advice quality across groups.

Group A: Panel (a) indicates substantial macro-level dispersion and pronounced downside risk. The empirical density spans approximately from -0.6 to 1.3, with a clear *heavier right tail* than the overlaid normal benchmark. Moreover, the mode is located around $c_k \approx -0.2$, implying that the most common realization is a case whose average advice quality is *below* the Group A mean. Taken together, the right-skewed tail behavior and the negative modal mass indicate that cross-case variation in Group A is characterized not only by dispersion but also by economically meaningful *downside risk*: a nontrivial fraction of cases delivers advice whose average assessed quality falls materially below the group norm.

Group B: Panel (b) displays a markedly tighter and more symmetric distribution. The empirical density spans roughly -1.2 to 1.2 and tracks the normal benchmark closely, suggesting comparatively symmetric and well-behaved cross-case variation in case-level advice quality. At the same time, the kernel density is relatively *flat in the middle range* (i.e., less sharply peaked than a normal curve), indicating that cases are distributed more evenly around the group mean rather than concentrating tightly at a single modal value. Among the three groups, Group B exhibits the strongest macro-level consistency in advice quality.

Group C: Panel (c) shows moderate dispersion combined with clear asymmetry consistent with *left-skewness* and a noticeable shift of mass toward the right side of the support. The empirical density spans approximately from -1.5 to 1.0. Relative to the normal benchmark, the distribution places substantial probability mass at $c_k > 0$ and displays *heavier upper tails*, indicating that a meaningful subset of cases attracts advice whose average assessed quality is materially *above* the Group C mean. In economic terms, Group C displays macro inconsistency that is tilted toward *upside heterogeneity*—greater dispersion driven in part by unusually high-quality cases—rather than primarily by downside realizations.

Pooled Sample: Panel (d) presents the pooled distribution of c_k, which appears approximately

symmetric and close to normal. However, this apparent regularity masks substantial heterogeneity across subreddit groups. Pooling combines the tight, near-normal distribution observed in Group B with the more dispersed and non-normal distributions in Groups A and C, thereby obscuring economically meaningful differences in macro inconsistency across advice-seeker contexts.

Overall, Figure 5 shows that macro inconsistency—measured as cross-case dispersion in c_k—varies sharply across subreddit groups. Group A exhibits the most salient downside risk (a heavier negative tail and a negative mode), Group B follows a near-normal shape with comparatively symmetric dispersion but a flatter center, and Group C exhibits asymmetry with substantial mass at positive c_k and heavier upper tails. These patterns indicate that the Reddit community's ability to deliver *uniform-quality guidance* across cases depends critically on the socioeconomic orientation of the subreddit.

Figure 6 plots kernel density estimates of c_k, the case-level mean of comment fixed effects, separately by advisory domain. **Borrowing (Panel a):** The borrowing domain exhibits a relatively symmetric distribution centered near zero, with support spanning approximately from -1.2 to 1.0. The empirical density closely tracks the normal benchmark in the central region, although mild deviations appear in the tails. This pattern suggests moderate macro-level dispersion with limited skewness: while some cases receive advice that is materially above or below the domain mean, extreme deviations are relatively rare. Overall, borrowing advice displays a moderate degree of cross-case heterogeneity but little evidence of pronounced downside or upside tail risk.

Saving (Panel b): The saving domain shows a noticeably flatter and bimodal distribution relative to the normal benchmark. The empirical density places substantial mass on both the negative and positive sides of zero, with reduced density at the center. This pattern indicates that cases in the saving domain tend to cluster into qualitatively different regimes—some attracting advice whose average quality is below the domain mean and others receiving advice that is above average. Such bimodality suggests macro inconsistency driven by heterogeneity in case types or advice-seeker characteristics, rather than random noise around a single central tendency.

Budgeting (Panel c): The budgeting domain exhibits a distribution that is *mode-below-zero* with pronounced right-tail behavior. The empirical density peaks at a negative value of c_k (roughly around 0 to -0.4), indicating that a large share of cases attracts advice whose case-level average quality is below the budgeting-domain mean. At the same time, the distribution features a long and relatively heavy *upper tail* extending well into positive values (approaching $c_k \approx 1.5$), with a secondary shoulder in the right tail. This pattern implies substantial macro-level heterogeneity: many budgeting cases deliver below-average advice quality relative to the domain mean, yet a nontrivial subset of cases receives advice that is materially above the domain norm.

Portfolio Allocation (Panel d): Portfolio allocation displays a qualitatively similar shape to

budgeting. Portfolio allocation displays a left-skewed distribution with substantial mass below zero and a thinner right tail. The empirical density peaks at negative values of c_k, indicating that many cases receive advice whose average assessed quality falls below the domain mean. The left tail extends to approximately -0.8, while the right tail is comparatively limited. This pattern points to meaningful downside risk at the case level and suggests instability in the delivery of uniformly high-quality guidance in this domain, consistent with the technical complexity and heterogeneity of portfolio-related questions.

Planning (Panel e): The planning domain exhibits a relatively smooth and symmetric distribution, albeit with somewhat wider support than the borrowing domain. The empirical density is centered close to zero and broadly follows the normal benchmark, with modest deviations in the tails. This pattern indicates moderate macro-level dispersion and relatively stable case-level advice quality around the domain mean. While some heterogeneity exists, planning advice appears comparatively uniform across cases.

Retirement (Panel f): The retirement domain shows substantial dispersion and marked departures from normality. The empirical density spans a wide range, from roughly -1.4 to 0.9, and displays a multimodal shape with pronounced mass in the negative region. This pattern suggests considerable cross-case heterogeneity, with many cases receiving advice whose average quality is well below the domain mean, alongside a smaller set of cases with above-average outcomes. The resulting distribution indicates pronounced macro inconsistency, driven largely by downside realizations.

Taken together, the domain-level densities indicate that borrowing and planning are comparatively well-behaved, saving exhibits a flatter (and potentially multi-peaked) shape, while both budgeting and portfolio allocation are characterized by mode-below-zero distributions with heavy upper tails, and retirement displays substantial dispersion and marked departures from normality. These patterns suggest that the Reddit community's ability to provide *uniform-quality guidance* varies systematically across advisory domains, with more technical and consequential topics exhibiting greater macro-level inconsistency.

3.1.2 Micro Inconsistency: Within-Case Dispersion in Advice Quality

Within-case dispersion (micro), by contrast, captures disagreement among Reddit commenters about the appropriate advice for a given financial situation, reflecting inconsistency in tailoring recommendations to a specific advice seeker. This reflects inconsistency in advice given to the same individual problem. Interpretation: *"If I post one question, how conflicting are the answers I receive?"* High dispersion here implies confusion, contradictory recommendations,

and high cognitive costs for the advice seeker. This captures tailoring failure at the micro level—disagreement about what the right advice is for a specific case.

Let $\mathcal{I}_k$ denote the set of comments in case k, with cardinality $n_k \leq 3$. We define within-case dispersion as the interquartile range (IQR) of the estimated comment fixed effects:

$$C_k \;=\; \text{IQR}(\{\widehat{\alpha}_i : i \in \mathcal{I}_k\}) \;=\; Q_{75,k} - Q_{25,k}. \tag{3}$$

Because each case contains at most three comments, the interquartile range effectively captures the spread between the highest- and lowest-quality responses within a thread, providing a robust and economically meaningful measure of within-case disagreement without relying on unstable moment-based statistics.

Figure 7 plots kernel density estimates of C_k, the within-case interquartile range (IQR) of comment fixed effects, separately for Subreddit Groups A, B, and C, as well as for the pooled sample. By construction, C_k measures the extent of disagreement in assessed advice quality *within* a case, after netting out adviser-specific leniency or stringency. Dispersion and shape differences in these distributions therefore capture *micro inconsistency*—the degree to which commenters provide conflicting guidance in response to the same financial situation.

Group A: Panel (a) indicates the highest level of micro inconsistency among the three groups. The empirical density is wide, with support extending from near zero to approximately $C_k \approx 4$, and exhibits clear multimodality. The mass is concentrated away from zero, with a pronounced peak around $C_k = 1$ and substantial probability mass above $C_k = 2$. This pattern implies that, within a given case, comments in Group A frequently differ markedly in assessed quality. From the advice seeker's perspective, this corresponds to a high likelihood of encountering sharply conflicting recommendations in response to a single question. The deviation from the normal benchmark further suggests that disagreement within cases is not well approximated by symmetric noise, but instead reflects structured heterogeneity in advice provision whitin a case.

Group B: Panel (b) shows a comparatively tighter and more regular distribution. While C_k remains strictly positive by construction, the density is more concentrated around moderate values (roughly between 0.5 and 2) and tracks the normal benchmark relatively closely. Extreme values of C_k are less common than in Group A. This pattern indicates lower within-case disagreement: comments responding to the same case in Group B tend to be more similar in assessed quality, implying greater coherence in the advice offered to a given advice seeker.

Group C: Panel (c) indicates the *lowest* degree of micro inconsistency among the three groups. The distribution is noticeably more compact than those of Groups A and B, with most of the mass concentrated in a narrow interval of approximately $C_k \in [0.7,\ 1.3]$ and a clear peak around $C_k \approx 1$. The tails are comparatively thin, with limited probability mass at high values of C_k, implying that

extreme within-case disagreement is relatively rare in Group C. Overall, this pattern suggests that, conditional on a given case, commenters in Group C tend to provide advice of more similar assessed quality, yielding the most coherent within-thread guidance among the three subreddit groups.

Pooled Sample: Panel (d) presents the pooled distribution of C_k, which appears approximately unimodal and close to normal, centered at $C_k = 1$. As in the macro analysis, pooling masks substantial heterogeneity across subreddit groups. In particular, the heavy upper tail observed in Group A and the tighter distributions in Groups B and C are averaged out in the pooled density, underscoring the importance of subgroup-specific analysis when assessing micro inconsistency.

Overall, Figure 7 reveals clear heterogeneity in micro-level inconsistency across subreddit groups. Group A exhibits the highest degree of within-case dispersion, with a wide support and substantial mass at high values of C_k, indicating frequent and pronounced disagreement among commenters responding to the same case. Group B displays moderate micro inconsistency: while disagreement is common, the distribution is more regular and closer to the normal benchmark, with fewer extreme realizations. In contrast, Group C exhibits the lowest degree of micro inconsistency. Its distribution is noticeably more compact than those of Groups A and B, with most mass concentrated between $C_k \approx 0.7$ and 1.3, indicating relatively coherent advice within threads. Overall, the figure indicates that the extent of disagreement faced by an individual advice seeker depends critically on the socioeconomic orientation of the subreddit (who is seeking advice), with Group A exhibiting the greatest within-thread conflict and Group C the most internally consistent guidance.

Figure 8 plots kernel density estimates of C_k, the within-case interquartile range (IQR) of comment fixed effects, separately by advisory domain. **Borrowing (Panel a):** The borrowing domain exhibits moderate micro inconsistency. The distribution is relatively smooth and unimodal, with most mass concentrated between approximately $C_k \approx 0.5$ and 1.5, and a peak near $C_k \approx 1$. The tails are present but not extreme, indicating that while disagreement among comments is common, sharply conflicting advice within the same case is relatively infrequent. Overall, borrowing-related discussions tend to generate moderately coherent advice within threads.

Saving (Panel b): Saving displays the highest degree of micro inconsistency across domains. The empirical density is wide and strongly multimodal, with distinct peaks at low, intermediate, and high values of C_k. The support extends well beyond $C_k = 3$, indicating frequent and substantial disagreement among commenters responding to the same saving-related question. This pronounced multimodality suggests that saving cases often elicit qualitatively different types of advice, leading to high within-thread conflict and elevated cognitive costs for advice seekers.

Budgeting (Panel c): The budgeting domain shows relatively low to moderate micro inconsistency. The distribution is more compact than that of saving and portfolio allocation, with

most mass concentrated around $C_k \approx 1$. While some dispersion is evident—particularly in the right tail—extreme values are comparatively rare. This pattern indicates that budgeting questions tend to elicit advice of broadly similar assessed quality within a case, yielding relatively coherent guidance at the micro level.

Portfolio Allocation (Panel d): Portfolio allocation exhibits substantial micro inconsistency, characterized by a wide distribution and noticeable multimodality. The density shows significant mass at both moderate and high values of C_k, with the upper tail extending beyond $C_k \approx 2$. This pattern implies frequent disagreement among commenters within the same case, consistent with the technical complexity and heterogeneity of portfolio-related advice. Advice seekers in this domain are therefore more likely to face conflicting recommendations.

Planning (Panel e): The planning domain exhibits moderate micro inconsistency with a relatively well-behaved distribution that closely mirrors the budgeting domain. The density is unimodal and concentrated around $C_k \approx 1$, with limited tail mass. Compared to saving and portfolio allocation, within-case disagreement is less pronounced, suggesting that planning-related discussions tend to generate more aligned advice among commenters.

Retirement (Panel f): Retirement exhibits a comparatively regular and compact distribution, closely approximating the normal benchmark. Most mass lies between $C_k \approx 0$ and 1.5, with thin tails. This indicates relatively low micro inconsistency: commenters responding to retirement-related cases tend to agree more closely in assessed advice quality, yielding more consistent guidance within threads.

Overall, Figure 8 reveals substantial heterogeneity in micro inconsistency across advisory domains. Saving and portfolio allocation exhibit the highest within-case dispersion, reflecting frequent disagreement among commenters responding to the same financial problem. Borrowing, budgeting, planning, and retirement display comparatively lower micro inconsistency, with more compact distributions and fewer extreme realizations. These patterns indicate that the degree of confusion and conflict faced by an individual advice seeker depends critically on the advisory domain (what the advice is about), with technically complex or normatively ambiguous topics generating the greatest within-thread disagreement.

3.2 Downside Risk

In this section, we assess downside risk—defined as the odds of receiving bad rather than good advice—and begin by examining how the distribution of socially sourced advice quality varies across online communities serving different socioeconomic groups. The unit of observation is a comment i posted in case k (a distinct advice-seeking post) and evaluated by professional adviser j.

Professional advisers initially rate each comment on a five-point ordinal scale $y^*_{ikj} \in \{1, 2, 3, 4, 5\}$, where 1 denotes *Very Poor* and 5 denotes *Excellent*. To focus on economically meaningful distinctions and reduce noise in fine grading, we coarse-grain these ratings into three ordered categories:

$$y_{ikj} = \begin{cases} 1 & \text{Bad advice (ratings 1–2),} \\ 2 & \text{Moderate advice (rating 3),} \\ 3 & \text{Good advice (ratings 4–5).} \end{cases} \tag{4}$$

Throughout, higher values correspond to higher advice quality (Bad $<$ Moderate $<$ Good).

We compare advice quality across subreddit groups that differ in socioeconomic orientation. Group A consists of subreddits targeting high earners and high-net-worth individuals; Group B comprises large general personal-finance communities spanning the socioeconomic spectrum; and Group C includes communities oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals.

Because the dependent variable is ordinal, we estimate an ordered response model. Rather than impose the proportional-odds (parallel-lines) restriction, we use a generalized ordered logit specification that allows covariate effects to differ across thresholds. Let $G_{ik} \in \{A, B, C\}$ denote the subreddit group for comment i in case k. The model is

$$\log\left(\frac{\Pr(y_{ikj} > c)}{\Pr(y_{ikj} \le c)}\right) = \kappa_c + \beta_c^B \mathbf{1}\{G_{ik} = B\} + \beta_c^C \mathbf{1}\{G_{ik} = C\} + \gamma_{c,j} + \delta_{c,d}, \qquad c \in \{1, 2\}, \tag{5}$$

where $c = 1$ contrasts "above Bad" (Moderate or Good) versus Bad, and $c = 2$ contrasts Good versus (Bad or Moderate). The model includes adviser indicators $\gamma_{c,j}$ and advisory-domain indicators $\delta_{c,d}$ (with one domain omitted as the base category– "borrowing"). Allowing coefficients to vary by threshold accommodates asymmetric effects across the lower and upper tails of the advice-quality distribution.

We estimate equation (5) using a generalized ordered logit model and compute standard errors clustered at the case (post) level to allow for arbitrary correlation across comments associated with the same advice request. There are 48 distinct cases (clusters) in the sample. The resulting generalized ordered logit estimates are reported in Table 7.

Specifically, Table 7 reports threshold-specific relative risk ratios (RRRs). Because the model estimates cumulative logits, coefficients in the first panel describe the likelihood of being rated at least Moderate (Moderate or Good versus Bad), while coefficients in the second panel describe the likelihood of being rated Good (Good versus Bad or Moderate). An RRR greater than one therefore indicates a higher likelihood of crossing the relevant quality threshold.

Several patterns emerge. Relative to high-net-worth subreddits (Group A), comments posted

in general personal-finance subreddits (Group B) are more likely to avoid the lowest quality category: the estimated RRR at the lower threshold is 1.82 ($p = 0.059$), implying an 82% higher likelihood of being rated at least Moderate rather than Bad. The corresponding coefficient for financially constrained subreddits (Group C) is also above one but is not statistically significant. At the upper threshold, the effect for Group C is more pronounced: the estimated RRR of 2.50 ($p = 0.010$) indicates a substantially higher likelihood that comments from financially constrained communities are rated Good rather than Bad or Moderate. These patterns suggest that socioeconomic heterogeneity in advice quality is concentrated toward the upper tail of the distribution. As we discuss later, this appears to be driven by a small set of contributors in Group C who consistently provide high-quality advice to financially constrained Reddit users [ADD TABLE FOR IT].

While threshold-specific coefficients indicate where differences arise, their economic interpretation is limited because they describe movements in cumulative odds rather than changes in category-specific probabilities. As such, our primary estimands are predicted probabilities and pairwise contrasts across subreddit groups. These quantities translate the model estimates into economically interpretable percentage-point differences and directly quantify how probability mass shifts across Bad, Moderate, and Good advice categories.

3.2.1 Predictive Margins and Pairwise Contrasts

For each advice-quality category $q \in \{1, 2, 3\}$ and subreddit group $g \in \{A, B, C\}$, we compute predictive margins

$$p_q(g) = \frac{1}{N} \sum_{i,k,j} \Pr(y_{ikj} = q \mid G_{ik} = g, X_{ikj}) \,, \tag{6}$$

where X_{ikj} includes adviser and advisory-domain indicators evaluated at their observed values. The results are presented in Figure 9. The figure conveys a clear socioeconomic gradient in predicted advice quality across subreddit groups. Group A (high earners/high-net-worth) has the lowest predicted probability of "good" advice and the highest predicted probability of "bad" advice. Group C (financially constrained) exhibits the opposite pattern, with the highest probability of "good" advice and a lower probability of "moderate," suggesting that quality differences operate primarily through an upper-tail shift rather than a uniform reduction in low-quality content. Group B (general personal finance) appears intermediate, with the lowest predicted probability of "bad" advice and the highest probability of "moderate," consistent with probability mass moving from "bad" to "moderate" in broad communities and from "moderate" to "good" in financially constrained communities. Confidence intervals are wide—particularly for Group C in the "bad"

and "good" panels—so some pairwise differences may not be significant, but the qualitative ranking is stable and economically meaningful.

We then report pairwise contrasts

$$\Delta_q(g, g') = p_q(g) - p_q(g'), \tag{7}$$

measured in percentage points. Inference is based on delta-method standard errors computed from the case-clustered variance–covariance matrix. The results are reported in Table 8 and yield three main findings.

Bad advice: Relative to high-net-worth communities (Group A), general personal-finance subreddits (Group B) exhibit a lower incidence of Bad advice. The estimated contrast is $\Delta_{\text{Bad}}(B, A) = -0.119$ (s.e. 0.065, $p = 0.066$), implying an 11.9-percentage-point reduction in the predicted probability of Bad advicee; the estimate is significant at the 10% level and economically meaningful in magnitude. The corresponding contrast for financially constrained communities is smaller and not statistically significant, $\Delta_{\text{Bad}}(C, A) = -0.068$ (s.e. 0.101, $p = 0.503$).

Moderate advice: Differences in Moderate advice primarily reflect reallocations needed for probabilities to sum to one. The most pronounced contrast is between financially constrained and general personal-finance communities: $\Delta_{\text{Moderate}}(C, B) = -0.168$ (s.e. 0.066, $p = 0.011$), indicating that comments in financially constrained subreddits are substantially less likely to be rated Moderate than those in general communities.

Good advice: The most robust and economically important result concerns Good advice. Surprisingly, subreddits oriented toward financially constrained users exhibit a substantially higher probability of generating Good advice than high-net-worth communities, with $\Delta_{\text{Good}}(C, A) = 0.199$ (s.e. 0.077, $p = 0.009$). This 19.9-percentage-point increase is statistically significant at the 1% level under conservative, case-clustered inference. The corresponding contrast for general personal-finance subreddits is positive but smaller and not statistically significan, $\Delta_{\text{Good}}(B, A) = 0.083$ (s.e. 0.061, $p = 0.175$).

The estimated magnitudes are economically large. For instance, if the probability of Good advice in high-net-worth subreddits is 25%, the estimated contrast $\Delta_{\text{Good}}(C, A) = 0.199$ implies an 80% relative increase in the likelihood of receiving Good advice. Even if the baseline probability were as high as 35%, the same estimate would correspond to a 57% relative increase.

Taken together, our results reveal pronounced socioeconomic heterogeneity in the quality of social-media financial advice. Relative to communities oriented toward high-net-worth individuals, general personal-finance forums are less likely to generate clearly poor advice, while communities serving financially constrained users are substantially more likely to produce high-quality advice,

as assessed by professionals after netting out persistent differences in advisers' rating tendencies (leniency versus stringency). In particular, financially constrained communities exhibit a markedly higher incidence of good advice. These patterns suggest that social media can be a comparatively valuable source of personal finance guidance for financially constrained households—who often face binding cost barriers to professional advice and lower baseline financial literacy—whereas high-net-worth households may be better served by professional advisers rather than relying primarily on peer-provided online guidance.

In the appendix, we present robustness checks using multinomial logit and binary logit specifications (Good vs. Not-Good; Bad vs. Not-Bad) with the same control structure and case-clustered inference (ADD TABLES to APPENDIX). These alternative specifications yield probability contrasts consistent in sign and magnitude with the generalized ordered logit results, indicating that the conclusions are not driven by functional-form assumptions.

4 Determinants of the Quality of Advice

The section studies the flip side of quality: *how professionals distinguish higher- from lower-quality online advice*. We study both observable and latent correlates of quality. On the observable side, we examine whether social endorsement, commenter reputation, the informational richness of the advice-seeking post, and the advice seeker's financial sophistication predict professional assessments. On the latent side, we combine an interpretable supervised text classifier with a large language model to analyze advisers' written justifications and recover the evaluative criteria embedded in professional reasoning. This moves beyond documenting heterogeneity in quality to identifying practical signals of good versus bad advice—information that is directly useful for retail consumers navigating social-media platforms.

4.1 Observable Features

Our findings from the previous section show that online advice can be substantially noisy at both micro (case-level) and macro (across cases). In this section, we conduct a fine-grained regression analysis to identify the factors that predict the quality of personal finance advice on social media. This question is central to household finance: many retail consumers turn to online communities because professional advice is costly and some consumers do not fully trust financial professionals, and identifying the features associated with higher-quality advice can help households navigate these platforms more effectively. We benchmark advice quality using evaluations by Certified Financial Planner (CFP®) professionals, including two adjunct finance professors, and thus provide an expert-grounded assessment of tailored financial guidance in online

settings. We begin by estimating the following three-way fixed-effects model of adviser-assessed advice quality:

$$\begin{aligned} y_{ikj} = {} & \beta_1 \ln(\text{CommentVotes}_i + 1) + \beta_2 \text{Top1\%Commenter}_i \\ & + \beta_3 \ln(\text{PostVotes}_k + 1) + \beta_4 \ln(\text{PostWords}_k + 1) \\ & + \beta_5 \text{FinLit}_k + \beta_6 \text{SubredditGroup}_k + \alpha_t + \delta_d + \gamma_j + \varepsilon_{ikj}. \end{aligned} \tag{8}$$

Where i indexes comments, k cases (posts), j professional advisers, t years, and d advisory domains. The dependent variable y_{ikj} is adviser j's quality rating of comment i in case k, measured on a discrete 1–5 scale ($1 =$ Very Poor, $5 =$ Excellent). The specification includes year fixed effects (α_t), domain fixed effects (δ_d), and adviser fixed effects (γ_j). Standard errors are clustered at the case level.

PostVotes and CommentVotes are net scores (upvotes minus downvotes) and capture the degree of social attention a post or comment receives from the Reddit community. We interpret these measures as proxies for crowd-based assessments of salience and perceived quality. FinLit measures the financial literacy of the advice seeker, as assessed by professional advisers on a 1–5 scale. Top1%Commenter is a binary indicator equal to one if, at the end of the previous month, the commenter ranked in the top 1% by cumulative comment upvotes within the platform, and zero otherwise. SubredditGroup classifies subreddits into three mutually exclusive categories: Group A (baseline), targeting high earners and high-net-worth individuals; Group B, comprising large general personal-finance subreddits spanning the socioeconomic distribution; and Group C, oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals.

Table 6 reports the regression estimates. Several findings emerge. First, comment-level popularity—measured by the logarithm of comment net votes—exhibits no statistically meaningful association with professional assessments of advice quality. The coefficient on $\ln(\text{CommentVotes}+1)$ is -0.074 ($p = 0.228$), indicating that crowd endorsement of individual comments is a poor predictor of adviser-rated quality. Specifically, a 1% increase in comment net votes is associated with a -0.00074 change in y, an effect that is both economically negligible and statistically insignificant. This result highlights a disconnect between crowd evaluations and professional judgments of advice quality.

Second, commenter reputation is positively associated with assessed quality. The coefficient on Top1%Commenter is 0.260 ($p = 0.089$), statistically significant at the 10% level. As this variable is binary, the coefficient reflects a discrete shift in the outcome. Specifically, holding constant year, domain, and adviser fixed effects, comments authored by top-percentile contributors receive quality ratings that are higher by 0.26 points on the 1–5 scale relative to comments authored

by other users. This finding suggests that persistent, platform-recognized expertise conveys information about advice quality beyond contemporaneous voting signals.

Third, post-level characteristics play a substantive role. Post popularity, captured by $\ln(\text{PostVotes} + 1)$, is positively associated with advice quality, with a coefficient of 0.094 ($p = 0.057$), statistically significant at the 10% level. Specifically, a 1% increase in post net votes corresponds to a 0.00094 increase in y. More importantly, informational richness—proxied by post length—exhibits a large and precisely estimated effect. The coefficient on $\ln(\text{PostWords} + 1)$ is 0.225 ($p = 0.001$), statistically significant at the 1% level. Specifically, a 1% increase in post length is associated with a 0.00225 increase in y, indicating that richer contextual information materially improves the quality of advice elicited from the crowd.

Fourth, advice-seeker financial literacy displays a strong, monotonic, and highly statistically significant relationship with assessed advice quality. Relative to the omitted literacy category, the estimated coefficients increase from 0.480 (10% level) to 1.136, 1.648, and 2.115 (all significant at the 1% level). Specifically, moving from the baseline literacy category to levels 2 through 5 is associated with increases in y of 0.48, 1.14, 1.65, and 2.12 points, respectively. These magnitudes are large relative to the scale of the outcome and are consistent with the household finance literature documenting that financially literate individuals are better able to articulate constraints, evaluate recommendations, and elicit higher-quality advice; for example see Lusardi and Mitchell (2023) and Schoar and Sun (2024).

Finally, we observe systematic heterogeneity across subreddit socioeconomic orientations. Using Group A (high earners/high-net-worth) as the reference category, advice posted in general-audience subreddits (Group B) receives ratings that are higher by 0.347 points (1% level), while advice in subreddits oriented toward middle-class and financially constrained users (Group C) is rated 0.289 points higher (10% level). These findings suggest that communities serving broader or financially constrained audiences generate advice that is, on average, assessed as higher quality by professionals than advice in high-net-worth–focused forums. This is encouraging for household finance because, as argued by Gomes, Haliassos, and Ramadorai (2021), professional financial advice—or delegation of personal financial decisions—is often accessible only to individuals able to pay substantial fees.

Overall, the results indicate that observable platform signals such as votes are weak substitutes for professional assessments of advice quality. Instead, structural features—commenter reputation, informational richness of advice-seeking posts, and the financial sophistication of advice seekers—play a central role in shaping the quality of socially sourced financial advice. Notably, higher-quality advice is disproportionately observed in general-audience and financially constrained communities, implying that individuals who lack access to professional advisers due to high fees may nonetheless improve financial decision-making by seeking peer-provided advice

on social media platforms.

4.2 Latent Features (Textual Analysis)

This subsection examines which observable features of online financial advice are systematically associated with higher or lower professional assessments of quality. The objective is to recover the criteria professionals apply when evaluating advice and to characterize these criteria using information contained in advisers' written explanations that justify their rating scores. We do not evaluate the objective correctness of advice or adviser performance in this subsection. Instead, we use modern textual analysis methods—combining classical machine-learning techniques with a large language model—as descriptive tools to extract and summarize how professionals distinguish higher- from lower-quality advice.

Our analysis uses 405 adviser assessments, each consisting of a discrete quality rating (scores 1–5) and a written explanation articulating the adviser's reasoning. These explanations constitute the primary textual data for the empirical analysis. We treat the paired observation—(adviser justification text, rated score)—as the fundamental unit of analysis.

To illustrate the structure of the data, the following examples reproduce four adviser assessments from the dataset. Each example consists of a numerical rating accompanied by a written justification:

1. **Rating score = 2 (poor):** "This certainly did not sound like professional advice and seemed like it lumped everything in one basket for investing. One does not just go low risk or high growth. It depends on the goal and time frame and of course the risk assessment of the saver."

2. **Rating score = 3 (moderate):** "Solid advice but just a bit shy on details."

3. **Rating score = 4 (good):** "BobSanchez47 gives good advice, recommending OP not invest exclusively in US stocks. Approximating the total world stock market is still going to heavily weight towards US stocks, but that's normal. VTSBX isn't an option in the list OP provided. BobSanchez47 might've meant VTSNX, which is the international stock index fund. Given OP's investment horizon and risk tolerance, an aggressive allocation is suitable. BobSanchez47 correctly states OP should not have more than 30% in bonds."

4. **Rating score = 5 (excellent):** "Very thorough with helpful context and details for the OP. They went a step further and explained how the OP should think about this question."

Methodology

We first estimate a classical supervised machine-learning (ML) text classifier and interpret the features underlying its predictions using SHAP (SHapley Additive exPlanations) values. We then estimate a large language model (LLM; GPT-4.1 API) on the same data to complement the textual analysis. In both approaches, each input observation is the paired *(adviser justification text, rating)*.

Before presenting model specifications and interpretive results, we compare the two models' out-of-sample predictive ability to classify assessments into three categories—*bad* (scores 1–2), *moderate* (score 3), and *good* (scores 4–5)—relative to advisers' ratings in a held-out test set. Table 9 reports predictive performance using a stratified 90/10 train–test split. The stratified approach is used to preserve the proportions of bad, moderate, and good advice for both training and prediction. Both models distinguish good from bad advice at rates substantially above chance. Predictive accuracy is highest at the extremes and lowest for moderate ratings, consistent with intermediate cases being less sharply defined.

This pattern motivates our empirical focus on good versus bad advice, where the mapping from textual justification to assessment is most stable. We do not interpret predictive performance causally; rather, it provides evidence that advisers' justifications contain systematic evaluative structure that can be recovered by different modeling approaches.

Machine-Learning Framework and the Role of SHAP Values

We begins with a classical supervised machine-learning text classifier to predict professional quality assessments from advisers' written justifications. The task is predictive: given an adviser's explanation, the model estimates the probability that the associated rating falls into a higher- versus lower-quality category.

Text is converted into numerical inputs using a term frequency–inverse document frequency (TF–IDF) representation. The TF–IDF vectorizer includes unigrams, bigrams, and trigrams and is restricted to the 1,000 most frequent features after removing standard and custom stop words. This representation captures both domain-specific terminology (e.g., references to risk, diversification, or time horizon) and short reasoning patterns without imposing semantic structure *ex ante*.

The predictive model is a logistic regression classifier estimated on the TF–IDF features. Logistic regression is a linear probabilistic model that maps text features to the probability that an assessment is classified as high quality rather than low quality. The linear specification is intentional. First, it provides a transparent mapping from textual features to predicted assessments. Second, it limits reliance on opaque nonlinear interactions that would complicate economic interpretation. Third, it aligns with our objective of recovering interpretable professional criteria rather than maximizing predictive accuracy *per se*.

The model is trained on a stratified training sample—constructed to preserve the proportions

of bad, moderate, and good advice—and evaluated on a held-out test set. Predictive performance provides evidence that advisers' justifications contain systematic information related to their ratings, but prediction is not the object of interest. Instead, the fitted model serves as a measurement device that summarizes how textual features (extracted by TF–IDF and weighted by the classifier) map into professional evaluations.

Coefficients from TF–IDF regressions are not readily interpretable because features are high dimensional, correlated, and scaled by corpus frequency. We therefore rely on SHAP (SHapley Additive exPlanations) values to interpret model predictions. SHAP values decompose a prediction into additive feature-level contributions, measuring each feature's average marginal contribution to the predicted classification. This provides both local explanations for individual assessments and global summaries of which textual elements professionals systematically reward or penalize.

Figure 11 reports mean absolute SHAP values from the classical machine-learning model, summarizing the average contribution of individual textual features to the model's classification of advice quality. Larger values indicate features that more strongly influence the predicted assessment, regardless of direction. Several patterns emerge.

First, features capturing specific, context-engaged advice are most influential: concrete topic terms (e.g., credit score, emergency fund, real estate) carry the highest mean SHAP values and are mainly associated with predictions of good and moderate advice.

Second, features capturing generic or underspecified advice—such as generic nature, offers little, don (don't) details, simply states, and rule thumb—also have high mean SHAP values, but their influence is concentrated in the poor-advice category. The prominence of these terms indicates that the absence of detail and reliance on vague heuristics are salient signals in professional evaluations, strongly shifting assessments toward lower quality.

Third, the plot highlights a clear distinction between process-oriented reasoning and conclusion-only statements. Features like time horizon, financial situation, retirement accounts, and lower expenses contribute to higher-quality classifications when embedded in substantive reasoning, whereas similar terms contribute to poor classifications when presented without supporting explanation. This pattern is consistent with professionals evaluating advice based on how recommendations are justified rather than on the presence of specific keywords alone.

Taken together, the SHAP results indicate that professional assessments of advice quality are strongly influenced by whether advice is specific, contextualized, and explanatory, as opposed to generic, underspecified, or heuristic-driven. The figure provides direct evidence that professionals reward engagement with the underlying financial problem and penalize advice that substitutes generalities for disciplined reasoning.

Large language Model Evidence

To complement the classical machine-learning analysis and to sharpen the interpretation of the extracted textual signals, we conduct a parallel analysis using a GPT-4.1 API.[4] The LLM is trained and evaluated on the same train–test split and is used primarily as an interpretive device to articulate professional standards in natural language, rather than as a stand-alone alternative estimator.

During training, the LLM is provided with professional evaluations of financial advice, each consisting of a discrete quality rating and a corresponding written justification. The model is instructed to infer the criteria professionals use to judge advice quality and to articulate the implicit rules underlying these judgments. Importantly, the LLM is not given any exogenous definitions of quality; instead, it learns evaluative criteria solely from the structure and content of professional explanations.

In the test phase, the LLM is then prompted to apply the inferred criteria to a new set of adviser justifications and to predict corresponding quality ratings. This two-stage prompting strategy—first eliciting professional standards, then applying them to unseen cases—mirrors the logic of the SHAP-based interpretation in the classical model while allowing criteria to be expressed directly in natural language.

Next, we prompt the LLM to summarize the principal patterns it inferred during training and relied on when generating predictions. Table 10 presents the most important features extracted and employed by the LLM to classify profesionals' assessment into bad and good advice. The table shows that qualitative analysis of the LLM's inferred evaluative criteria aligns closely with the SHAP-based findings. The LLM consistently identifies tailoring to individual circumstances, coherent risk–return reasoning, and explanatory depth as hallmarks of good advice. Conversely, it flags generic prescriptions, lack of nuance, and unsupported assertions as defining features of bad advice.

5 Robustness of the Methodology

Our empirical design relies on professional advisers' evaluations of social-media advice. Advisers rate each comment on a five-tier ordinal scale $\{1, 2, 3, 4, 5\}$, where 1 denotes *Very Poor* and 5 denotes *Excellent*. Methodologically, this is an ordinal classification task: each comment (the subject) is assigned to one of a finite set of ordered categories, in which three professional advisers independently rate the same set of comments.

High agreement among advisers implies that advisers are largely interchangeable as evaluators: the measured quality of a comment is not driven by the composition of the evaluation panel. In

[4]Training prompt (system role): The model is instructed that it receives professional evaluations of financial advice, each consisting of a 1–3 rating (1 = poor, 2 = moderate, 3 = good) and a textual justification, and is asked to infer the criteria professionals use to judge advice quality and to explain the reasoning behind these inferred criteria. Test prompt: Based on the inferred criteria, the model is asked to predict ratings for a new set of textual justifications.

our setting, this matters for interpretation because it keeps the focus on socially sourced advice rather than idiosyncratic adviser tendencies. Accordingly, this section evaluates the robustness of our methodology by quantifying inter-adviser agreement, which also addresses our fourth research question: ***How much do financial professionals agree on diagnosis and treatment when presented with the same individual situation and information set?***

5.1 Inter-Adviser agreement (Inter-Rater Reliability Study)

In this section, we assess the robustness of our experimental design by estimating inter-rater reliability (IRR) metrics, which summarize the extent of inter-adviser agreement in a single statistic while correcting for chance agreement given the number of raters, subjects (observations), and rating categories. We report a suite of IRR measures to ensure that inferences about inter-adviser reliability are not driven by the idiosyncrasies of any single coefficient.

Before turning to the empirical IRR estimates, we evaluate the suitability of the design—in particular, whether the combination of raters, subjects, and categories is sufficient to limit spurious (chance-driven) agreement—using the Monte Carlo random-ratings (RR) framework in Gwet (2014). Figure 12 reports the 95th percentile of Conger's π under an RR model in which raters assign categories to subjects purely at random. The RR-based 95th percentile serves as a critical value: under the null of random ratings, Conger's π exceeds this threshold with probability 5%. Accordingly, when the agreement coefficient computed from the observed ratings exceeds the RR-based 95th percentile, we reject the null of purely chance-driven agreement and interpret the observed agreement as nonrandom. Panel (a) plots the RR-based 95th percentile of Conger's π for $r = 2$ as a function of the number of subjects (n) and categories (q). Panel (b) plots the RR-based 95th percentile of Conger's π for $q = 2$ as a function of n and the number of raters (r).

Given our design—three raters ($r = 3$), five ordered categories ($q = 5$), and at least 20 subjects per subsample at the domain or group levels ($n \geq 20$)—the RR simulations imply that chance-corrected agreement exceeding approximately 0.2 is unlikely under random ratings (probability below 5%). Therefore, agreement estimates above this threshold are difficult to attribute to sampling noise alone, underscoring that the study is well-powered to detect nonrandom inter-adviser agreement at the domain and group levels. Notably, Retirement accounts for the minimum number of subjects (n = 20) in our study compared to other advisory domains and subreddit groups.

We next quantify inter-adviser agreement in our experiment by advisory domain and by socioeconomic group (subreddit group) using standard IRR metrics, including Conger's κ (Conger's π), Fleiss's π, Krippendorff's α, Brennan–Prediger, and Gwet's AC, defined as follows.

Inter-rater reliability Metrics

Let n denote the number of items (subjects) rated by at least two raters, and let r_i be the number of raters who scored item i (here typically $r_i = 3$). Ratings take ordered categories $k \in \{1, \ldots, q\}$ with $q = 5$. Let r_{ik} be the number of raters who assign item i to category k, so that $\sum_{k=1}^{q} r_{ik} = r_i$. We use an ordinal weight matrix $W = \{w_{kl}\}_{k,l=1}^{q}$ with $w_{kk} = 1$ and $w_{kl} \in [0, 1]$ decreasing in $|k - l|$ (e.g., linear or quadratic weights). All chance-corrected coefficients take the form

$$\hat{\theta} = \frac{p_o - p_e}{1 - p_e}, \tag{9}$$

and differ only in how the expected chance agreement p_e is defined.

» Observed (weighted) agreement:

For multi-rater ordinal ratings, a convenient representation of observed agreement is the average weighted pairwise agreement within item:

$$p_o \equiv p_a = \frac{1}{n} \sum_{i=1}^{n} p_{a|i}, \qquad p_{a|i} = \sum_{k=1}^{q} \frac{r_{ik} \left(r_{ik+} - 1 \right)}{r_i (r_i - 1)}, \qquad r_{ik+} \equiv \sum_{l=1}^{q} w_{kl} \, r_{il}. \tag{10}$$

This construction counts, for each item, all ordered rater pairs and credits partial agreement according to w_{kl}. Because the rating scale is ordinal, we permit *partial agreement*: two raters may be treated as agreeing to some extent even when their assigned categories do not coincide. Operationally, instead of scoring agreement as a binary indicator (one for an exact match and zero otherwise), we assign a weight $w_{kl} \in [0, 1]$ to each pair of categories (k, l) that gives full credit to exact matches and partial credit to near misses. The ordinal weight matrix $W = \{w_{kl}\}_{k,l=1}^{q}$ is symmetric ($w_{kl} = w_{lk}$), satisfies $w_{kk} = 1$ for all k as follows:[5]

Ordinal weighting matrix W (partial-agreement weights)

	1	2	3	4	5
1	1.0000	1.0000	0.0000	0.0000	0.0000
2	1.0000	1.0000	0.5000	0.0000	0.0000
3	0.0000	0.5000	1.0000	0.5000	0.0000
4	0.0000	0.0000	0.5000	1.0000	1.0000
5	0.0000	0.0000	0.0000	1.0000	1.0000

» Fleiss's π (common marginals):

[5]The analysis uses a discrete, nonstandard partial-agreement weighting matrix that assigns full credit to exact matches ($w_{kk} = 1$), full credit to select adjacent-category matches (e.g., 1 vs. 2 and 4 vs. 5), partial credit to select two-step disagreements (weight 0.5), and zero credit otherwise. Consequently, *Percent Agreement* should be interpreted as a *weighted* agreement rate under this partial-agreement definition rather than a strict exact-match rate.

Fleiss's π uses a single pooled marginal distribution across raters to define chance agreement. Define pooled category shares

$$\pi_k \equiv \frac{\sum_{i=1}^{n} r_{ik}}{\sum_{i=1}^{n} r_i} \quad (k = 1, \ldots, q), \tag{11}$$

and compute expected (weighted) agreement under independent draws from the pooled distribution:

$$p_e^{(\text{Fleiss } \pi)} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}\, \pi_k\, \pi_l. \tag{12}$$

» Conger's κ (rater-specific marginals; corrected for rater heterogeneity):

Conger's κ allows each rater to have a distinct marginal propensity to use each category. Let $g \in \{1, \ldots, r\}$ index raters, and let n_g be the number of items rated by rater g. Let n_{gk} be the number of items that rater g assigns to category k, and define rater-specific marginals

$$p_{gk} \equiv \frac{n_{gk}}{n_g}, \qquad p_{+k} \equiv \frac{1}{r} \sum_{g=1}^{r} p_{gk}. \tag{13}$$

Conger's chance agreement uses a covariance adjustment term (capturing dispersion in rater-specific marginals) and is

$$p_e^{(\text{Conger})} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \left(p_{+k} p_{+l} - \frac{s_{kl}}{r} \right), \quad s_{kl} \equiv \frac{1}{r-1} \sum_{g=1}^{r} \left(p_{gk} p_{gl} - r\, p_{+k} p_{+l} \right). \tag{14}$$

» Brennan–Prediger (uniform-marginal baseline):

The Brennan–Prediger coefficient uses a uniform chance baseline that is invariant to prevalence in observed marginals. For ordinal (weighted) agreement, the chance agreement is

$$p_e^{(\text{BP})} = \frac{1}{q^2} \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}, \tag{15}$$

which reduces to $p_e = 1/q$ in the unweighted case ($w_{kk} = 1$, $w_{kl} = 0$ for $k \neq l$).

» Gwet's AC (ordinal: AC2; prevalence-robust baseline)

Gwet's AC2 modifies chance agreement to reduce sensitivity to category prevalence and marginal imbalance. Let $T_w \equiv \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}$ be the total weight mass. Using the pooled marginal

shares π_k above, the expected chance agreement for AC2 is

$$p_e^{(\text{AC2})} = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k(1-\pi_k). \tag{16}$$

» Krippendorff's α (ordinal; unbalanced r_i allowed):

Krippendorff's α can be written in the same chance-corrected agreement form. Let

$$\bar{r} \equiv \frac{1}{n}\sum_{i=1}^{n} r_i \qquad \text{and} \qquad \pi_k \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_{ik}}{r_i}\right). \tag{17}$$

Define the finite-sample adjustment term

$$\varepsilon_n \equiv \frac{1}{n\,\bar{r}}. \tag{18}$$

Expected (weighted) chance agreement is

$$p_e^{(\alpha)} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}\,\pi_k\,\pi_l, \tag{19}$$

and the observed agreement is adjusted as

$$\hat{p}_o^{(\alpha)} = (1-\varepsilon_n)\,p_o + \varepsilon_n. \tag{20}$$

Krippendorff's alpha is then

$$\hat{\alpha} = \frac{\hat{p}_o^{(\alpha)} - p_e^{(\alpha)}}{1 - p_e^{(\alpha)}}. \tag{21}$$

Inter-rater reliability Metrics by Domain

Results are reported in Table 11. Our findings indicate substantial heterogeneity in inter-adviser agreement across domains. After correcting for chance agreement, agreement is highest in *Saving*, moderate in *Investing* (and to a lesser extent *Budgeting*), and weak—often statistically indistinguishable from zero—in *Debt*, *Planning*, and *Retirement*. We use *Landis–Koch benchmark scale* for interpreting IRR metrics: Poor (<0); Slight ($[0, 0.2)$); Fair ($[0.2, 0.4)$); Moderate ($[0.4, 0.6)$); Substantial ($[0.6, 0.8)$); Almost Perfect ($[0.8, 1]$).

Borrowing ($n = 22$): Weighted percent agreement is 0.583 (SE 0.126), but the chance-corrected coefficients are small and statistically indistinguishable from zero: Brennan–Prediger 0.256

($p = 0.254$), Cohen/Conger's κ 0.162 ($p = 0.389$), Scott/Fleiss' κ 0.110 ($p = 0.673$), Gwet's AC 0.324 ($p = 0.158$), and Krippendorff's $\alpha = 0.124$ ($p = 0.628$). Under the Landis–Koch scale, these estimates fall in the slight-to-fair range at best, but confidence intervals include zero throughout, implying that—under the chosen weighting scheme and with 22 subjects—agreement is not reliably different from chance.

Saving ($n = 23$): Agreement is high and precisely estimated. Weighted percent agreement is 0.761 (SE 0.083). All chance-corrected measures are economically large and statistically significant: Brennan–Prediger 0.573 ($p < 0.001$), Cohen/Conger's κ 0.520 ($p = 0.003$), Scott/Fleiss' κ 0.529 ($p = 0.002$), Gwet's AC 0.596 ($p < 0.001$), and Krippendorff's $\alpha = 0.536$ ($p = 0.001$). These estimates lie predominantly in the moderate-to-substantial range, indicating a stable and nonrandom consensus among advisers when evaluating Saving-related advice.

Budgeting ($n = 22$): Budgeting exhibits intermediate reliability. Weighted percent agreement is 0.614 (SE 0.092). Chance-corrected agreement is in the fair range: Brennan–Prediger 0.310 (marginal, $p = 0.058$), Cohen/Conger's κ 0.281 ($p = 0.108$), Scott/Fleiss' κ 0.244 ($p = 0.171$), and Krippendorff's $\alpha = 0.255$ ($p = 0.088$). Notably, Gwet's AC is 0.336 and statistically significant ($p = 0.038$), consistent with the interpretation that agreement is present but modest, and that prevalence and marginal effects may attenuate kappa-type statistics.

Investing ($n = 24$): Investing (portfolio allocation) shows consistent, statistically detectable agreement. Weighted percent agreement is 0.632 (SE 0.067). Chance-corrected coefficients fall in the fair range but are more significant than in Budgeting: Brennan–Prediger 0.343 ($p = 0.004$), Scott/Fleiss' κ 0.247 ($p = 0.047$), Gwet's AC 0.384 ($p = 0.001$), and Krippendorff's $\alpha = 0.258$ ($p = 0.030$). Cohen/Conger's κ is marginal at conventional levels (0.296, $p = 0.061$). Collectively, these results imply nontrivial consensus among advisers when rating Investing advice, although the magnitude remains below that in Saving.

Planning ($n = 24$): Planning resembles Debt. Weighted percent agreement is 0.569 (SE 0.134), while chance-corrected coefficients are small and statistically insignificant (e.g., Brennan–Prediger 0.231, $p = 0.335$; Cohen/Conger's κ 0.156, $p = 0.478$; Gwet's AC 0.269, $p = 0.263$; Krippendorff's $\alpha = 0.150$, $p = 0.537$). Although confidence intervals are wide, the point estimates do not support meaningful agreement beyond chance under the chosen weighting scheme.

Retirement ($n = 20$): Retirement exhibits relatively strong reliability. Weighted percent agreement is 0.708 (SE 0.127). Chance-corrected measures are sizable: Brennan–Prediger 0.479 ($p = 0.035$) and Gwet's AC 0.522 ($p = 0.010$) are statistically significant; Cohen/Conger's κ 0.407 is marginal ($p = 0.067$); Scott/Fleiss' κ 0.391 and Krippendorff's $\alpha = 0.401$ are not significant at the 5% level but are economically meaningful with wide confidence intervals. Overall, Retirement falls in the moderate range, with statistical significance depending on the specific chance model.

Notably, Krippendorff's α is most often used to quantify inter-rater reliability when multiple coders assign labels to the same set of items, especially when data come from content analysis / coding tasks (text, disclosures, news, filings, transcripts, survey responses) where coder agreement is needed to validate the coding scheme. In empirical finance, it is used when researchers code qualitative or semi-structured information—e.g., tone, topics, risk disclosures, governance features, or ESG classifications, as a reliability metric that accommodates ordinal scores and incomplete coder–item matrices (e.g., Berg, Kölbel, and Rigobon, 2022; Kara et al., 2022).

Taken together, inter-adviser agreement is highest and most robust for Saving, moderate for Investing and (and somewhat weaker for Budgeting), and low for Borrowing, Planning, and Retirement once corrected for chance. This pattern is consistent with the interpretation that professional norms and "accepted practice" are more sharply defined in Saving and portfolio allocation evaluations than in Debt and Planning evaluations, where adviser judgments appear more dispersed. The difference is informative and might reflect task ambiguity rather than rater competence.

Interestingly, in medical research, Krippendorff's $\alpha \approx 0.4$–0.75 is often viewed as credible reliability for multi-rater ordinal coding in pathology-style tasks, whereas $\alpha \geq 0.75$ more commonly arises in narrower, highly standardized imaging classifications (e.g., Dano et al., 2020; Crimì et al., 2020; Solass et al., 2019; Choi et al., 2019; Travis et al., 2013). In our setting, agreement reaches this "credible" range only in the *Saving* domain; in the remaining advisory domains, chance-corrected agreement is materially lower, typically falling in the fair-to-slight range under conventional benchmarks.

Inter-rater reliability Metrics by Socioeconomic Group

Results by subreddit group are reported in Table 12. The IRR estimates by subreddit group exhibit economically meaningful heterogeneity in inter-adviser agreement across user socioeconomic profiles. Overall, agreement is strongest in Group A (high earners and high-net-worth users), weaker and imprecisely estimated in Group B (broad, heterogeneous personal-finance communities), and intermediate in Group C (middle-class and financially constrained users). This pattern is consistent with the industry structure of professional advice: advisers are typically trained and employed as wealth managers, and the resulting greater familiarity with high-net-worth contexts is reflected in higher inter-adviser agreement (i.e., greater competency as captured by consensus) when evaluating advice directed at wealthier users.

Group A (high earners / high-net-worth; $n = 36$): Weighted percent agreement is 0.681 (SE 0.079, $p < 0.001$). All chance-corrected coefficients are statistically significant and economically sizable: Brennan–Prediger 0.430 ($p = 0.002$), Cohen/Conger's $\kappa = 0.374$ ($p = 0.021$), Scott/Fleiss' $\kappa = 0.355$ ($p = 0.027$), Gwet's AC 0.463 ($p = 0.001$), and Krippendorff's

$\alpha = 0.361$ ($p = 0.014$). Interpreted through conventional benchmarks, these estimates lie in the fair-to-moderate range, indicating a stable and nonrandom professional consensus when rating advice for high-socioeconomic-status users.

Group B (general personal finance; $n = 67$): Weighted percent agreement is 0.624 (SE 0.116, $p < 0.001$), yet chance-corrected agreement is materially weaker and statistically indistinguishable from zero for most coefficients: Brennan–Prediger 0.329 ($p = 0.113$), Cohen/Conger's $\kappa = 0.236$ ($p = 0.310$), Scott/Fleiss' $\kappa = 0.221$ ($p = 0.308$), Gwet's AC 0.377 (marginal, $p = 0.064$), and Krippendorff's $\alpha = 0.225$ ($p = 0.263$). The divergence between moderate weighted percent agreement and weak chance-corrected measures is consistent with prevalence and marginal-distribution effects: once corrected for chance agreement under the observed marginals, residual consensus is limited. Substantively, the breadth and heterogeneity of Group B user circumstances appear to generate more dispersed professional judgments.

Group C (middle-class / financially constrained; $n = 32$): Weighted percent agreement is 0.641 (SE 0.108, $p < 0.001$). Chance-corrected coefficients are positive and economically nontrivial but estimated less precisely than in Group A: Brennan–Prediger 0.358 (marginal, $p = 0.064$), Cohen/Conger's $\kappa = 0.327$ ($p = 0.087$), Scott/Fleiss' $\kappa = 0.292$ ($p = 0.167$), Gwet's AC 0.383 ($p = 0.040$), and Krippendorff's $\alpha = 0.299$ ($p = 0.083$). Overall, Group C falls in the fair range, with statistical significance depending on the chance model; nonetheless, the consistently positive estimates across measures indicate nontrivial agreement.

5.2 Inter-Adviser Disagreement

In this section, we cross-check the evidence from the IRR analysis using the two-way fixed-effects specification in Equation 1 ($y_{ikj} = \mu + \alpha_i + \gamma_j + \varepsilon_{ikj}$), which we estimate separately by each of the six advisory domains and, in a parallel set of analyses, by each subreddit group. Indices i, k, and j denote comments, cases (threads), and professional advisers, respectively. The dependent variable, y_{ikj}, is the quality score assigned by adviser j to comment i in case k, measured on a discrete 1–5 scale (1 = "Very Poor," 5 = "Excellent"). The specification includes a constant, μ, capturing the mean quality level within a domain, comment fixed effects, α_i, and adviser fixed effects, γ_j:

Comment fixed effects, α_i, capture each comment's average assessed quality—common across advisers—relative to the mean quality within a given advisory domain (subressit group). Adviser fixed effects, γ_j, estimate adviser j's systematic leniency or stringency—common across comments.

The residual, ε_{ikj}, is the component of adviser j's rating of comment i in case k that remains after netting out (i) the comment's average quality, α_i, and (ii) adviser j's systematic

leniency/stringency, γ_j. We interpret ε_{ikj} as idiosyncratic disagreement at the rating-comment level. Conceptually, it reflects inter-adviser randomness and inconsistency (disagreement)—our object of interest—along with measurement noise (e.g., miscoding, day-to-day judgment variation, fatigue, or inattention).

Our approach rests on the premise that, under high reliability in professional training and competence, advisers should assign similar quality ratings to the same piece of advice tailored to a given client's circumstances, after netting out comment fixed effects and adviser fixed effects. Accordingly, disagreement should be centered near zero and exhibit low dispersion, reflecting uniform and reproducible advisory standards. Conversely, greater dispersion—manifested in higher variance, skewness, or heavy tails—signals limited reliability and substantial adviser-specific variation.

Specifically, we measure adviser disagreement using the regression residual,

$$\hat{\varepsilon}_{ikj} \;=\; y_{ikj} \;-\; \hat{\alpha}_i \;-\; \hat{\gamma}_j. \tag{22}$$

Because this measure is based on residuals, it can be positive or negative. A value of zero indicates perfect agreement with the comment's average assessed quality (after accounting for adviser leniency or stringency). Negative (positive) values indicate that adviser j assigns a lower (higher) score than implied by the comment's average, *i.e.*, is relatively more stringent (lenient) for that rating event. The underlying intuition is that, for a given piece of advice, dispersion in quality ratings reflects limited reliability in professional assessment. Figure 13 plots $\hat{\varepsilon}_{ikj}$ across all comments and advisory domains and reveals substantial dispersion in ratings for the same comments within the same cases (ADD EXPLANATION).

Differences in residual ratings may reflect genuine disagreement in advisers' quality assessments or measurement noise (e.g., miscoding, day-to-day judgment variation, fatigue, or inattention). To attenuate idiosyncratic noise, we average residuals for each adviser within each case. Let n_{kj} denote the number of comments evaluated by adviser j in case k. We define adviser j's average disagreement in case k as

$$d_{kj} = \frac{1}{n_{kj}} \sum_{i \in \mathcal{I}_k} \widehat{\varepsilon}_{ikj}, \tag{23}$$

Figure 14 plots the distribution of d_{kj}. We find substantial cross-adviser differences in average residual ratings, with dispersion that varies across advisory domains. Under high reliability, d_{kj} should be centered near zero (at both the mean and median) with low dispersion; in contrast, greater dispersion—manifested in higher variance, skewness, or heavy tails—signals limited reliability and substantial adviser-specific variation.

All seven panels use the same kernel (Epanechnikov) and bandwidth (0.18), so cross-domain

differences mainly reflect the data, not smoothing. In the pooled panel ("All advisory domains"), the kernel density is sharply concentrated near zero and broadly tracks the normal reference, suggesting that—aggregating across domains—mean adviser–case deviations from the comment- and adviser-adjusted benchmark are modest and approximately symmetric. Specifically, in terms of centrality, the distribution is sharply concentrated near zero, with a peak density of roughly 0.80–0.85 at $d_{kj} \approx 0$. In terms of dispersion, most probability mass lies within approximately [-1.0,+1.0], with relatively thin tails extending to about -1.8and +1.2. While this pooled evidence suggests limited average disagreement, the domain-specific panels reveal economically meaningful heterogeneity in both spread and asymmetry.

The **Saving** domain exhibits the clearest signature of high reliability: its density is strongly peaked at (or extremely close to) zero with comparatively limited tail mass, indicating that adviser–case mean residuals are tightly clustered and that systematic adviser-specific departures are small for most observations. Specifically, its density is most sharply concentrated at zero, with a peak of about 1.0–1.1 at $d_{kj} \approx 0$.

Portfolio allocation also shows substantial mass near zero, consistent with high agreement for the majority, but stands out for a more pronounced extreme negative tail; this combination implies generally strong reliability punctuated by a minority of adviser–case pairs exhibiting unusually large negative mean residuals. Specifically, Portfolio allocation is characterized by substantial central mass near zero (peak density roughly 0.7–0.8) combined with marked asymmetry in the tails: the left tail extends to approximately −2.0 (including a small far-left bump), whereas the right tail largely dissipates by about +1.1. This pattern is consistent with high agreement for most adviser–case pairs, punctuated by a minority of extreme negative deviations.

Planning appears moderately concentrated around zero yet with a slight rightward shift of the modal mass, consistent with mild positive asymmetry in mean residuals (i.e., a greater incidence of positive d_{kj} than would be expected under a symmetric benchmark), while still remaining relatively close to the normal overlay. Specifically, the modal mass is shifted to the right, with the mode occurring around $d_{kj} \approx$ +0.2 to +0.4, and material probability extending from roughly -1.1 to +0.9, producing a comparatively flat central plateau rather than a sharp peak at zero.

Budgeting likewise maintains nontrivial central mass but displays more visible structure—most notably a secondary concentration on the negative side—suggesting greater heterogeneity in adviser-specific tendencies within this domain, as would arise if subsets of advisers systematically diverge in their mean residual patterns. Specifically, the kernel density has a slightly positive mode (approximately +0.05 to +0.15, peak density around 0.75–0.80) but also features a distinct left-side shoulder near $d_{kj} \approx$ -0.7 (density roughly 0.30–0.35), consistent with heterogeneity in adviser–case mean deviations. Its support extends to approximately -1.2 and +1.2, with heavier tails than implied by the normal overlay.

Two domains, borrowing and retirement, exhibit comparatively weaker reliability under the residual-based disagreement metric. In **Borrowing**, the density is shifted left of zero and shows thicker tails relative to the normal reference, indicating both a systematic negative tilt in mean residuals and a higher prevalence of adviser–case pairs with sizable departures from the benchmark. The density is systematically shifted left of zero, with a modal region around $d_{kj} \approx$ -0.15 to -0.25 (peak density near 0.8), indicating a tendency toward negative mean residuals (i.e., relatively more stringent assessments). Borrowing also exhibits thicker tails than the normal benchmark, with material mass spanning roughly [-0.8,+0.8] and visible right-tail probability out to about +1.0.

Retirement displays the most pronounced evidence of limited reliability: the density is less tightly centered at zero and visibly multi-modal, with meaningful mass on both sides of zero and a notably heavy left tail. Such multi-modality is consistent with substantively different adviser "styles" or interpretive frameworks that generate distinct clusters of mean residuals, rather than a single homogeneous error distribution. Specifically, the density is visibly multi-modal, with local peaks near $d_{kj} \approx 0$ and around +0.3 to +0.5(each with density on the order of 0.65–0.75), together with a long negative tail reaching approximately -1.5 and nontrivial mass below -1.0.

Taken together, the kernel densities imply strongest reliability in **Saving** (and broadly in the pooled distribution), high but tail-sensitive reliability in **Portfolio allocation**, intermediate reliability in **Planning** and **Budgeting** (with Budgeting showing more heterogeneity), and the weakest reliability—marked by asymmetry, heavy tails, and/or multi-modality—in **Borrowing** and especially **Retirement**.

Next, we replicate the analysis for socioeconomic heterogeneity by estimating Equation 1 separately within each of the three subreddit groups (A, B, and C). This stratification allows us to examine how adviser disagreement varies with the socioeconomic profile of advice seekers. Figure Figure 15 plots the distribution of d_{kj} across subreddit groups.

All three panels employ the same Epanechnikov kernel and a common bandwidth of 0.18; accordingly, cross-group differences in the estimated densities primarily reflect differences in the underlying distribution of adviser–case mean residuals d_{kj}, rather than differential smoothing. Subreddit **Group A** (high earners/high net worth) exhibits the strongest signature of agreement: the distribution is sharply centered at $d_{kj} = 0$, with a peak density of approximately 1.10–1.15 at $d_{kj} \approx 0$. Most probability mass lies within roughly $[-0.5, +0.5]$, with relatively thin tails tapering toward about -0.9 and $+0.9$, and the kernel closely tracks the normal overlay, indicating only limited departures from normality.

In contrast, **Group B** (the largest general personal-finance subreddits spanning a broad socioeconomic spectrum) displays weaker concentration at the benchmark and greater directional

asymmetry. The density at $d_{kj} = 0$ declines to about 0.84–0.86, and the mode shifts rightward to approximately $d_{kj} \approx +0.15$ to $+0.30$, with a peak near 0.88–0.90. This rightward shift implies a higher incidence of positive mean residuals (directional leniency). Dispersion is also materially higher: the support widens, with nontrivial mass extending from roughly -1.2 to $+1.0$. Relative to the normal benchmark, the right tail is visibly heavier (the empirical density exceeds the normal for $d_{kj} \gtrapprox 0.3$), consistent with an elevated frequency of large positive deviations.

Group C (middle-class and financially constrained, low-income–oriented subreddits) shows the weakest clustering at zero and the clearest evidence of heterogeneity. The center is flatter and the peak lower than in Groups A and B; the density at $d_{kj} = 0$ is approximately 0.75–0.80, with the mode close to zero but slightly positive ($d_{kj} \approx 0$ to $+0.10$). The distribution remains broad, with substantial mass spanning roughly $[-0.8, +0.8]$ and tails reaching about -1.1 and $+1.0$, indicating elevated disagreement relative to Group A. Correspondingly, the normal approximation performs less well: the empirical density is flatter around the center and exhibits thicker tails (exceeding the normal in portions of both tails), consistent with substantial dispersion and adviser–case heterogeneity.

Taken together, the panels imply a clear reliability gradient across socioeconomic strata. Agreement is highest in **Group A** (peak $\approx$ 1.10–1.15 at zero and tight central mass), intermediate in **Group B** (lower central density, a right-shifted mode around $\approx +0.2$, and wider tails), and lowest in **Group C** (flatter center, broader dispersion, and more prominent tail mass). Under the interpretation of d_{kj} as adviser–case mean disagreement, adviser assessments are most reproducible for advice seekers in higher-SES subreddits and become progressively more dispersed—and more directionally tilted toward positive mean residuals—in the broad and financially constrained groups, particularly Groups B and C. This pattern is unsurprising, as professional advisers' training and practice are typically oriented toward serving higher-wealth clients, rather than middle-class and low-income households.

References

Andersen, Steffen, John Y Campbell, Kasper Meisner Nielsen, and Tarun Ramadorai (2020). "Sources of inaction in household finance: Evidence from the danish mortgage market". *American Economic Review* 110.10, pp. 3184–3230.

Andries, Marianne, Maxime Bonelli, and David Alexandre Sraer (2025). "Financial advisors and investors' bias". *The Review of Financial Studies, forthcoming*.

Ayers, John W., Adam Poliak, Mark Dredze, Eric C. Leas, Zixin Zhu, John B. Kelley, Douglas J. Faix, Alyssa M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Dave M. Smith

(2023). “Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum”. *JAMA Internal Medicine* 183.6, pp. 589–596. DOI: 10.1001/jamainternmed.2023.1838. URL: https://pubmed.ncbi.nlm.nih.gov/37115527/.

Babolmorad, Nazanin, Peter Bossaerts, and Nadia Massoud (2025). *Viral but Vanishing: Investment Advisors, Social Media, and Regulation*. Tech. rep. Faculty of Economics, University of Cambridge.

Berg, Florian, Julian F Kölbel, and Roberto Rigobon (2022). “Aggregate confusion: The divergence of ESG ratings”. *Review of Finance* 26.6, pp. 1315–1344.

Bergstresser, Daniel, John MR Chalmers, and Peter Tufano (2008). “Assessing the costs and benefits of brokers in the mutual fund industry”. *The Review of Financial Studies* 22.10, pp. 4129–4156.

Bradley, Daniel, Jan Hanousek Jr, Russell Jame, and Zicheng Xiao (2024). “Place your bets? The value of investment research on Reddit’s Wallstreetbets”. *The Review of Financial Studies* 37.5, pp. 1409–1459.

Campbell, John Y (2006). “Household finance”. *The journal of finance* 61.4, pp. 1553–1604.

Campbell, John Y and Tarun Ramadorai (2026). *Household finance in retrospect and prospect*. Tech. rep. Journal of Finance: Insights and Perspectives, forthcoming.

CFP Board (Jan. 2022). *Guide to Satisfying the Duty of Care When Providing Financial Advice That Does Not Require Financial Planning*. PDF. URL: https://www.cfp.net/-/media/files/cfp-board/standards-and-ethics/compliance-resources/duty-of-care-guide.pdf.

— (2026). *Code of Ethics and Standards of Conduct*. Web page. Accessed: 2026-03-11. URL: https://www.cfp.net/ethics/code-of-ethics-and-standards-of-conduct.

Chalmers, John and Jonathan Reuter (2020). “Is conflicted investment advice better than no advice?” *Journal of Financial Economics* 138.2, pp. 366–387.

Chen, Shuaiyu, Lin Peng, and Dexin Zhou (2025). “Wisdom or Whims? Decoding the Language of Retail Trading with Social Media and AI”. *Decoding the Language of Retail Trading with Social Media and AI (April 02, 2025)*.

Choi, Eun Jung, Eun Hye Lee, You Me Kim, Yun-Woo Chang, Jin Hwa Lee, Young Mi Park, Keum Won Kim, Young Joong Kim, Jae Kwan Jun, and Seri Hong (2019). “Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography

Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study". *Ultrasonography* 38.2, pp. 172–180.

Christoffersen, Susan EK, Richard Evans, and David K Musto (2013). "What do consumers' fund flows maximize? Evidence from their brokers' incentives". *The Journal of Finance* 68.1, pp. 201–235.

Cole, Jennifer, Chris Watkins, and Dorothea Kleine (2016). "Health Advice from Internet Discussion Forums: How Bad Is Dangerous?" *Journal of Medical Internet Research* 18.1, e4. DOI: 10.2196/jmir.5051. URL: https://www.jmir.org/2016/1/e4/.

Cookson, J Anthony, Runjing Lu, William Mullins, and Marina Niessner (2024). "The social signal". *Journal of Financial Economics* 158, p. 103870.

Crimì, F, A Varotto, G Orsatti, C Lacognata, D Cecchin, AC Frigo, P Zucchetta, R Stramare, and F Pomerri (2020). "Lung visualisation on PET/MRI: implementing a protocol with a short echo-time and low flip-angle volumetric interpolated breath-hold examination sequence". *Clinical Radiology* 75.3, 239–e15.

Dano, Hélène, Serdar Altinay, Laurent Arnould, Noella Bletard, Cecile Colpaert, Franceska Dedeurwaerdere, Benjamin Dessauvagie, Valérie Duwel, Giuseppe Floris, Stephen Fox, et al. (2020). "Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study". *Modern Pathology* 33.3, pp. 354–366.

Drozd, Brandy, Emily Couvillon, and Andrea Suarez (2018). "Medical YouTube Videos and Methods of Evaluation: Literature Review". *JMIR Medical Education* 4.1, e3. DOI: 10.2196/mededu.8527. URL: https://mededu.jmir.org/2018/1/e3/.

Egan, Mark, Gregor Matvos, and Amit Seru (2024). "The Problem of Good Conduct among Financial Advisers". *Journal of Economic Perspectives* 38.4, pp. 193–210.

FINRA (2026). *FINRA Rule 2111 (Suitability) FAQ*. Web page. Accessed: 2026-03-11. URL: https://www.finra.org/rules-guidance/key-topics/suitability/faq.

Foerster, Stephen, Juhani T Linnainmaa, Brian T Melzer, and Alessandro Previtero (2017). "Retail financial advice: does one size fit all?" *The Journal of Finance* 72.4, pp. 1441–1482.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2015). "Money doctors". *The Journal of Finance* 70.1, pp. 91–114.

Gomes, Francisco, Michael Haliassos, and Tarun Ramadorai (2021). "Household finance". *Journal of Economic Literature* 59.3, pp. 919–1000.

Greig, Fiona, Tarun Ramadorai, Alberto G Rossi, Stephen P Utkus, and Ansgar Walther (2025). "Human financial advice in the age of automation". *Available at SSRN 4301514*.

Gwet, Kilem L (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hirshleifer, David, Lin Peng, and Qiguang Wang (2025). "News diffusion in social networks and stock market reactions". *The Review of Financial Studies* 38.3, pp. 883–937.

Kara, Alper, Aziidah Nanteza, Aydin Ozkan, and Yilmaz Yildiz (2022). "Board gender diversity and responsible banking during the COVID-19 pandemic". *Journal of Corporate Finance* 74, p. 102213.

Linnainmaa, Juhani T, Brian T Melzer, and Alessandro Previtero (2021). "The misguided beliefs of financial advisors". *The Journal of Finance* 76.2, pp. 587–621.

Lusardi, Annamaria and Olivia S Mitchell (2014). "The economic importance of financial literacy: Theory and evidence". *American Economic Journal: Journal of Economic Literature* 52.1, pp. 5–44.

— (2023). "The importance of financial literacy: Opening a new field". *Journal of Economic Perspectives* 37.4, pp. 137–154.

Schoar, Antoinette and Yang Sun (2024). *Financial Advice and Investor Beliefs: Experimental Evidence on Active vs. Passive Strategies*. Tech. rep. MIT Sloan Working Paper 7155-24.

SEC: IA-5248 (June 2019). *Commission Interpretation Regarding Standard of Conduct for Investment Advisers*. SEC Issue Date: June 5, 2019. URL: https://www.sec.gov/files/rules/interp/2019/ia-5248.pdf.

Solass, Wiebke, Christine Sempoux, Norman J Carr, Frederic Bibeau, Daniel Neureiter, Tarkan Jäger, Tina Di Caterino, Christophe Brunel, Eckhard Klieser, Claus W Fristrup, et al. (2019). "Reproducibility of the peritoneal regression grading score for assessment of response to therapy in peritoneal metastasis". *Histopathology* 74.7, pp. 1014–1024.

Sterz, Jasmina, Yvonne Beaugé, Pia Tueckmantel, Lena Bepler, Armin N. Flinspach, Yves Gramlich, Ralf Verboket, Priscilla Bintaro, Maren Janko, Maria H. Flinspach, Moritz Merker, Simon Bepler, Jan T. Vollrath, Sebastian H. Voß, and Martin Rüsseler (2025). "Didactic and Content Quality of Basic Life Support Videos on YouTube: Cross-Sectional Study". *JMIR Formative Research* 9, e69103. DOI: 10.2196/69103. URL: https://formative.jmir.org/2025/1/e69103.

Travis, Simon PL, Dan Schnell, Piotr Krzeski, Maria T Abreu, Douglas G Altman, Jean–Frédéric Colombel, Brian G Feagan, Stephen B Hanauer, Gary R Lichtenstein, Philippe R Marteau,

et al. (2013). “Reliability and initial validation of the ulcerative colitis endoscopic index of severity”. *Gastroenterology* 145.5, pp. 987–995.

6 Figures and Tables

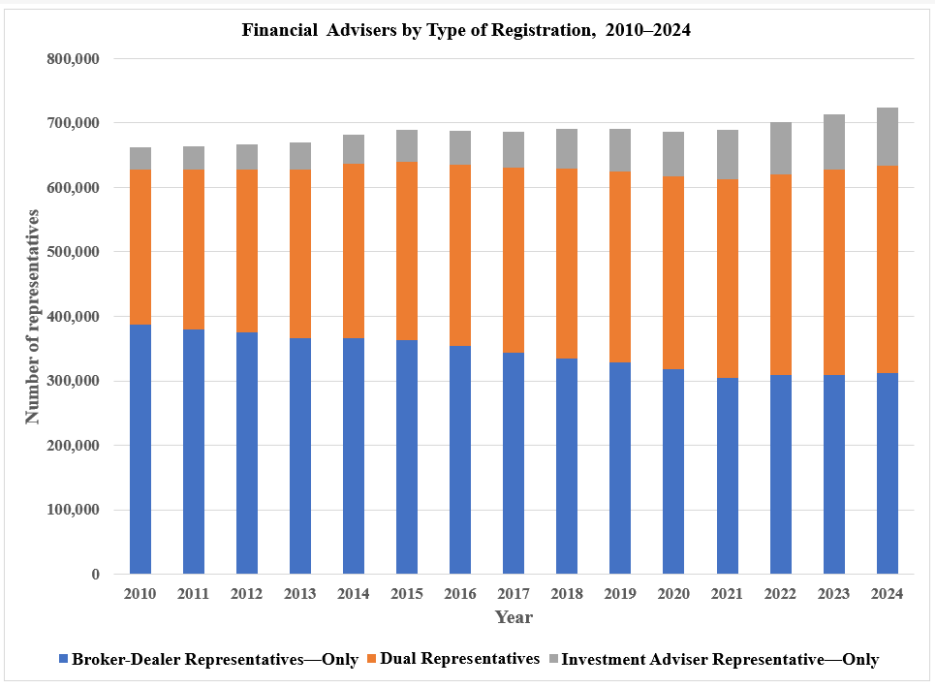


Figure 1: Financial Advisers by Type of Registration, 2010–2024

This figure shows the number of registered representatives/brokers, investment adviser representatives, and dual registered advisers over time; Source: FINRA Industry Snapshot (2025).

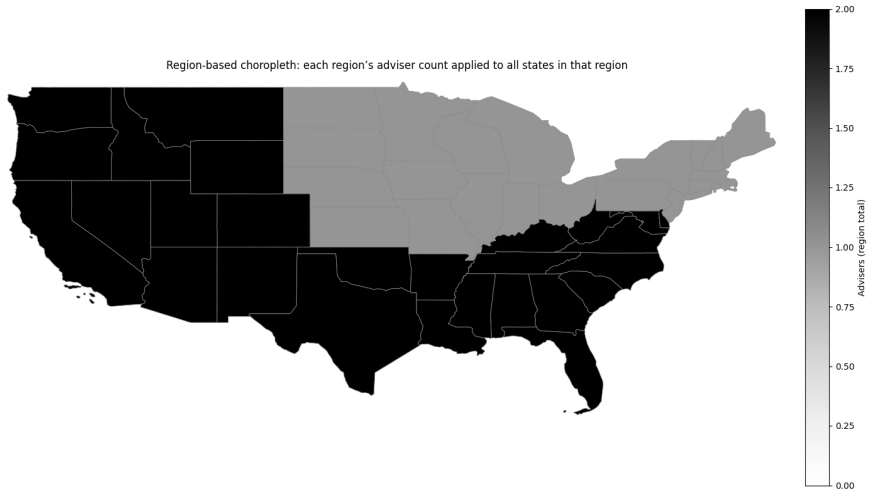


Figure 2: Geographic Distribution of Selected Advisers by U.S. Region

This figure maps the cross-sectional distribution of expert advisers across U.S. regions (State-Level Imputation). Each adviser is assigned to a Census-style region based on their reported state, and the region-level adviser count is then imputed to all states within that region. Shading intensity (grayscale) increases with the number of advisers in the corresponding region; state borders are shown for geographic reference.

Row	Group	Sub Reddit	Title
1	A	**Fire**	FI/RE - Financial Independence & Retiring Early
2		**RichPeoplePF**	RichPeoplePF
3		**HENRYfinance**	High Earner, Not Rich Yet
4		**fatFIRE**	Wealth and Financial Independence/Retire Early
5		**FIREyFemmes**	FIREyFemmes
6	B	**FinancialPlanning**	Financial Planning, Personal Finance, Frugality, Money, and More!
7		**personalfinance**	Personal Finance
8		**Bogleheads**	Passive Indexing Community for Long-Term Lazy Investors
9	C	**MiddleClassFinance**	Middle Class Finance
10		**povertyfinance**	Personal Finance For The Financially Challenged

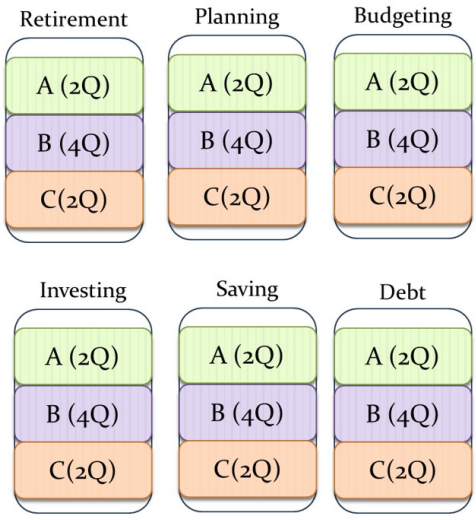
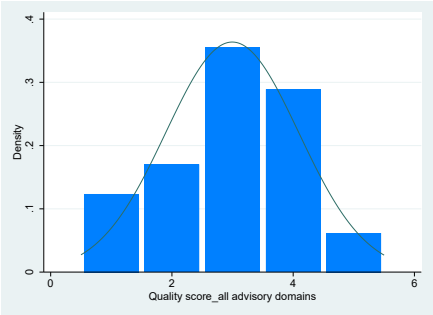
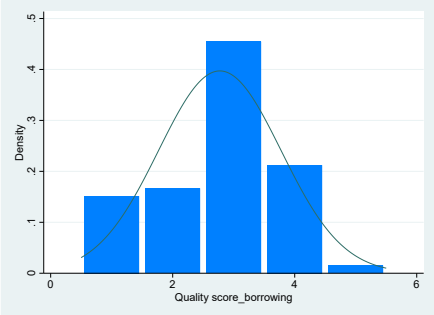


Figure 3: Construction of the 48-Case Reddit Sample

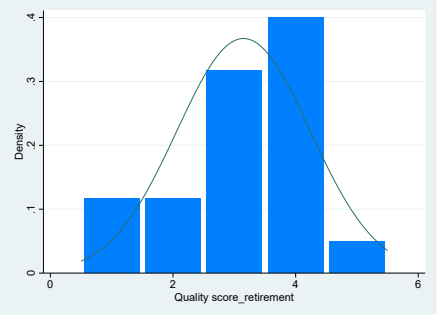
We draw 48 Reddit cases as a representative stratified random sample from 156,944 posts and associated comment threads across the ten largest personal-finance forums between January 2022 and December 2023. Stratification uses LLM-based classification (ChatGPT-4.0.1) into six advisory domains and three subreddit groups defined by users' socioeconomic profiles. Group A represents high earners and high-net-worth individuals; Group B comprises the three largest general personal-finance subreddits open to users across the socioeconomic spectrum; and Group C consists of subreddits oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals.



a) All advisory domains



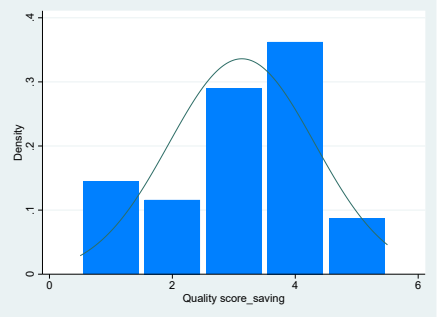
b) Borrowing



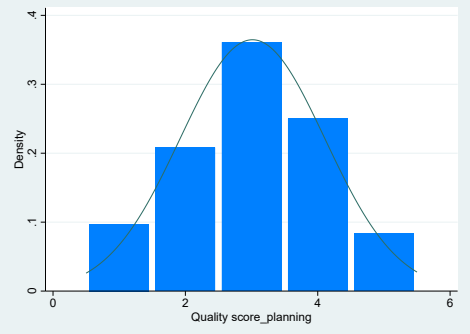
c) Retirement



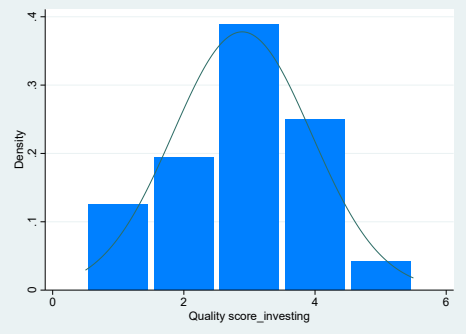
d) Budgeting



e) Saving



f) Planning



g) Investing

Figure 4: Distribution of Adviser Quality Scores Across Advisory Domains.

Histograms show empirical densities of 1–5 quality ratings by advisers; the overlaid curve is the fitted normal distribution for each panel (pooled and by domain).

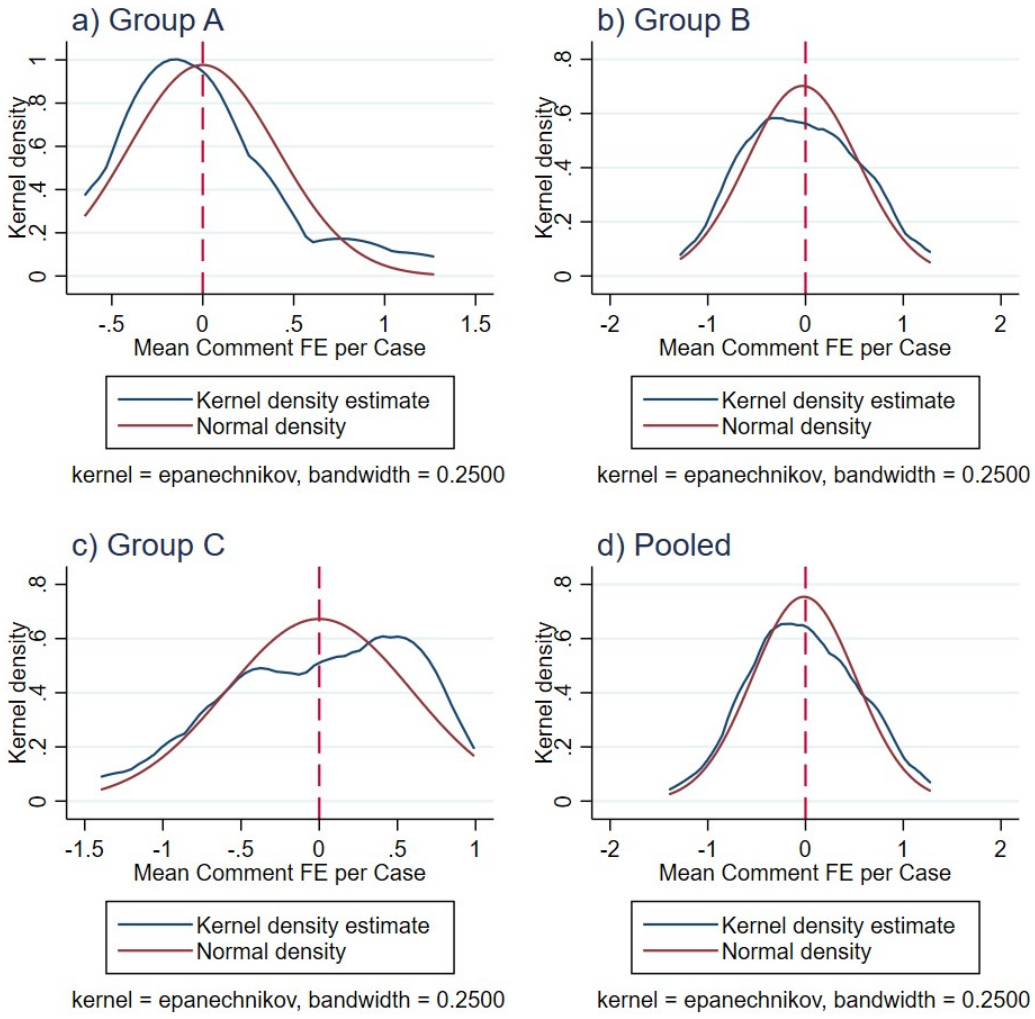


Figure 5: Macro Inconsistency in Advice Quality by Subreddit Group.

This figure plots kernel density estimates of c_k, the case-level mean of estimated comment fixed effects, separately for Subreddit Groups A, B, and C, as well as for the pooled sample. Comment fixed effects $\widehat{\alpha}_i$ are obtained from the two-way fixed-effects model in Equation 1 and capture comment i's average assessed quality relative to the group-specific mean, after netting out adviser-specific leniency or stringency. The case-level mean is defined as $c_k = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \widehat{\alpha}_i$, where n_k is the number of comments in case k and $\mathcal{I}_k$ denotes the set of comments in that case. Dispersion in c_k across cases captures macro-level inconsistency in advice quality within a subreddit group.

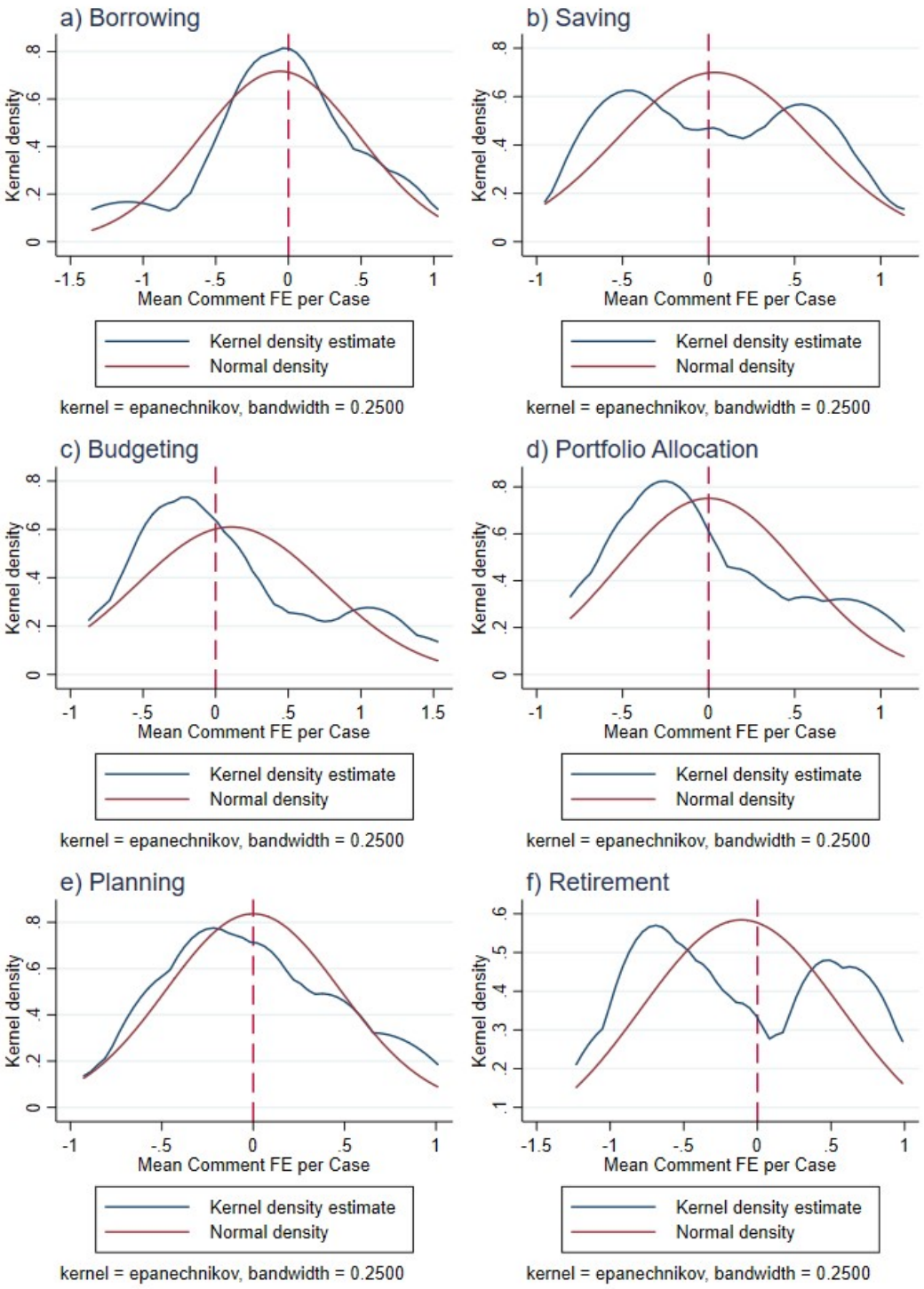


Figure 6: Macro Inconsistency in Advice Quality by Advisory Domain.

This figure plots kernel density estimates of c_k, the case-level mean of estimated comment fixed effects, separately for each advisory domain. Comment fixed effects $\widehat{\alpha}_i$ are obtained from the two-way fixed-effects model in Equation 1 and capture comment i's average assessed quality relative to the domain-specific mean, after netting out adviser-specific leniency or stringency. The case-level mean is defined as $c_k = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \widehat{\alpha}_i$, where n_k is the number of comments in case k and $\mathcal{I}_k$ denotes the set of comments in that case. Dispersion in c_k across cases captures macro-level inconsistency in advice quality within a domain.

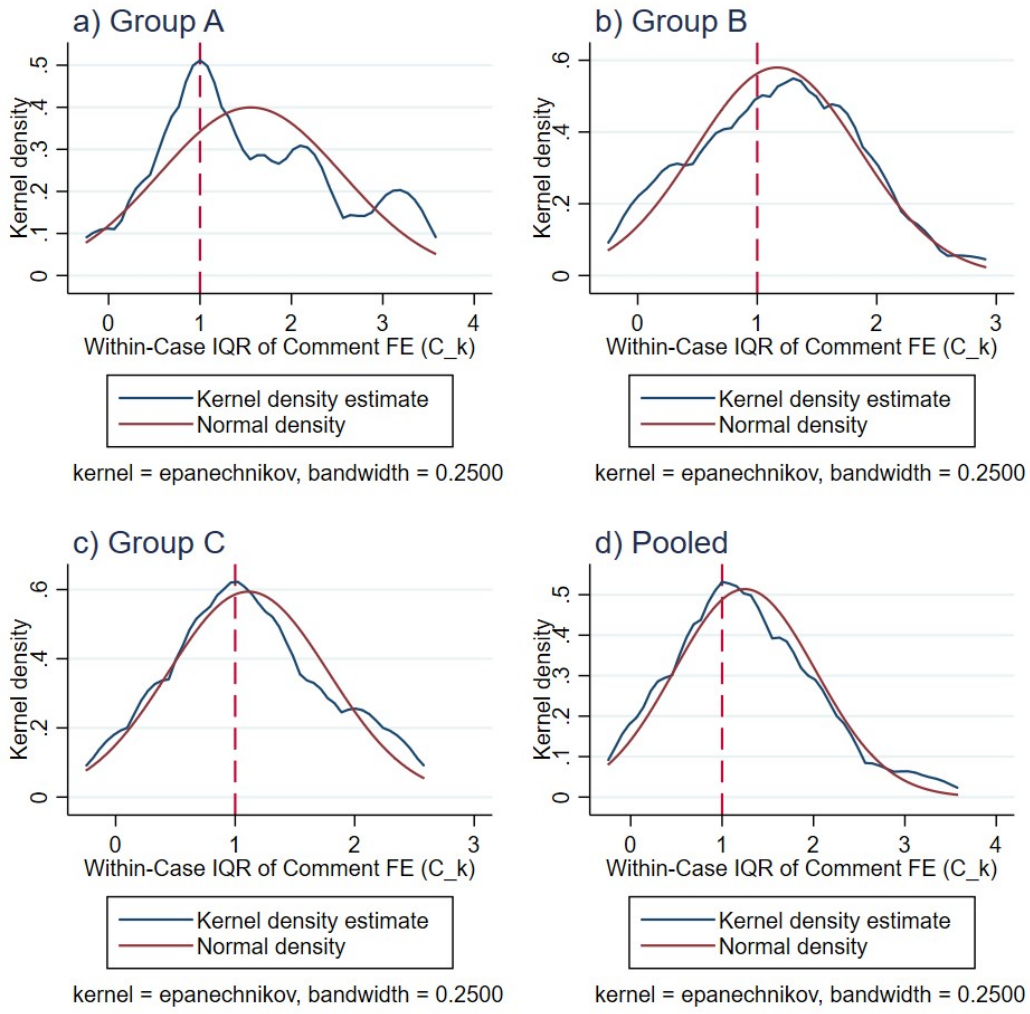


Figure 7: Micro-Level Inconsistency in Comment-Level Advice Quality by Group

This figure plots kernel density estimates of C_k, the within-case interquartile range (IQR) of estimated comment fixed effects, separately for Subreddit Groups A, B, and C, as well as for the pooled sample using Equation 3. C_k measures the dispersion in assessed advice quality across comments responding to the same case.

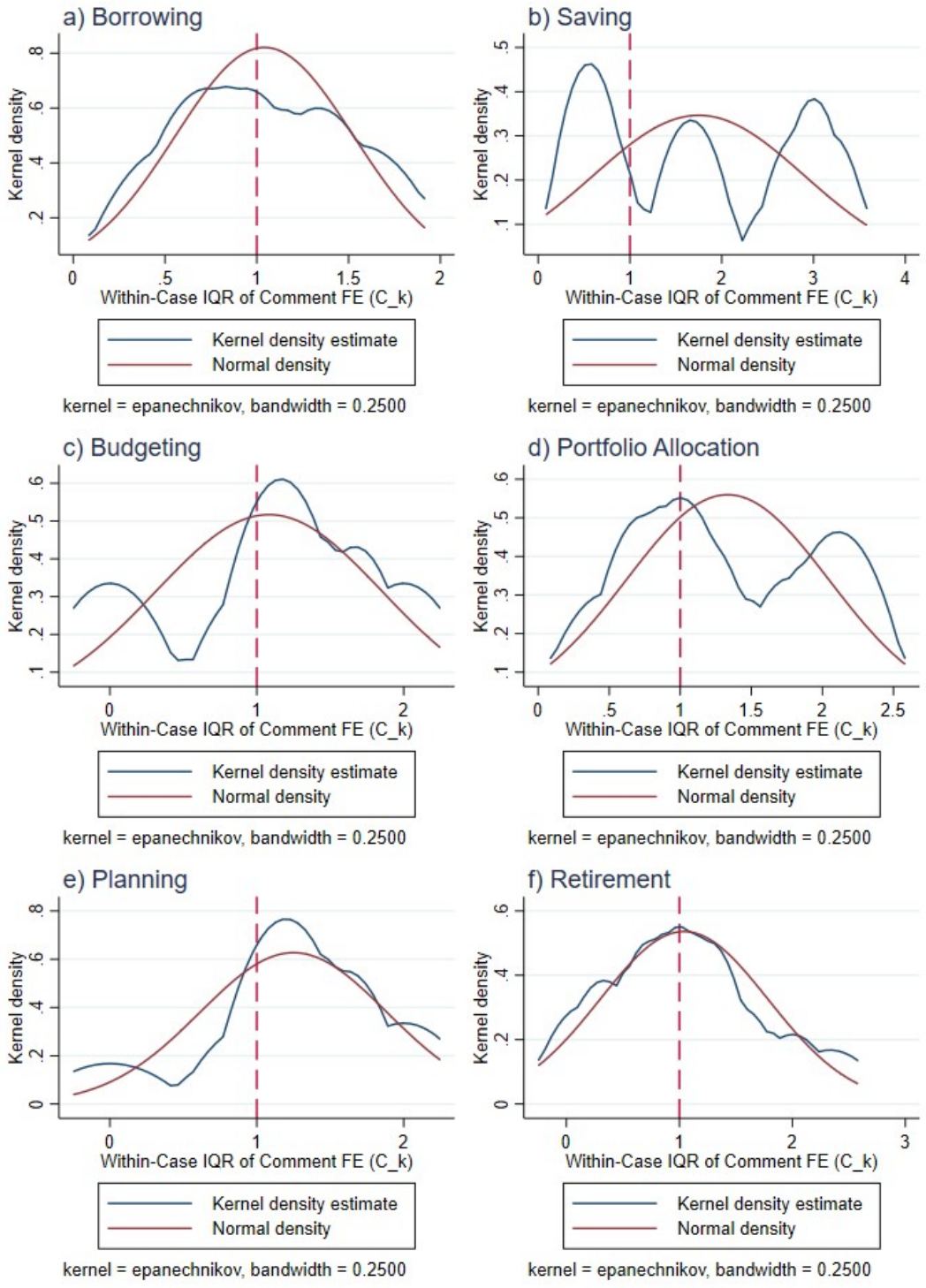


Figure 8: Micro-Level Inconsistency in Comment-Level Advice Quality by domain

This figure plots kernel density estimates of C_k, the within-case interquartile range (IQR) of estimated comment fixed effects, separately across advisory domains using Equation 3. C_k measures the dispersion in assessed advice quality across comments responding to the same case.

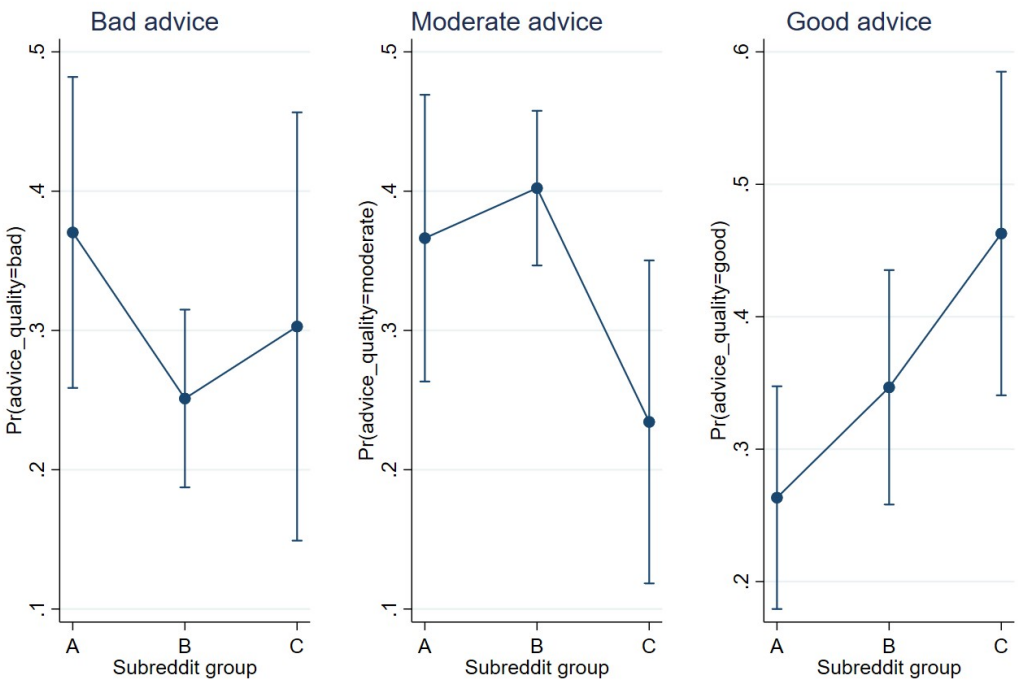


Figure 9: Predicted Probabilities by Subreddit Group (Generalized Ordered Logit).

This figure plots predicted probabilities of advice quality categories (Bad, Moderate, and Good) across subreddit groups, computed from the generalized ordered logit model in Equation (5). Predicted probabilities are obtained by averaging over the observed distribution of covariates, including adviser indicators and advisory-domain indicators, rather than evaluating covariates at their means, using Equation (6). Group A consists of subreddits targeting high earners and high-net-worth individuals; Group B comprises large general personal-finance subreddits spanning the socioeconomic spectrum; and Group C includes subreddits oriented toward middle-class and financially constrained users. Standard errors are clustered at the case (post) level.

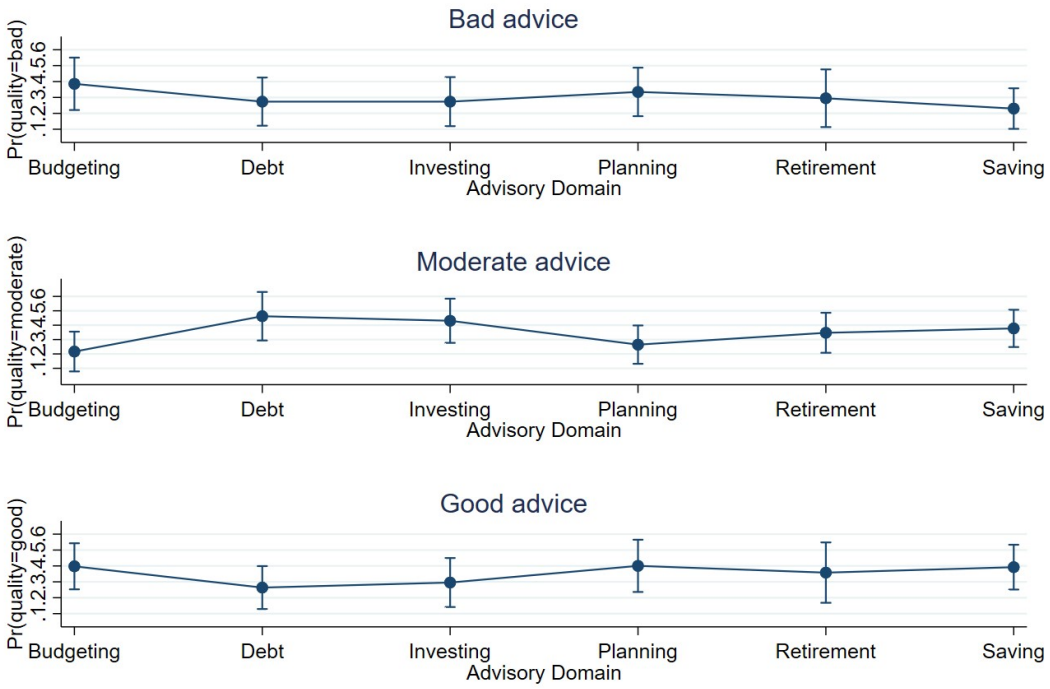


Figure 10: Predicted Probabilities by Advisory Domain (Generalized Ordered Logit).

This figure plots predicted probabilities of advice quality categories (Bad, Moderate, and Good) across advisory domains, computed from the generalized ordered logit model in Equation (5). Predicted probabilities are obtained by averaging over the observed distribution of covariates, including adviser indicators and advisory-domain indicators, rather than evaluating covariates at their means, using Equation (6)[Edit EQUATION]. Standard errors are clustered at the case (post) level.

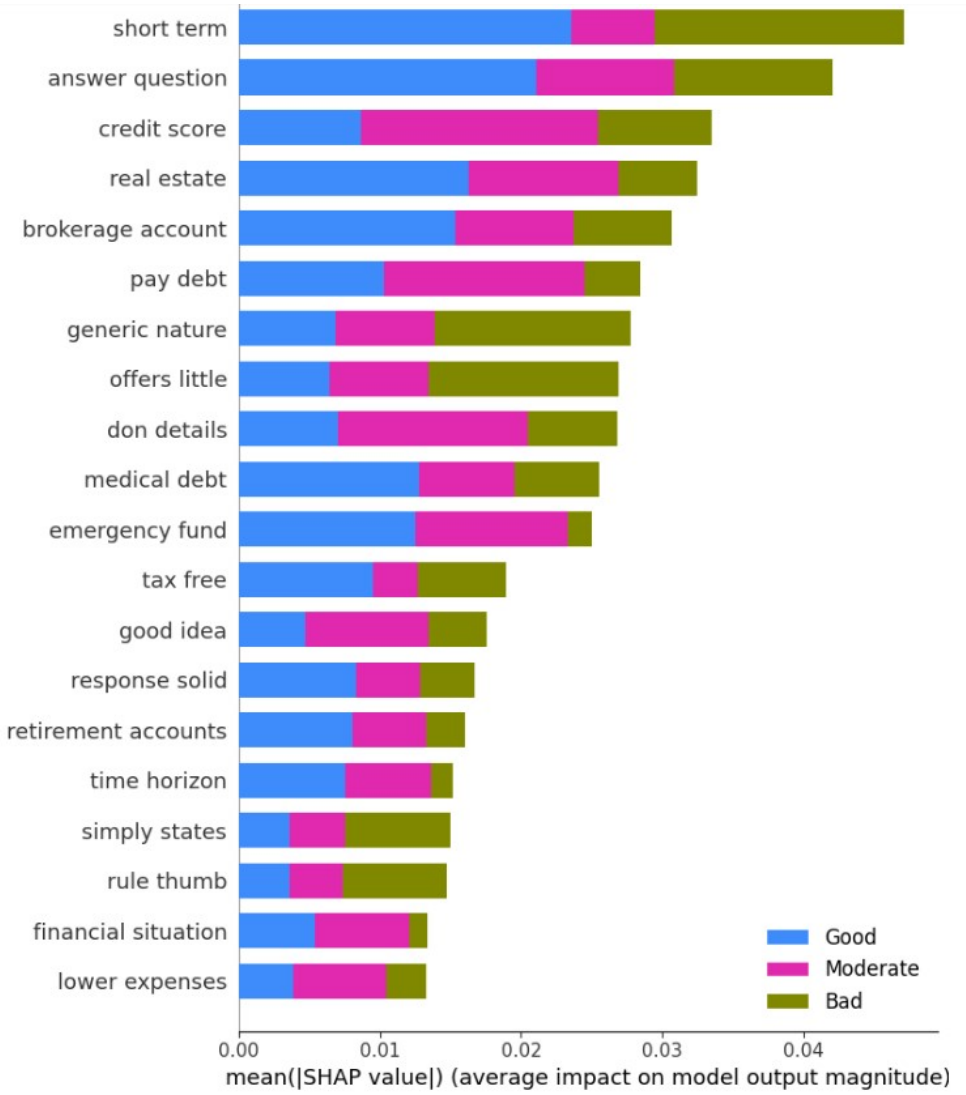


Figure 11: Mean Absolute SHAP Values by Advice-Quality Category (Classical ML)

This figure reports mean absolute SHAP values from the classical supervised text-classification model (TF–IDF features with logistic regression), computed on the held-out test set. Bars show the average absolute contribution of each textual feature to the model's predicted classification, aggregated by predicted advice-quality category (Good, Moderate, Bad). Larger values indicate features that have greater average influence on the model output (i.e., greater predictive importance), irrespective of sign.

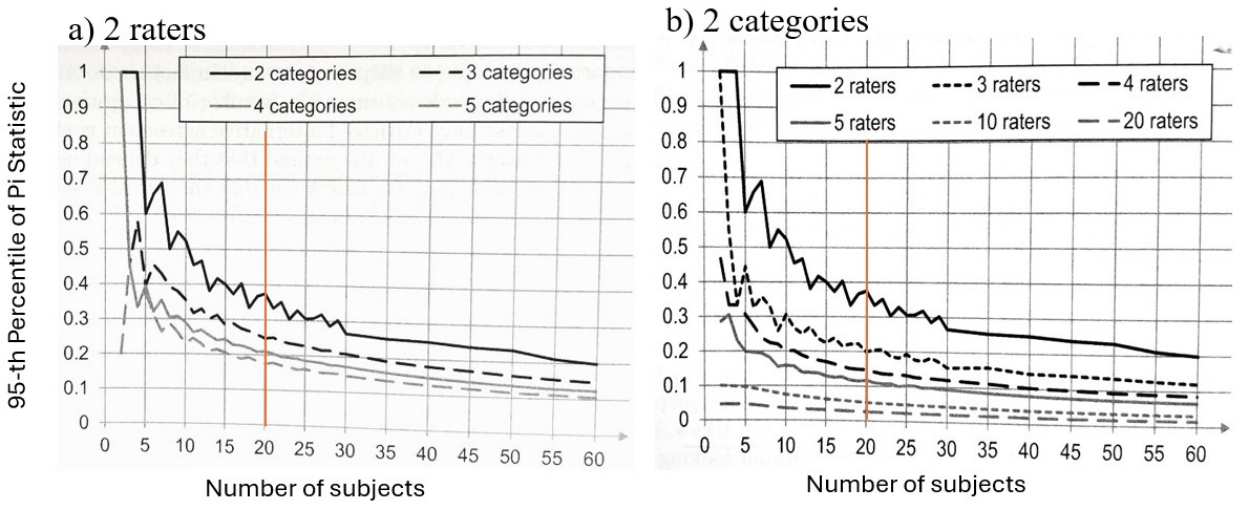
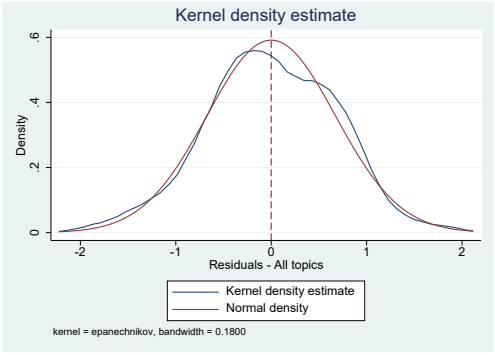
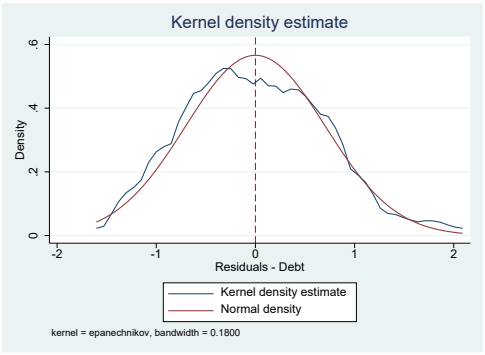


Figure 12: Random-Ratings Critical Values for Conger's π.

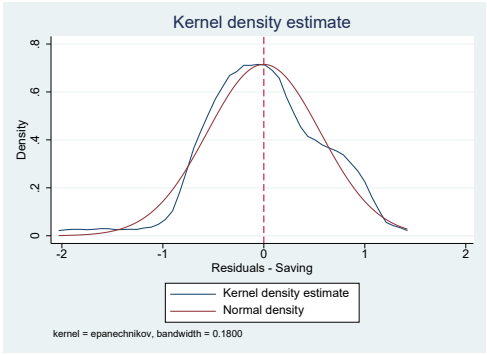
The figure reports the 95th percentile of Conger's π under a random-ratings (RR) model in which raters assign categories to subjects purely at random. The figure examines how the number of subjects (n), raters (r), and categories (q) affect the upper-tail behavior of the agreement coefficient. The RR-based 95th percentile provides a critical value: under random ratings, Conger's π exceeds this threshold with probability 5%. Accordingly, when the agreement coefficient computed from the observed ratings exceeds the RR-based 95th percentile, we reject the null of random (purely chance-driven) ratings and interpret the observed agreement as nonrandom. Panel (a) plots the RR-based 95th percentile of Conger's π for $r = 2$ as a function of n and q. Panel (b) plots the RR-based 95th percentile of Conger's π for $q = 2$ as a function of n and r; Source:Gwet (2014).



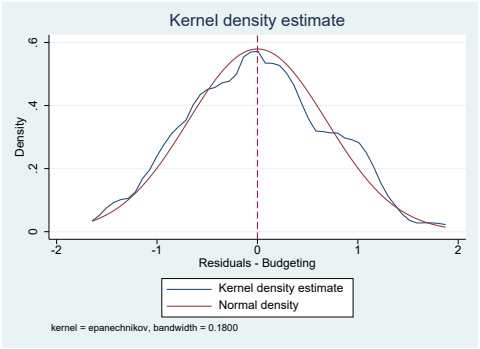
a) All Advisory Domains



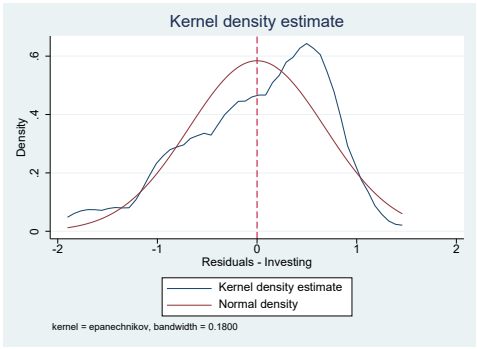
b) Borrowing



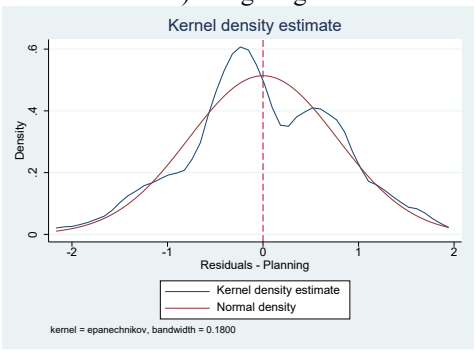
c) Saving



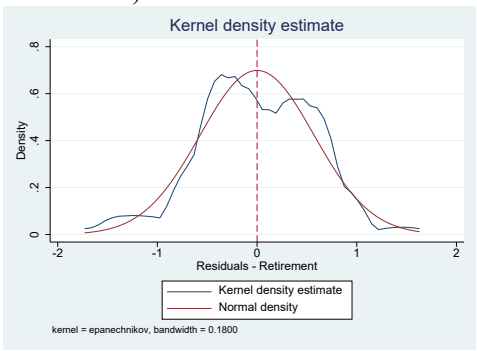
d) Budgeting



e) Portfolio Allocation



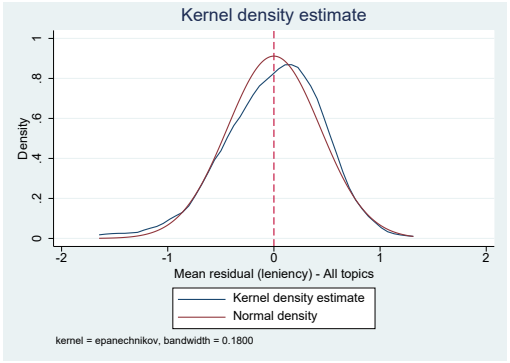
f) Planning



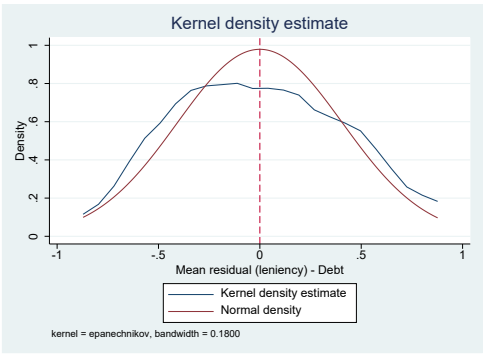
g) Retirement

Figure 13: Kernel Densities of Adviser Disagreement Residuals across Advisory Domain.

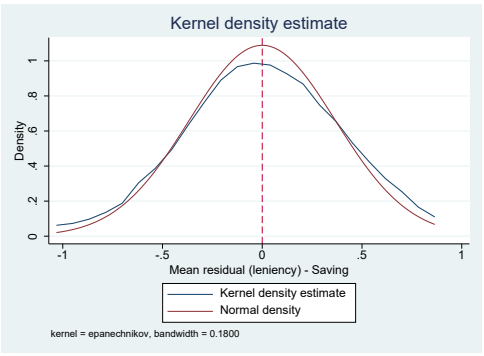
The figure plots kernel density estimates of the residuals $\hat{\varepsilon}_{ikj}$ from domain-specific regressions of adviser j's 1–5 quality rating y_{ikj} on a domain mean (μ), comment fixed effects (α_i), and adviser fixed effects (γ_j): $y_{ikj} = \mu + \alpha_i + \gamma_j + \varepsilon_{ikj}$. Panels report residual distributions for financial planning, budgeting, saving, investing (portfolio allocation), retirement saving, borrowing, and the pooled sample. Each panel also overlays the corresponding fitted normal density for visual benchmarking.



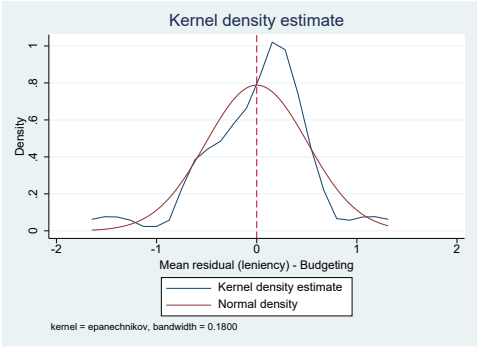
a) All Advisory Domains



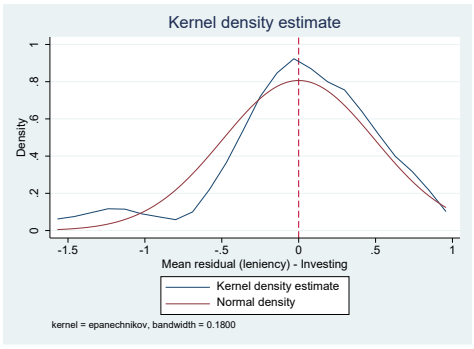
b) Borrowing



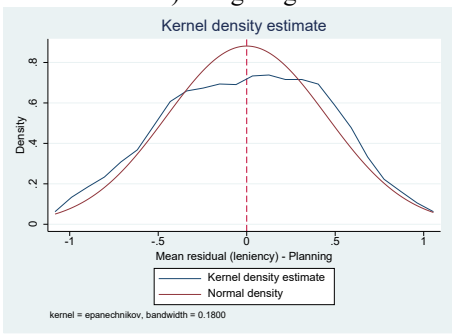
c) Saving



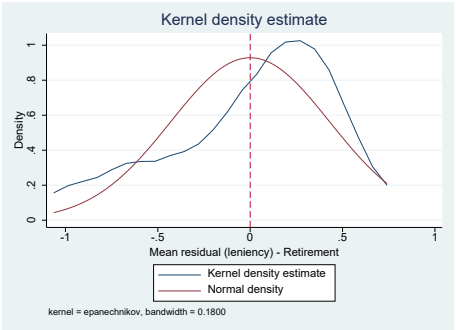
d) Budgeting



e) Portfolio Allocation



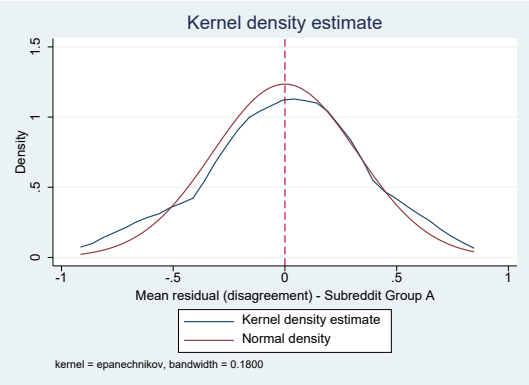
f) Planning



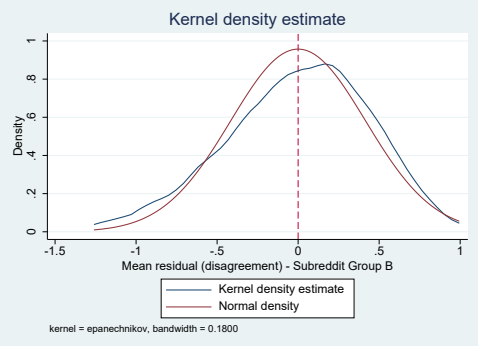
g) Retirement

Figure 14: Kernel Densities of Adviser-Level Mean Residual Disagreement (d_{kj}) by Domain.

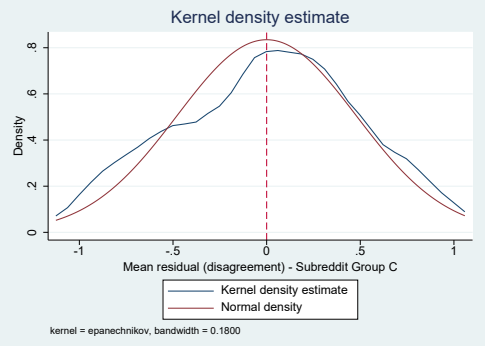
This figure plots kernel density estimates of d_{kj}, the average residual for adviser j within case k by advisory domain. $d_{kj} = \frac{1}{n_{kj}} \sum_i \hat{\varepsilon}_{ikj}$, where n_{kj} is the number of comments evaluated by adviser j in case k, using Equation 1. Greater dispersion of the densities indicates stronger cross-adviser disagreement in assessed advice quality.



a) Subreddit Group A



b) Subreddit Group B



c) Subreddit Group C

Figure 15: Kernel Densities of Adviser-Level Mean Residual Disagreement (d_{kj}) by Subreddit Group.

This figure plots kernel density estimates of d_{kj}, the average residual for adviser j within case k by subreddit group (A, B, and C). We compute $d_{kj} = \frac{1}{n_{kj}} \sum_i \hat{\varepsilon}_{ikj}$, where n_{kj} is the number of comments evaluated by adviser j in case k, using Equation 1. Group A represents high earners and high-net-worth individuals; Group B comprises the three largest general personal-finance subreddits open to users across the socioeconomic spectrum; and Group C consists of subreddits oriented toward middle-class and financially constrained users, with an emphasis on low-income individuals. Greater dispersion of the densities indicates stronger cross-adviser disagreement in assessed advice quality.

Table 1: Reddit Data

The table summarizes advice seekers attributes for 156,944 posts drawn from the ten largest personal-finance forums on Reddit between January 2022 and December 2023, with demographic and financial characteristics extracted from the post text using ChatGPT-4.0.1.

a) Age

Age	Count	Percentage
20 years and younger	5,252	3.5%
21–30years	22,213	14.3%
31–40years	10,557	5.9%
41–50years	2,757	1.3%
51–60years	1,507	1.0%
61–70years	923	0.7%
71 years and older	325	0.3%
Not mentioned	129,533	73.1%
Total	156,944	100.0%

b) Gender

Gender	Count	Percentage
Female	4,804	2.9%
Male	9,387	5.0%
non-binary	10	0.0%
Not mentioned	142,743	92.2%
Total	156,944	100.0%

c) Domain

Domain (topic)	Count	Percentage
Budgeting	14,964	9.1%
Debt	23,791	22.0%
Investing	42,304	20.2%
Other	16,311	0.4%
Planning	22,250	11.3%
Retirement	26,258	25.7%
Saving	11,066	11.3%
Total	156,944	100.0%

d) Net income

Net income	Count	Percentage
Lower class	3,246	2.4%
Lower-middle class	4,855	3.7%
Middle class	7,448	5.9%
Upper-middle class	6,022	4.3%
Upper class 1	1,472	0.8%
Upper class 2	5,984	2.9%
Not mentioned	127,917	79.9%
Total	156,944	100.0%

e) Net worth

Net worth	Count	Percentage
Lower class	580	0.4%
Lower-middle class	5,011	3.6%
Middle class	3,848	2.4%
Upper-middle class	2,351	1.0%
Upper class	2,851	0.5%
Not mentioned	142,303	92.1%
Total	156,944	100.0%

f) Risk tolerance

Risk Tolerance	Count	Percentage
Aggressive	1,858	0.9%
Moderate	760	0.4%
Conservative	1,697	1.3%
Not mentioned	152,629	97.4%
Total	156,944	100.0%

Table 2: Expert Advisers: Background and Assignment load

Adviser	Experience (Years)	Area(s) of expertise	Years as CFP®	Role	Gender	U.S. Region	Assigned Cases
1	33	All	19	Retired Senior Vice President, Wealth Advisor	Female	South	12
2	11	All	7	Wealth Advisor	Female	Midwest	24
3	17	Planning, Budgeting, Investing, Saving, Retirement	10	Lead Advisor & Finance Adjunct Faculty	Female	South	12
4	30	All	32	Executive Vice President	Male	Pacific	12
5	35	All	36	Retired Wealth Advisor	Female	Mountain	12
6	33	All	21	Retired Senior Compliance Consultant	Male	Northeast	36
7	20	All	11	Senior Vice President, Wealth Management Advisor	Male	Pacific	24
8	10	All	5	Senior Financial Planner & Adjunct Professor of Finance	Male	Mountain	12

Notes: (a) Affiliations are omitted to preserve adviser anonymity. (b) State identifiers are aggregated to regions for the same purpose. (c) "All" in the "Area(s) of expertise" column denotes expertise spanning all six advisory domains in the study. (d) "Assigned Cases" reports the number of cases assigned to each adviser.

Table 3: Descriptive Statistics

 # POSTS

Panel A) Quality Score of posts across advisory domains

Advisory domain	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
Budgeting	3	3	3.5	4.5	5	5	5	4.250	0.886	8
Debt	4	4	4	5	5	5	5	4.625	0.518	8
Investing	3	3	4	4	4.5	5	5	4.125	0.641	8
Planning	2	2	3.5	4	5	5	5	4.000	1.069	8
Retirement	3	3	4	4	5	5	5	4.250	0.707	8
Saving	2	2	3	4	4.5	5	5	3.750	1.035	8
Total	2	3	4	4	5	5	5	4.167	0.834	48

Panel B) Quality Score of posts across subreddit groups

Subreddit group	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
A	3	4	4	4	5	5	5	4.333	0.651	12
B	2	3	3	4	4.5	5	5	3.875	0.900	24
C	3	4	4	5	5	5	5	4.583	0.669	12
Total	2	3	4	4	5	5	5	4.167	0.834	48

 # COMMENTS

Panel C) Quality Score of Comments across advisory domains

Advisory domain	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
Budgeting	1	1	2	3	4	4	5	3.030	1.136	66
Debt	1	1	2	3	3	4	5	2.773	1.005	66
Investing	1	1	2	3	4	4	5	2.889	1.056	72
Planning	1	2	2	3	4	4	5	3.014	1.094	72
Retirement	1	1	3	3	4	4	5	3.150	1.087	60
Saving	1	1	2	3	4	4	5	3.130	1.187	69
Total	1	1	2	3	4	4	5	2.995	1.097	405

Panel D) Quality Score of Comments across subreddit groups

Subreddit group	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
A	1	1	2	3	3.5	4	5	2.713	1.068	108
B	1	1	2	3	4	4	5	3.055	1.050	201
C	1	1	2	3	4	5	5	3.188	1.173	96
Total	1	1	2	3	4	4	5	2.995	1.097	405

Panel E) Similarity of Comments across subreddit groups

Subreddit group	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
A	1	2	2	3	4	4	5	3.056	0.975	108
B	1	1	2	3	4	4	5	2.844	1.110	199
C	1	2	3	3	4	4	5	3.308	0.939	91
Total	1	1	2	3	4	4	5	3.008	1.052	398

Panel F) Quality Score of Comments across advisers

Adviser	Min	p10	p25	p50	p75	p90	Max	Mean	SD	N
1	1	1	2	3	4	5	5	3.143	1.216	35
2	1	1	2	3	4	5	5	2.851	1.197	67
3	1	2	3	3.5	4	5	5	3.344	1.125	32
4	1	3	3	3.5	4	4	5	3.469	0.879	32
5	1	1	2	3	4	5	5	2.941	1.254	34
6	1	2	3	3	4	4	5	3.049	0.772	103
7	1	1	2	3	4	4	5	2.809	1.162	68
8	1	1	1	3	3	4	5	2.618	1.280	34
Total	1	1	2	3	4	4	5	2.995	1.097	405

Table 4: Two-Way Fixed-Effects Regressions by Subreddit Group

This table reports results from OLS regressions with fixed effects modeling advisers' quality scores for socially sourced advice (comments) in the 48 Reddit cases. We estimate $y_{ikj} = \mu + \alpha_i + \gamma_j + \varepsilon_{ikj}$ for the full sample and separately by subreddit group (Groups A, B, and C), where y_{ikj} is the 1–5 quality rating ($1 =$ "Very Poor," $5 =$ "Excellent") assigned by adviser j to comment i in case k, μ captures the domain mean, α_i are comment fixed effects, and γ_j are adviser fixed effects. Group A consists of subreddits targeting high earners and high-net-worth individuals; Group B comprises large general personal-finance subreddits spanning the socioeconomic spectrum; and Group C includes subreddits oriented toward middle-class and financially constrained users.

Advisory Domain	**All**	**Group A**	**Group B**	**Group C**
Constant	2.995***	2.713***	3.055***	3.188***
t-statistic	(72.12)	(41.00)	(51.76)	(32.37)
Comment FEs	Yes	Yes	Yes	Yes
Adviser FEs	Yes	Yes	Yes	Yes
R^2	0.622	0.748	0.597	0.594
Adjusted R^2	0.419	0.586	0.365	0.323
Observations	405	108	201	96

Notes: Comment and adviser fixed effects included in all specifications. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 5: Two-Way Fixed-Effects Regressions by Advisory Domain

This table reports results from OLS regressions with fixed effects modeling advisers' quality scores for socially sourced advice (comments) in the 48 Reddit cases. We estimate $y_{ikj} = \mu + \alpha_i + \gamma_j + \varepsilon_{ikj}$ separately for each advisory domain—financial planning, budgeting, saving, investing (portfolio allocation), retirement saving, and borrowing (debt management)—where y_{ikj} is the 1–5 quality rating ($1 =$ "Very Poor," $5 =$ "Excellent") assigned by adviser j to comment i in case k, μ captures the domain mean, α_i are comment fixed effects, and γ_j are adviser fixed effects.

Advisory Domain	**Retirement**	**Saving**	**Budgeting**	**Planning**	**Investing**	**Borrowing**
Constant	3.15***	3.13***	3.03***	3.014***	2.889***	2.773***
t-statistic	(34.297)	(37.062)	(28.05)	(26.511)	(28.577)	(25.697)
Comment FEs	Yes	Yes	Yes	Yes	Yes	Yes
Adviser FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.724	0.779	0.633	0.496	0.582	0.508
Adjusted R^2	0.571	0.651	0.404	0.223	0.340	0.239
Observations	60	69	66	72	72	66

Notes: Comment and adviser fixed effects included in all specifications. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 6: Determinants of the Quality of Socially Sourced Financial Advice on Reddit

This table reports estimates from the following three-way fixed-effects regression, estimated using 405 observations: $y_{ikj} = \beta_1 \ln(\text{CommentVotes} + 1) + \beta_2 \text{Top1\%Commenter}_i + \beta_3 \ln(\text{PostVotes} + 1) + \beta_4 \ln(\text{PostWords} + 1) + \beta_5 \text{FinLit}_k + \beta_6 \text{SubredditGroup}_k + \alpha_t + \delta_d + \gamma_j + \varepsilon_{ikj}$.
Where y_{ikj} is the quality score ($1 =$ Very Poor; $5 =$ Excellent) assigned by adviser j to comment i in case (post) k. CommentVotes and PostVotes denote net scores (upvotes minus downvotes) for comments and posts, respectively, capturing crowd-based attention. Top1%Commenter is an indicator equal to one if the commenter ranked in the top 1% by cumulative comment upvotes at the end of the previous month. PostWords measures post length. FinLit is the adviser-assessed financial literacy of the advice seeker (1–5 scale). SubredditGroup classifies subreddits into Group A (baseline; high earners/high-net-worth), Group B (large general personal-finance), and Group C (middle-class/financially constrained). The specification includes year fixed effects (α_t), advisory-domain fixed effects (δ_d), and adviser fixed effects (γ_j). Standard errors are clustered at the case level.

Regressor	**Coefficient**	**Standard error**	**t-statistic**	**P$>\|t\|$**	**[95% conf. interval]**	
					Lower	**Upper**
ln(Comment's vote + 1)	−0.074	0.061	−1.22	0.228	−0.196	0.048
Top 1% Commenter	0.260	0.150	1.74	0.089	−0.041	0.560
ln(Post's vote+1)	0.094	0.048	1.95	0.057	−0.003	0.192
ln(Post's word count + 1)	0.225	0.065	3.46	0.001	0.094	0.356
Financial Literacy						
2	0.480	0.263	1.83	0.074	−0.048	1.009
3	1.136	0.267	4.25	0.000	0.599	1.674
4	1.648	0.261	6.32	0.000	1.124	2.172
5	2.115	0.347	6.09	0.000	1.417	2.813
Subreddit						
B	0.347	0.105	3.32	0.002	0.137	0.557
C	0.289	0.162	1.79	0.080	−0.036	0.614
Intercept	0.305	0.443	0.69	0.494	−0.586	1.197
Year fixed effects			Yes			
Domain fixed effects			Yes			
Adviser fixed effects			Yes			
R^2			0.252			
Adjusted R^2			0.207			
Within R^2			0.215			

Table 7: Generalized Ordered Logit Estimates of Advice Quality by Subreddit Group

This table reports relative risk ratios (RRRs) for advice quality across subreddit groups (capturing the socioeconomic status of advice seekers) from the generalized ordered logit model in Equation (5) with non-parallel lines. The dependent variable is a three-category advice-quality classification constructed from professional advisers' 1–5 ratings: Bad (1–2), Moderate (3), and Good (4–5). Panel A reports the cumulative threshold comparing Moderate/Good versus Bad; Panel B reports the cumulative threshold comparing Good versus Bad/Moderate. Group A (high earners and high-net-worth subreddits) is the omitted subreddit category. The specification includes adviser indicators and advisory-domain indicators (with "borrowing" as the base category); these are entered as dummy variables and are not absorbed fixed effects. Standard errors are clustered at the case (post) level. The sample consists of 405 observations across 48 cases. Wald $\chi^2 = 100.82$ and pseudo $R^2 = 0.0674$.

	Relative Risk Ratio (RRR)					
	RRR	Std. Err.	z	p	95% CI	
					Lower	Upper
Panel A: Threshold 1 (Moderate/Good vs. Bad)						
Subreddit group (omitted: Group A)						
Group B	1.8172	0.5754	1.89	0.059	0.9769	3.3802
Group C	1.3821	0.6774	0.66	0.509	0.5289	3.6117
Panel B: Threshold 2 (Good vs. Bad/Moderate)						
Subreddit group (omitted: Group A)						
Group B	1.5082	0.4649	1.33	0.183	0.8243	2.7597
Group C	2.5043	0.8895	2.58	0.010	1.2484	5.0238

Table 8: Pairwise Contrasts in Predicted Advice Quality Across Subreddit Groups

This table reports pairwise contrasts in predicted probabilities of advice quality categories across subreddit groups, computed from the generalized ordered logit model in equation (5). The dependent variable is a three-category advice-quality classification constructed from professional advisers' 1–5 ratings: Bad (ratings 1–2), Moderate (rating 3), and Good (ratings 4–5). Group A consists of subreddits targeting high earners and high-net-worth individuals; Group B comprises large general personal-finance subreddits spanning the socioeconomic spectrum; and Group C consists of subreddits oriented toward middle-class and financially constrained users. Reported contrasts correspond to differences in predicted probabilities between subreddit groups (e.g., B vs. A), expressed in percentage points. Predicted probabilities are computed by averaging over the observed distribution of covariates, including adviser indicators and advisory-domain indicators (with "borrowing" as the base category), which enter the model as dummy variables and are not absorbed fixed effects. Standard errors are computed using the delta method and are clustered at the case (post) level to account for correlation across comments associated with the same advice request. The sample consists of 405 observations across 48 cases.

a) Pr(Advice Quality Category = bad)

Subreddit Group	Contrast	Delta-method std.	Z	P$>\|z\|$	[95% conf. interval]	
					Lower	Upper
B vs A	−0.119	0.065	−1.84	0.066	−0.246	0.008
C vs A	−0.068	0.101	−0.67	0.503	−0.265	0.130
C vs B	0.052	0.086	0.60	0.546	−0.116	0.220

b) Pr(Advice Quality Category = moderate)

Subreddit Group	Contrast	Delta-method std.	Z	P$>\|z\|$	[95% conf. interval]	
					Lower	Upper
B vs A	0.036	0.058	0.62	0.537	−0.078	0.150
C vs A	−0.132	0.082	−1.62	0.106	−0.292	0.028
C vs B	−0.168	0.066	−2.55	0.011	−0.297	−0.039

c) Pr(Advice Quality Category = good)

Subreddit Group	Contrast	Delta-method std.	Z	P$>\|z\|$	[95% conf. interval]	
					Lower	Upper
B vs A	0.083	0.061	1.36	0.175	−0.037	0.204
C vs A	0.199	0.077	2.60	0.009	0.049	0.350
C vs B	0.116	0.077	1.51	0.132	−0.035	0.267

Table 9: Out-of-Sample Classification Performance: Classical ML vs. LLM

This table reports out-of-sample classification performance using a stratified 90/10 train–test split. Panel A presents results from the classical machine-learning model (TF–IDF features with logistic regression). Panel B reports results from the large language model (GPT-4.1). Classes correspond to rofessional ratings: Bad (1), Moderate (2), and Good (3). Accuracy, precision, recall, and F1-scores are computed on the held-out test set. Precision measures the fraction of observations classified into a given category that truly belong to that category. Recall measures the fraction of true observations in a category that are correctly classified. The F1-score is the harmonic mean of precision and recall and summarizes the trade-off between false positives and false negatives for each category. Accuracy measures the fraction of observations in the test set that are correctly classified across all categories. Weighted accuracy measures overall classification accuracy after weighting class-specific accuracies by each category's support.

	Precision	Recall	F1-score	Support
Panel A: Classical Machine Learning (Logistic Regression + TF–IDF)				
Bad	0.50	0.17	0.25	12
Moderate	0.38	0.67	0.49	15
Good	0.64	0.50	0.56	14
Accuracy			0.46	41
Macro Avg.	0.51	0.44	0.43	41
Weighted Avg.	0.50	0.46	0.44	41
Panel B: Large Language Model (GPT-4.1)				
Bad	0.80	0.67	0.73	12
Moderate	0.61	0.65	0.63	15
Good	0.69	0.75	0.72	14
Accuracy			0.68	41
Macro Avg.	0.70	0.69	0.69	41
Weighted Avg.	0.69	0.68	0.68	41

Table 10: GPT-4.1 API: Main Criteria Used for Predicting Professionals' Ratings

Good Advice	**Bad Advice**
Specific, tailored, actionable steps	Generic, vague, or one-size-fits-all
Detailed rationale, not just conclusions	Lacking explanation or detail
Incorporates all relevant facts/context	Ignores (or is unaware of) the OP's situation
Quantifies impacts, uses numbers	Speculative, unsubstantiated, "rule of thumb"
Considers alternatives and trade-offs	Presents only binary or limited options
Highlights caveats, limitations, risks	Glosses over risks or gives misleading assurance
Professional and nonjudgmental tone	Judgmental, condescending, dismissive
Corrects errors, avoids misinformation	Contains factual errors or outdated advice
Suggests further resources/expertise	Fails to refer when out of depth
Encourages planning and understanding	Relies on anecdotes or own experience only

Table 11: Inter-Rater Reliability Metrics by Advisory Domain

This table reports inter-rater reliability (IRR) statistics for advisers' ordinal (five-category) quality ratings, computed separately by advisory domain. Each subject (comment) is rated by three independent advisers. Each panel corresponds to one domain and lists the number of subjects (n), along with the estimated agreement coefficient (Coef.), its unconditional standard error, the associated z-statistic and p-value, and the 95% confidence interval. "Percent agreement" is the observed *weighted* agreement rate under the study's partial-agreement weighting matrix, which assigns full credit to exact matches and partial credit to near matches. The remaining measures—Brennan–Prediger, Cohen/Conger's κ, Scott/Fleiss' κ, Gwet's AC, and Krippendorff's α—are chance-corrected agreement coefficients computed using the same weighting scheme.

Panel A: Borrowing ($n = 22$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.583	0.126	4.64	0.000	[0.337, 0.830]
Brennan–Prediger	0.256	0.224	1.14	0.254	[-0.184, 0.696]
Cohen/Conger's kappa	0.161	0.188	0.86	0.389	[-0.206, 0.529]
Scott/Fleiss' kappa	0.110	0.261	0.42	0.673	[-0.401, 0.622]
Gwet's AC	0.324	0.230	1.41	0.158	[-0.126, 0.774]
Krippendorff's alpha	0.124	0.256	0.48	0.628	[-0.378, 0.625]

Panel B: Saving ($n = 23$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.761	0.083	9.20	0.000	[0.599, 0.923]
Brennan–Prediger	0.573	0.148	3.88	0.000	[0.284, 0.862]
Cohen/Conger's kappa	0.519	0.172	3.02	0.003	[0.182, 0.857]
Scott/Fleiss' kappa	0.529	0.168	3.15	0.002	[0.200, 0.858]
Gwet's AC	0.596	0.143	4.17	0.000	[0.315, 0.876]
Krippendorff's alpha	0.536	0.166	3.24	0.001	[0.211, 0.860]

Panel C: Budgeting ($n = 22$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.614	0.092	6.70	0.000	[0.434, 0.793]
Brennan–Prediger	0.310	0.164	1.89	0.058	[-0.011, 0.631]
Cohen/Conger's kappa	0.281	0.175	1.61	0.108	[-0.062, 0.624]
Scott/Fleiss' kappa	0.244	0.178	1.37	0.171	[-0.105, 0.593]
Gwet's AC	0.336	0.162	2.08	0.038	[0.019, 0.654]
Krippendorff's alpha	0.255	0.150	1.71	0.088	[-0.038, 0.548]

Panel D: Investing ($n = 24$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.632	0.067	9.48	0.000	[0.501, 0.763]
Brennan–Prediger	0.343	0.119	2.88	0.004	[0.109, 0.576]
Cohen/Conger's kappa	0.296	0.158	1.88	0.061	[-0.013, 0.604]
Scott/Fleiss' kappa	0.247	0.124	1.99	0.047	[0.004, 0.491]
Gwet's AC	0.384	0.121	3.19	0.001	[0.148, 0.621]
Krippendorff's alpha	0.257	0.118	2.18	0.030	[0.026, 0.489]

Panel E: Planning ($n = 24$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.569	0.134	4.24	0.000	[0.306, 0.833]
Brennan–Prediger	0.231	0.240	0.96	0.335	[-0.239, 0.701]
Cohen/Conger's kappa	0.156	0.220	0.71	0.478	[-0.275, 0.587]
Scott/Fleiss' kappa	0.138	0.248	0.56	0.577	[-0.347, 0.624]
Gwet's AC	0.269	0.241	1.12	0.263	[-0.202, 0.740]
Krippendorff's alpha	0.150	0.243	0.62	0.537	[-0.327, 0.627]

Panel F: Retirement ($n = 20$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.708	0.127	5.57	0.000	[0.459, 0.958]
Brennan–Prediger	0.479	0.227	2.11	0.035	[0.034, 0.924]
Cohen/Conger's kappa	0.407	0.222	1.83	0.067	[-0.028, 0.842]
Scott/Fleiss' kappa	0.391	0.252	1.55	0.120	[-0.102, 0.885]
Gwet's AC	0.521	0.203	2.57	0.010	[0.123, 0.920]
Krippendorff's alpha	0.401	0.247	1.62	0.104	[-0.083, 0.886]

Table 12: Inter-rater Reliability by Subreddit Group (Ordinal, Weighted)

This table reports inter-rater reliability (IRR) statistics for advisers' ordinal (five-category) quality ratings, computed separately by subreddit group: Group A (high earners and high-net-worth users), Group B (broad, heterogeneous personal-finance communities), and Group C (middle-class and financially constrained users). Each subject (comment) is rated by three independent advisers. Each panel corresponds to one domain and lists the number of subjects (n), along with the estimated agreement coefficient (Coef.), its unconditional standard error, the associated z-statistic and p-value, and the 95% confidence interval. "Percent agreement" is the observed *weighted* agreement rate under the study's partial-agreement weighting matrix, which assigns full credit to exact matches and partial credit to near matches. The remaining measures—Brennan–Prediger, Cohen/Conger's κ, Scott/Fleiss' κ, Gwet's AC, and Krippendorff's α—are chance-corrected agreement coefficients computed using the same weighting scheme.

Panel A: Group A ($n = 36$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.681	0.079	8.61	0.000	[0.526, 0.835]
Brennan–Prediger	0.430	0.141	3.04	0.002	[0.153, 0.706]
Cohen/Conger's kappa	0.374	0.162	2.31	0.021	[0.056, 0.692]
Scott/Fleiss' kappa	0.355	0.161	2.21	0.027	[0.040, 0.669]
Gwet's AC	0.463	0.133	3.47	0.001	[0.201, 0.725]
Krippendorff's alpha	0.361	0.146	2.47	0.014	[0.074, 0.647]

Panel B: Group B ($n = 67$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.624	0.116	5.37	0.000	[0.396, 0.852]
Brennan–Prediger	0.329	0.208	1.59	0.113	[-0.078, 0.736]
Cohen/Conger's kappa	0.236	0.233	1.01	0.310	[-0.220, 0.693]
Scott/Fleiss' kappa	0.221	0.217	1.02	0.308	[-0.204, 0.647]
Gwet's AC	0.377	0.203	1.85	0.064	[-0.022, 0.775]
Krippendorff's alpha	0.225	0.201	1.12	0.263	[-0.169, 0.619]

Panel C: Group C ($n = 32$)

Metric	Coef.	Std. Err.	z	p	95% CI
Percent agreement	0.641	0.108	5.92	0.000	[0.429, 0.853]
Brennan–Prediger	0.358	0.193	1.85	0.064	[-0.020, 0.737]
Cohen/Conger's kappa	0.327	0.191	1.71	0.087	[-0.047, 0.701]
Scott/Fleiss' kappa	0.292	0.211	1.38	0.167	[-0.122, 0.705]
Gwet's AC	0.383	0.187	2.05	0.040	[0.018, 0.749]
Krippendorff's alpha	0.299	0.173	1.73	0.083	[-0.039, 0.637]

Internet Appendix

A.1 Instructions for the Experiment

Purpose of the Study

This study examines the quality of financial advice offered on social media. Specifically, the focus of this study is social forums on Reddit for personal finance, which are dedicated to topics including financial planning, budgeting, investing, saving, retirement, and debt management.

Socially Sourced Financial Advice on Reddit

Reddit is a social platform for news and discussion, organized into communities called subreddits, each dedicated to a specific topic. In personal finance subreddits, users seek tailored financial advice by submitting online posts that outline their financial situations and goals while avoiding the disclosure of sensitive personal information. Community members then discuss and vote on these posts, influencing their visibility within the subreddit.

Structure of Discussion on a Subreddit

A "thread" refers to a conversation initiated by a post. Users can reply to the original post and to one another, creating a hierarchical, tree-like structure. The original post appears at the top, with subsequent replies branching below, allowing users to engage in focused discussions by responding to specific comments. Each thread consists of the following elements, as illustrated in Figure A.1:

- *Originating Poster (OP)*: The author of the initial post that starts the thread.
- *Comments*: Replies to the OP or other comments within the thread.
- *Hierarchical Structure*: Comments are nested, enabling direct replies to specific comments, forming a visual hierarchy. This structure facilitates interaction, as community members provide advice, seek clarification, or refine recommendations. The OP can engage by responding to comments, offering further details, and contributing to the evolving discussion.
- *Voting System*: Community members, including the OP, can vote on comments, influencing their ranking. Comments with the most votes are often perceived as the most valuable or actionable, shaping the prominence of advice within the thread.

Study Methodology

The study design involves two main components:

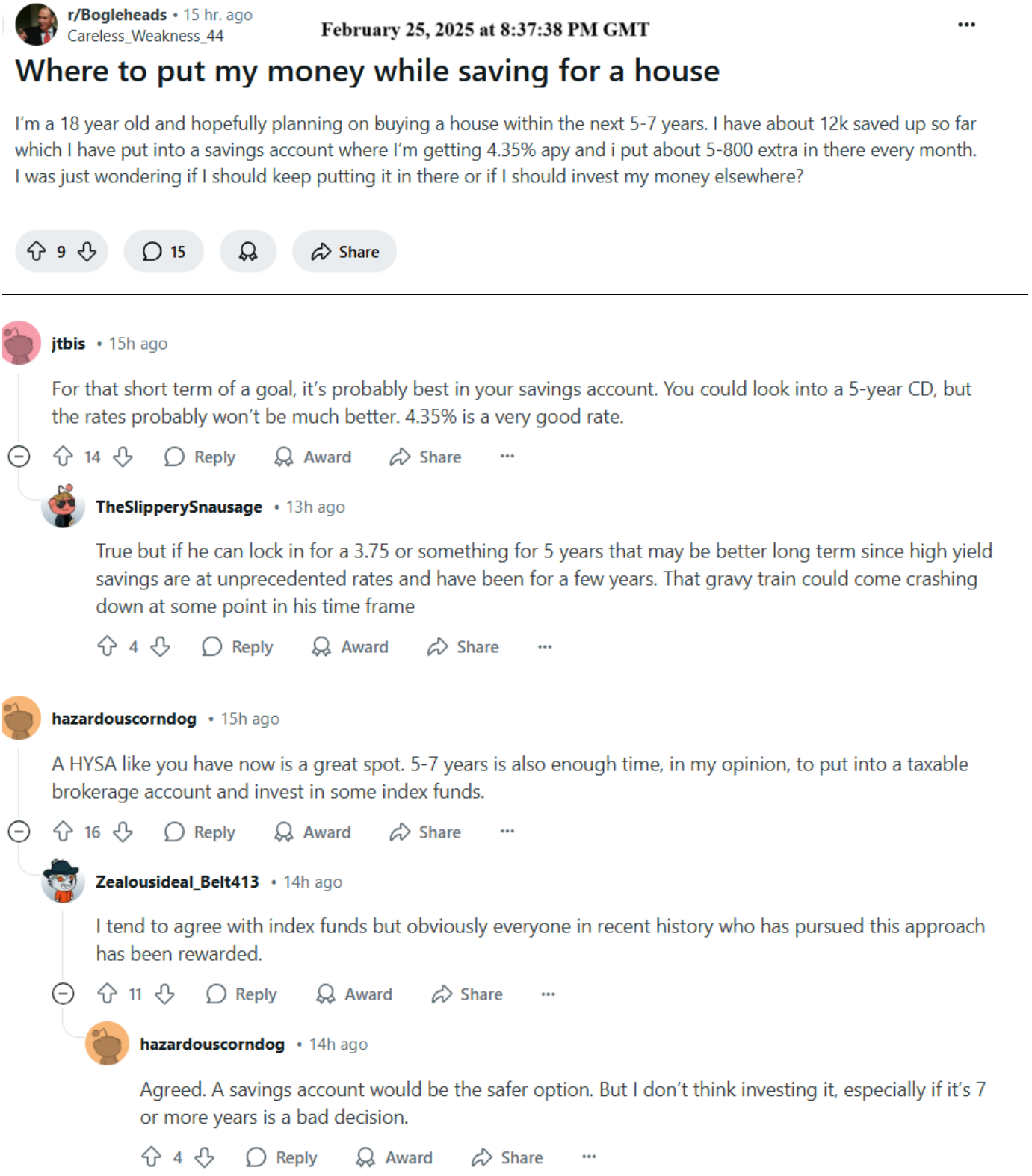


Figure A.1: An Example of Reddit Threads.

- *Selected Cases (threads)*: This study examines the quality of financial advice offered on social media, using cases drawn from the most active Reddit forums (subreddits) focused on personal finance from January 2022 to December 2023.[1] Each case (thread) includes an online request for tailored financial advice (original post), accompanied by the three top community responses (comments). Notably, some posts may have fewer than three comments.

- *CFP® Professional Evaluation*: CFP® Board has partnered with us to help identify CFP® professional evaluators. Selected CFP® evaluators will be asked to produce brief reports on 12 cases aligned with their area(s) of expertise (financial planning, budgeting, investing, saving, retirement, or debt management). Cases will be distributed individually in PDF format, each labeled with a unique identification number (Case ID). Evaluations are to be submitted online via a Google Form (click here to access the form). Each evaluation consists of six rating questions and four open-ended questions, the latter requiring justifications for the ratings provided.

Your Role in This Study

As a CFP® professional, your expertise is essential in assessing the quality of financial advice provided on social media platforms. Beyond academic guidelines, your experience allows you to evaluate advice for its real-world applicability and nuance. You will contribute to this study by evaluating 12 cases as follows:

1. *Advisor ID*: Your unique identification number (Advisor ID) is "———". It will be used to preserve your anonymity.

2. *Case Files*: You will receive 12 PDF files, each containing a case with a unique identification number (Case ID).

3. *Assessment Submission*: For each case, you will provide a brief evaluation and submit it via a Google Form (click here to access the form). Each evaluation is expected to take approximately 30 minutes. While some cases may require less time and others slightly more, the total time commitment for evaluating all 12 cases should not exceed 6 hours.

4. *Unbiased Evaluation*: Each case is to be evaluated at least by 3 CFP® professionals. We would like the evaluations to be professional, as if they were solicited by a court as expert witnesses, which is why we (and our sponsor, the Leverhulme Foundation) are prepared to compensate the members that would participate in our study.

[1] Our study includes the following subreddits: Personal Finance, Financial Planning, Bogleheads, FIRE, HENRY Finance, FatFIRE, FIREyFemmes, RichPeople Personal Finance, Middle Class Finance, and Poverty Finance.

5. *Context-Based Evaluation*: Each case includes a date-time stamp to serve as a reference point for your evaluation. You are required to assess each case as if you were the advice giver at the time the post was made, considering only the regulatory, political, and economic context that existed at that moment. Evaluations should be based solely on contemporaneous conditions; hindsight must not influence your assessment.

6. Finally, all 12 reports must be submitted no later than May 15, 2025.

Please note the following important points:

- *Hyperlinks*: Posts and comments may contain hyperlinks. It is recommended to check these links for a complete understanding.
  - Open PDF files in an internet browser rather than Adobe Reader, as Adobe may block hyperlinks by default. Adjust settings in Adobe if necessary to enable link functionality.
  - Some links may lead to Reddit.com or other websites. If a Reddit page displays an error initially, wait a few seconds. If it does not load automatically, refresh the page and wait briefly.
- *Identification of the Original Poster (OP)*: You can recognize the OP's replies through the "OP" tag next to their username. The OP often responds to comments to provide additional information or clarification, thus facilitating a discussion between the advice seeker (OP) and advice givers.
- *Community Voting on Advice*: Disagreement among advice givers may occur. Consider that the Reddit community evaluates advice through a voting system. Comments with more upvotes are generally regarded as more valuable, while those with negative scores often indicate advice that the majority find incorrect or unhelpful.
- *Case ID and Advisor ID*: Please ensure that you enter both the Case ID and your Advisor ID correctly in the assessment form. Errors in these entries may result in the deletion of your case assessment from the system

Confidentiality of Professionals Engaged in the Study

We are committed to protecting the confidentiality and privacy of all CFP® professionals engaged in this study. To achieve this, we will: Data Protection: Any data or insights provided by professionals will be securely stored and accessible only to authorized personnel involved in the project.

Anonymity: The identities of all professionals will be protected throughout the study. Any publications, reports, or presentations will ensure anonymity by using aggregated information or pseudonyms.

Data Retention and Disposal: We will adhere to data protection requirements during collection, storage, and use. In line with best practices and relevant regulations, any retained data will be safeguarded and disposed of securely when no longer needed.

Data Usage: Any data provided will be used only for the purpose of informing the study and achieving its objectives. It will not be used for any purpose outside the scope of this work without prior consent.

Honorarium

As part of your participation in this study, you will be awarded an honorarium of XXX to be paid directly by XXX. This honorarium is provided as a token of appreciation for your professional insights and the time you dedicate to supporting this research project. Please note that your involvement in this study is entirely voluntary and is intended to support academic research.

Final Note

We extend our sincere thanks to the CFP® Board for their valuable partnership in identifying CFP® professionals for this study. As a CFP® professional, your expertise is central to the success of this research, and we greatly appreciate your time and insights.

Thank you for your participation.

Assessment Form

Please use this form to submit your evaluation for each case, referencing your Adviser ID and the corresponding Case ID.

* Indicates required question

Advisor ID: *

Your answer

Case ID: *

Your answer

1) On a scale of 1 to 5, where 1 is "Very Poor" and 5 is "Excellent," how would you rate the quality of the original post: *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

2) Please explain the reasoning behind your assessment of the original post: *

Your answer

3) On a scale of 1 to 5, where 1 is "Very Poor" and 5 is "Excellent," how would you rate the quality of the advice provided by First Comment: *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

4) Please explain the reasoning behind your assessment of the first comment: *

Your answer

5) On a scale of 1 to 5, where 1 is "Very Poor" and 5 is "Excellent," how would you rate the quality of the advice provided by Second Comment:

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

6) Please explain the reasoning behind your assessment of the second comment:

Your answer

7) On a scale of 1 to 5, where 1 is "Very Poor" and 5 is "Excellent," how would you rate the quality of the advice provided by Third Comment:

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

8) Please explain the reasoning behind your assessment of the third comment:

Your answer

9) On a scale of 1 to 5, with 1 being "Very Different" and 5 being "Very Similar," how would you rate the similarity of the advice provided in comments 1 to 3? *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

10) On a scale of 1 to 5, where 1 indicates "Very Poor" and 5 indicates "Excellent," how would you rate the general financial literacy of the advice seeker in this case? *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Submit Clear form

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. - Contact form owner - Terms of Service - Privacy Policy

Does this form look suspicious? Report

Google Forms

A.2 An Illustrative Case from the 48-Case Sample

Retirement # Case ID: 18fn0da

Original Post:



r/Fire • 1 yr. ago
investurug

Date-Time: "Monday, December 11, 2023 at 5:14:42 AM GMT"

$4.8M NW, what do with money in the 401ks, IRAs after retirement?

Advice Request

Couple, I'm 51, wife 45. 3 young kids. Plan to retire in about 10 years. NW breakdown:

- Real estate ~ $1.1M (no mortgage)
- Cash/Stocks: ~$2.5M
- Retirement accounts: $1.2M

In the retirement accounts, $720k are index funds. Rest are dividend paying stocks. Our retirement strategy is all dividend stocks/ETFs income. All $2.5M stocks/cash port yields average ~6% right now. Question is what we do with the $720k index funds when the time comes. Do we buy dividend stocks all in or would you recommend gradually DCA a portfolio of stocks/EFTs over a period of time?

edit: I don't think income info matters but if needed, will provide. Our projected annual spend in 10 years would be around: $150k-$200k

0 19 Share

Comment 1:

FatFiredProgrammer • 1y ago

Top 1% Commenter

Why the focus on dividends? I'm not aware of any research showing they outperform. I do know that they reduce diversity and force a taxable event (at least in non tax advantaged accounts).

I'm strongly against holding individual stocks. GE was the godfather of dividend paying stocks until it wasn't. You're gambling. Period. And, imo, you're gambling based on the mistaken assumption that there's something unique or magical about dividends.

In my view, a dividend is a company throwing up its hands saying "We couldn't find a way to grow your money. Here's some money back for you to pay tax on."

If you want a dividend portfolio then, of course, that is your choice. My choice is - distilled - buy VTI or VOO.

11 Reply

investurug OP • 1y ago

Not against ETFs, will sure have a mix of them. The individual stocks we currently have are a basket of 25 stocks. Not as diversified as ETFs for sure, I got your point though.

1 Reply

Comment 2:

Key-Ad-8944 • 1y ago

Top 1% Commenter

I don't follow the issue with having $720k in index funds. You gradually liquidate your retirement assets in the most tax efficient manner. regardless of whether they are index funds or dividend stocks. Consider the following hypothetical.

- Index fund has 10%/year return with dividends reinvested, and pays 2% dividends. You do not reinvest the dividends and sell a total of 3%, so the principle increases by ~5%.
- Dividend stock has 10%/year return with dividends reinvested, and pays 5% dividends. You do not reinvest the 5% dividends, so the principle increases by ~5%.

Why is the first scenario a problem for you, but not the 2nd?

4 Reply

Comment 3:

noahsarc21 • 1y ago

The 2.5M are in taxable accounts ?

3 Reply

investurug OP • 1y ago

Yes.

1 Reply

noahsarc21 • 1y ago

I don't understand your question, is your question what to do with your funds when you withdraw them out of the retirement account ?

3 Reply

investurug OP • 1y ago

Pretty much. Because the funds in the retirement accounts don't generate dividend income. Suppose we could keep a portion of that still in the market to diversify after retirement. Our strategy is to all in on dividend generating stocks after retirement. We could be all wrong though.

-2 Reply

noahsarc21 • 1y ago

You are intending on withdrawing the funds slowly vs all at once correct ?

3 Reply

investurug OP • 1y ago

Either one. Not sure which one makes sense.

0 Reply

noahsarc21 • 1y ago

It doesn't make sense to withdraw all at the same time as it puts you into a much higher marginal bracket. You want to slowly withdraw while not incurring income from other sources to reduce your taxable impact

4 Reply

3 more replies

noahsarc21 • 1y ago

It doesn't make sense to withdraw all at the same time as it puts you into a much higher marginal bracket. You want to slowly withdraw while not incurring income from other sources to reduce your taxable impact

4 Reply ...

investurug OP • 1y ago

So a scheduled withdrawal would make more sense.

1 Reply

noahsarc21 • 1y ago

Yes

3 Reply ...

noahsarc21 • 1y ago

But that is secondary to your goal, you are saying here you have 2.5M cash and stocks doing 6% dividend yield or appreciation at the moment?

1 Reply ...

Collegeroids • 1y ago

What's your rational behind going all in on dividend stocks? Also what ETFs are you currently invested in? If the are broad marked ETFs then they definitely produce dividend income.

I am not aware of any research that shows that for using on dividend income outperforms focusing on total returns but if you're choosing to do that route for another reason let us know.

2 Reply ...

investurug OP • 1y ago

TBH, we don't have any rationale and feel stupid about it. We have an advisor meeting and will bring it up.

2 Reply ...

Collegeroids • 1y ago

Don't feel stupid. It's not a bad strategy depending on your goals/how you like to generate income from investments (some people prefer to receive income from dividends like receiving rent if you're landlord). Personally (and what a majority of this sub follows) is total returns from broad market ETFs.

2 Reply ...

noahsarc21 • 1y ago

And they do generate dividend income at a lower rate if they are ETFs around the 1-2% range of its VOO

1 Reply ...