Do AI Assistants Reinforce Investor Beliefs? Evidence from Human and AI Replies*

Swaminathan Balasubramaniam* Jorge Sabat† Soumyajit Ray‡

First prepared: 2025-09-21
This version: 2026-05-04

Abstract

Using Reddit `r/wallstreetbets` posts, we compare the sentiment of AI-generated replies with actual human top-level replies to the same investor-authored posts. We find that counterfactual replies generated by a user-facing AI model are systematically more bullish than human replies. Prompting the model to respond as a professional investor narrows the gap, but does not eliminate it. To study potential mechanisms, we compare the user-facing model with a matched base model that shares the same architecture and pretraining corpus but lacks instruction-following post-training and chat-format deployment. The instruction-tuned model is consistently more bullish than the matched base model, suggesting that the post-trained conversational layer amplifies bullishness beyond what is already present in the base model. We rationalize the evidence with a reduced-form model in which replies reflect a learned bullish default, responsiveness to the user's thesis, and categorical expression. Finally, we examine whether AI-augmented signals retain return-relevant content. The return evidence is descriptive: AI-augmented signals remain positively associated with near-horizon returns in the main specifications, but the relation is not uniform across samples or horizons.

Keywords: AI and LLMs; Investor Communication; Information Intermediaries; Sentiment; Retail Investors

JEL: G10, G12, G41, C55.

*Balasubramaniam: Finance Department, NEOMA Business School, Rouen, France; s.balasubramaniam@neoma-bs.fr. Sabat: Department of Economics and Business, Universidad Andrés Bello, Santiago, Chile, jorge.sabat@unab.cl. Ray: Department of Electrical Engineering, Johns Hopkins University, Baltimore, MD, rays@jhu.edu. We thank seminar participants at the 2025 AI in Finance Conference at Concordia University for useful suggestions. Sabat acknowledges financial support from FONDECYT Iniciación No. 11240130.

1 Introduction

Retail investors increasingly use conversational AI assistants to evaluate financial information, interpret market narratives, and form views about individual stocks.[1] These systems do more than retrieve information. When presented with an investment thesis, they generate an opinion: they may challenge the thesis, qualify it, or reinforce it. This matters in finance because beliefs, sentiment, and disagreement can affect prices and trading volume (De Long et al., 1990; Baker and Wurgler, 2006; Hong and Stein, 2007). If AI-generated opinions systematically frame investment arguments differently from human opinions, then conversational AI may affect financial communication through the way it responds, even holding fixed the investor's original information set. We ask whether user-facing AI replies differ systematically from human replies when responding to the same investment thesis.

We study this question using Reddit `r/wallstreetbets` (WSB), an online forum where retail investors discuss individual stocks. WSB provides two objects that are central to our design: investor-written posts about individual firms and actual human top-level replies to those same posts. We generate AI replies to the same posts using Mistral-7B-Instruct-v0.1, a seven-billion-parameter conversational language model in the style of current user-facing assistants, and compare the sentiment of the AI reply with the sentiment of the human replies observed in the original thread.

Our outcome is the reply's effect on conversation-level sentiment. Following Chen et al. (2025), we use a probabilistic probing procedure: a language model is asked to infer, from a given text, whether the associated stock is more likely to move up or down in the short run.[2] For each post, we compute $p_A(\text{post})$, the probability assigned to an upward move based on the post alone, and $p_A(\text{post}+\text{reply})$, the corresponding probability after adding a reply. The reply-induced sentiment shift is the difference between the two. Our baseline probe, which we call the instruct probe, is Mistral-7B-Instruct-v0.1, the instruction-tuned model in the matched model family. We also report all main contrasts using Mistral-7B-v0.1 as a second probe and discuss probe choice in detail in the empirical section.

Our first finding is that AI replies are systematically more bullish than human replies to the same posts. Throughout, the AI-human gap compares, for the same post, the sentiment score of the AI-augmented conversation with the average sentiment score of the human-augmented conversation. Under the baseline probe, the AI reply raises conversation-level

[1] McKinsey & Company. (2025). *New front door to the internet: Winning in the age of AI search.* Retrieved from https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/new-front-door-to-the-internet-winning-in-the-age-of-ai-search

[2] The probe is shown a post or post-plus-reply and asked to choose between token "A" (upward move) and token "B" (downward move). We extract the model's next-token logits for these two options and apply a two-class softmax: $p_A = \exp(\ell_A)/[\exp(\ell_A)+\exp(\ell_B)]$. If the A logit exceeds the B logit by 2, for example, the implied upward probability is $\exp(2)/(1+\exp(2)) \approx 0.88$. See Chen et al. (2025) for details.

sentiment by approximately +0.03 on average relative to the post alone, while the average human reply lowers it by approximately −0.23. The resulting AI-human gap is about +0.26 probability units and is highly significant across all generation temperatures. This difference is economically large: the same investment thesis is read as substantially more bullish when paired with an AI-generated opinion than when paired with the human opinions observed in the original thread.

This gap is not simply an agreement effect. If the AI assistant merely agreed more strongly with users than human commenters did, it should move upward on bullish posts and downward on bearish posts. Instead, the AI-human gap remains positive even for posts classified as bearish. This pattern motivates our interpretation of the result as a bullish bias rather than merely stronger agreement with the investor's thesis.

Our second finding is that prompt framing narrows, but does not eliminate, this gap. When the model is prompted to respond as a professional investor, the AI-human bullish gap falls by roughly 14 percent under the baseline probe. About 86 percent of the gap remains. We refer to the reducible portion as the prompt-sensitive component and the remaining portion as the prompt-insensitive component of the AI-human gap.

Our third finding concerns signal content. We ask whether the excess bullishness documented above simply adds uninformative optimism, or whether AI-augmented signals retain associations with subsequent returns. This exercise is descriptive: Reddit posts are often contemporaneous with news, AI replies are generated ex post, and the tests are not a trading-strategy evaluation. Under stock and date fixed effects, the instruction-tuned AI signals are positively associated with near-horizon returns in the main specifications. The association is strongest over one- to five-day horizons and fades by ten days. However, it is not uniform across subsamples: it is weaker before 2020, for small-cap stocks, and for posts matched to next-day-or-later returns. Thus the bullish bias does not appear to erase return-relevant variation in the underlying thesis, but the evidence should not be read as causal or implementable return predictability.

What explains the AI-human bullish gap? One possibility is that the human replies on WSB are unusually skeptical. Another is that the AI model's pretraining data may embed bullish language, optimistic market narratives, or other domain-specific priors. A third possibility is that the gap is related to the way a base language model is converted into a user-facing conversational assistant. A base language model is trained to generate statistically plausible text continuations. It is not, by itself, trained to be helpful, responsive, or instruction-following. Instruction tuning is a post-training process that uses human feedback, demonstrations, and related optimization methods to make the model behave more like a cooperative conversational assistant (Christiano et al., 2017; Ouyang et al., 2022). In a financial setting, this process may affect the model's learned default orientation, the weight

it places on the user's thesis, and the decisiveness with which it expresses the resulting view.

To assess the role of this post-trained conversational layer, we generate parallel replies from Mistral-7B-v0.1, the matched base model. The base and instruction-tuned models share the same architecture and pretraining corpus, but the base model lacks the instruction-following post-training and associated conversational format (Jiang et al., 2023).[3] We refer to this added component as the **post-trained conversational layer**. Our fourth finding is that the instruction-tuned model is consistently more bullish than the matched base model. Under the base probe, the base model's AI-human gap is small on average and becomes statistically indistinguishable from zero at the highest generation temperature, whereas the instruction-tuned model's gap remains large. Under the instruct probe, both generated models exceed the human benchmark, but the instruction-tuned model still produces the larger gap. Thus the most robust conclusion is not that the base model is uniformly human-like, but that the post-trained conversational layer amplifies bullishness relative to the matched base model.

The base model also gives us a second sentiment probe. Because the base and instruct probes have different distributional properties, we treat raw levels as probe-specific and use sign agreement as a conservative robustness check. We therefore repeat the main comparison using both the instruction-tuned probe and the base probe, and classify a shift as robust only when both probes agree on its sign. This conservative classification gives the same conclusion. Both probes agree that the AI reply shifts sentiment upward in 38.9 percent of posts, compared with 17.3 percent for human replies. Conversely, both probes agree on a downward shift in only 7.6 percent of AI cases, compared with 25.2 percent of human cases.

We interpret these findings using a reduced-form reply model. The model separates three forces. The first is a learned default orientation, denoted p_0, which captures the model's baseline tendency in the financial domain. The second is responsiveness, τ, which captures how much weight the assistant places on the user's thesis relative to that default. The third is categorical expression, γ, which captures the tendency to express the resulting assessment decisively rather than hedge around the midpoint. This distinction is important because responsiveness alone is an agreement mechanism: it pulls the reply toward the user's thesis or toward the model's default. A systematic bullish bias, including for bearish posts, requires a sufficiently bullish learned default, with categorical expression amplifying the resulting assessment. The model is not a structural account of language-model training; it is a simple framework for organizing the empirical patterns in finance terms.

[3]Prompting and post-training operate at different levels. A prompt changes the input to a fixed model, for example by asking it to respond as a professional investor. Post-training changes the model's weights: the numerical parameters learned during training that map input words into probabilities over next words. This distinction matters because our prompt experiment shows that role framing can reduce the AI-human gap, but cannot remove most of it; if the post-trained conversational layer changes the model's response rule, some of the gap may remain after prompt-level interventions.

The paper connects to the economics literature on biased communication, advice, and audience-facing intermediation. In classic models, senders slant reports toward the receiver's prior or preferences because doing so increases perceived accuracy, demand, or career value (Prendergast, 1993; Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006). The mechanism here is not strategic in the same sense: a conversational AI assistant need not have preferences over investor beliefs. But a system deployed as a helpful, responsive, and clear assistant may nevertheless behave as if it combines a domain-specific default with extra weight on the user's framing and more categorical expression. Our reduced-form model captures this through a learned default orientation, responsiveness to the user's message, and pressure toward categorical expression, linking the paper to work on coarse communication and limited attention (Sims, 2003; Mackowiak and Wiederholt, 2009).

Relatedly, Bini et al. (2026) study whether large language models exhibit systematic behavioral biases in economic and financial decision tasks. They document that LLM behavior varies across model families, model scale, and task type, and that prompting can partly reduce some biases. Our paper studies a different but complementary margin: whether a user-facing assistant changes the directional tone of financial communication when replying to investor theses, and whether that behavior is associated with the post-trained conversational layer.

The paper also contributes to the growing literature on large language models in finance. Liu et al. (2025) study disagreement across frontier models using anonymized earnings-call transcripts, emphasizing differences in embedded information sets. We study a different margin: how a user-facing AI opinion changes the sentiment of an investor conversation relative to human opinions responding to the same investment thesis. Bhagwat et al. (2025) use persona prompts to elicit heterogeneous investor narratives from language models. We also vary prompt framing, but use it to test whether role conditioning moderates the AI-human reply gap.

Our design is complementary to work on AI content in investor discourse. Hirshleifer et al. (2025) study AI adoption in settings including Seeking Alpha and `r/wallstreetbets` using AI-content detection and observational variation. We instead generate replies to the same parent post and compare them with the thread's actual human replies. This gives a direct benchmark for the sentiment movement produced by organic investor discussion. More broadly, the paper relates to work comparing AI and human financial communication in earnings calls and investor information processing (Bai et al., 2025; Blankespoor et al., 2025), as well as to conversational measurement methods that exploit thread structure (Balasubramaniam and Sabat, 2025).

A related literature studies temporal consistency and lookahead bias in language models. He et al. (2025) develop methods to ensure chronological consistency in historical analysis,

while Engelberg et al. (2025) remove firm identifiers to limit memorized outcome information. Our object is different. We study how a user-facing assistant responds to an investor's framing, not whether it recalls firm-specific outcomes. The results therefore speak to conversational behavior rather than data leakage or training-set contamination.

The paper proceeds as follows. Section 2 discusses investor-facing AI as a form of conversational intermediation and why its opinions may differ from human discussion. Section 3 presents the main empirical evidence on the AI-human bullish gap, prompt framing, robustness, and the base-model diagnostic. Section 3.6 develops the reduced-form theoretical interpretation. Section 4 examines whether AI-augmented signals retain return-relevant content in descriptive return regressions. Section 5 concludes.

2 Base Models, Instruction Tuning, and User-Facing Financial Replies

This section provides the technical background needed to interpret our empirical design. The distinction that matters for the paper is between a *base* language model and an *instruction-tuned* model. Both are large language models, but they are trained and used in different ways. A base model is trained primarily to predict the next token in a sequence of text. An instruction-tuned model starts from a base model and is then further trained to follow user instructions and produce useful assistant-style responses. Our empirical design exploits this distinction by comparing Mistral-7B-v0.1, the base model, with Mistral-7B-Instruct-v0.1, its matched instruction-tuned counterpart (Jiang et al., 2023).

2.1 Base Models

A base language model is trained on a large corpus of text to predict plausible continuations. Given an input sequence, the model assigns probabilities to possible next tokens and generates text by repeatedly sampling or selecting from those probabilities. In this sense, the model is not trained primarily to answer a user's question, follow an instruction, or behave as a financial advisor. It is trained to continue text in a way that is statistically consistent with the patterns learned during pretraining.

For example, if a base model is given a stock-related statement, it may continue the discussion in the style of financial text it has seen before. But the base model does not, by construction, treat the input as a request from a user who expects an organized answer. It has no direct objective to be helpful, balanced, cautious, agreeable, or actionable. Its behavior reflects the pretraining corpus and the next-token prediction objective.

The model's behavior is governed by its *weights*: the numerical parameters learned during

training. These weights determine how the model maps an input sequence into probabilities over possible next tokens. Once the weights are fixed, changing the prompt changes the input supplied to the model, but not the model itself. This distinction is important below because our prompt experiment changes only the input given to the model, whereas instruction tuning changes the model weights.

2.2 Instruction Tuning and User-Facing Assistants

The models used in consumer-facing AI assistants are typically not raw base models. They are further post-trained so that they respond to user requests in a helpful, organized, and conversational manner. This post-training stage is commonly referred to as *instruction tuning*. In broad terms, instruction tuning exposes the model to examples of prompts and desirable responses, so that the model learns to treat an input as a user request rather than merely as text to be continued (Wei et al., 2022; Ouyang et al., 2022).

A concrete example from the instruction-following literature illustrates the distinction. In Ouyang et al. (2022), a base GPT-3 model is given a user instruction asking it to explain the moon landing to a six-year-old, but its response behaves more like an unconstrained text continuation. The instruction-tuned model instead gives a short answer tailored to the user's request. The relevant distinction is not the topic itself, but the change in behavior: post-training teaches the model to treat the input as an instruction from a user and to produce a response that looks like an assistant's answer.

Many user-facing assistants are also trained using reinforcement learning from human feedback (RLHF). In RLHF, human annotators compare alternative responses to the same prompt. These rankings are used to train a reward model, and the language model is then updated so that it assigns higher probability to responses that receive higher predicted reward (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). The result is a model that is more likely to produce responses judged by human evaluators to be helpful, appropriate, and responsive to the user's request.

This process can also introduce systematic conversational tendencies. If human evaluators tend to prefer responses that are supportive, agreeable, or aligned with the user's stated view, the model may learn to validate the user's framing rather than provide a more independent assessment. This tendency is often described in the computer science literature as *sycophancy* (Sharma et al., 2023). Training assistants to be helpful and acceptable to users may also favor polite, constructive, and non-confrontational replies (Bai et al., 2022). These properties are useful in many settings, but in financial applications they may affect the directional tone of the response.

2.3 Prompting versus Post-Training

Prompting and post-training operate at different levels. A prompt changes the text given to a fixed model. For example, asking the model to respond as a professional investor changes the input context and may lead the model to produce a more cautious or analytical answer. But the prompt does not change the model's weights. The same underlying model is being used; only the words supplied to it have changed.

Post-training changes the model itself. Instruction tuning and RLHF update the model's weights so that, across many possible prompts, it is more likely to produce responses that resemble desired assistant answers. Thus, role prompting can partly moderate behavior, but it need not undo the behavioral tendencies embedded during post-training. This distinction motivates our prompt experiment. If professional-investor prompting narrows the AI-human gap but leaves most of it intact, this suggests that part of the gap may be tied to the post-trained model rather than to surface-level prompt wording alone.

2.4 The Mistral Base–Instruct Comparison

Our empirical design uses the Mistral model family because it provides a matched base and instruction-tuned pair. Mistral-7B-v0.1 is the base model. Mistral-7B-Instruct-v0.1 is the instruction-tuned version of the same model family (Jiang et al., 2023). The two models share the same architecture and pretraining corpus. The instruct variant adds the post-training and interaction format that make it usable as a conversational assistant.

We refer to this added component as the **post-trained conversational layer**. This layer includes instruction tuning and the associated chat-template format used when interacting with the instruct model. In practice, the instruct model is not used by simply passing raw text into the model. The prompt is wrapped in a chat template that marks the user's message and asks the model to respond in assistant form. The base model, by contrast, is used as a raw continuation model. Appendix A.1 describes the prompt formatting used in our experiments.

The base–instruct comparison helps distinguish two explanations for the AI-human bullish gap. One possibility is that the gap reflects the pretraining corpus or synthetic text generation more generally. If so, the base model should also produce replies that are similarly more bullish than human replies. A second possibility is that the gap is associated with the post-trained conversational layer. If so, the instruction-tuned model should differ from human replies more strongly than the base model does. The empirical results support the second interpretation: the base model produces less bullish replies than the instruction-tuned model and is much closer to the human benchmark.

We do not claim that this comparison separately identifies supervised fine-tuning, RLHF,

chat-template formatting, or other elements of post-training. The comparison is best interpreted as evidence on the combined behavioral role of the post-trained conversational layer. This is the economically relevant object for our setting because investors encounter user-facing assistants in their post-trained conversational form, not as raw base models.

3 Empirical Design and Main Evidence

3.1 Empirical Setting and Sample Construction

We study whether user-facing AI replies differ systematically from human replies when responding to the same investment thesis. The empirical design requires two objects: investor-authored posts that present views about individual stocks, and human replies to those same posts. Reddit provides such a setting.

We draw on `r/wallstreetbets` (WSB), a prominent online forum where retail investors discuss individual equities and trading strategies. We use the full archive of WSB posts and comments from 2013 through 2022, focusing on posts tagged as "Due Diligence" (DD). DD posts contain structured investment theses with relatively clear directional views, making them well suited for measuring whether a reply increases or decreases the bullishness of the surrounding conversation (Bradley et al., 2024). The value of this setting is the pairing of real investor-authored stock theses with human top-level replies to those same theses.

To construct the sample, we extract ticker symbols from these threads and match them to the Center for Research in Security Prices (CRSP) database.[4] We identify 2,495 unique tickers that are the subjects of explicit investment arguments on the platform. We match alphanumeric tokens to CRSP tickers after removing tokens that coincide with common English words using the Zipf frequency scale,[5] and we allow additional matches based on distinctive tokens from company names. We retain only threads that map to a single public company to avoid ambiguity.

Reddit organizes comments by depth: depth-1 comments are direct replies to the original post, and higher-depth comments are replies to those comments. We restrict attention to depth-1 replies throughout, because our AI-generated replies are also produced in response to the original post. Both AI and human replies therefore address the same investment thesis, making the comparison well defined.

Sample structure. Our analysis uses three related samples. The **post universe** consists

[4]The CRSP database provides historical identifiers, ticker symbols, and return data for publicly traded firms and is widely used in empirical asset pricing research.

[5]The Zipf frequency scale measures the logarithmic frequency of words in natural language corpora. Following common practice in natural language processing, we treat tokens with a Zipf frequency above 4 as common English words and exclude them from ticker matching. See van Heuven et al. (2014) for details.

of 28,922 DD posts that pass quality filters and map uniquely to a CRSP-listed firm. The **generation universe** consists of all AI-generated replies to these posts under two prompt conditions, three temperature settings, and two replications, yielding approximately 173,500 generated replies per generator-probe cell. The **AI-versus-human comparison sample** consists of the 15,206 posts in the post universe that also have at least one depth-1 human comment. All AI-versus-human comparisons use this sample.

3.2 Experimental Design

For each post in the generation universe, we generate a reply from Mistral-7B-Instruct-v0.1, a conversational AI model used here as the user-facing AI assistant. In the first part of the analysis we refer to this simply as the AI model. Section 3.4 introduces its matched base model, Mistral-7B-v0.1, to assess the role of post-training and to provide a second sentiment probe for robustness.

AI response generation. For each of the 28,922 posts in the post universe, we generate AI replies under two prompt conditions. Under the **neutral prompt**, the model is given no additional role framing. Under the **professional investor prompt**, the generation instruction is prefaced with the sentence "You are a professional investor." We use the professional-investor condition to test whether role conditioning narrows the AI-human bullish gap. The full text of both prompts appears in Appendix A.1. We generate responses at temperatures 0.0, 0.5, and 1.0, with two independent replications at each temperature, yielding six AI responses per post per prompt condition. Responses are capped at 120 tokens.

Unit of observation. In all AI-versus-human shift comparisons, we first average the two replications within each post-by-temperature-by-prompt cell and then compare the resulting cell-level AI shift to the human benchmark. The observation count of 15,206 reflects this averaging.

Measuring conversation-level sentiment. We measure the bullishness or bearishness of text using the probabilistic probing method of Chen et al. (2025). A language model (the probe) is presented with a forced-choice question: based on the text, is an upward or downward stock move more likely in the short run? Rather than sampling a response, we extract the probe's raw next-token logits for the tokens "A" (upward) and "B" (downward) and apply a two-class softmax. For any text x:

$$p_A(x) \equiv \Pr_{\text{probe}}(A \mid x) = \frac{\exp\big(\ell_A(x)\big)}{\exp\big(\ell_A(x)\big) + \exp\big(\ell_B(x)\big)}, \tag{1}$$

where $\ell_A(x)$ and $\ell_B(x)$ are the logits for tokens "A" and "B" respectively at the first output position. Probing is deterministic: logits are read directly without sampling, so $p_A(x) \in (0, 1)$

is a fixed scalar for any given text and probe. In the main analysis we use Mistral-7B-Instruct-v0.1 as the probe model (instruct probe). Section 3.4 introduces the matched base model Mistral-7B-v0.1 as a second probe (base probe), which is used extensively in the robustness analysis of Section 3.5.

Reply-induced sentiment shift. For post i, the probe first assigns an upward-move probability to the post text alone, $p_A(\text{Post}_i)$. It then assigns an upward-move probability to the post plus a response j:

$$p_A(\text{Post}_i + \text{Response}_{ij}) \;=\; \Pr_{\text{probe}}\big(A \mid \text{Post}_i,\, \text{Response}_{ij}\big), \tag{2}$$

where the post and response are concatenated as a single input with a separator. The **joint shift** for response j to post i is then:

$$\delta_{ij}^{\text{joint}} \;=\; p_A(\text{Post}_i + \text{Response}_{ij}) \;-\; p_A(\text{Post}_i). \tag{3}$$

A positive $\delta_{ij}^{\text{joint}}$ means the response makes the conversation more bullish than the post alone; a negative value means it makes the conversation more bearish. This is the primary measure used throughout the paper.

We also construct a **response-relative measure** that presents the response text in isolation to the probe:

$$\delta_{ij}^{\text{resp}} \;=\; p_A(\text{Response}_{ij}\text{ alone}) \;-\; p_A(\text{Post}_i). \tag{4}$$

This measure does not ask how the reply changes the joint post-plus-reply reading. Instead, it asks whether the reply text itself is more or less bullish than the post it answers. The two measures have a post-level correlation of approximately 0.93 under the base probe and 0.91 under the instruct probe, and agree on sign in 94% and 89% of cases respectively, suggesting that the main results are not driven mechanically by how the post and reply interact inside the joint probing prompt.

Human benchmark. For post i with K_i depth-1 human comments, we average the joint post-plus-reply probe scores equally across all comments:

$$\bar{p}_A(\text{Post}_i + \text{Human}) \;=\; \frac{1}{K_i}\sum_{k=1}^{K_i} p_A(\text{Post}_i + \text{Human}_{ik}). \tag{5}$$

The human discussion benchmark shift is then:

$$\delta_i^{H} \;=\; \bar{p}_A(\text{Post}_i + \text{Human}) \;-\; p_A(\text{Post}_i). \tag{6}$$

Averaging occurs before the shift is computed, so δ_i^H represents the average reply-induced shift in joint sentiment from the observed depth-1 human replies to the post. Posts with no depth-1 comments are excluded from all AI-versus-human comparisons. An analogous average is computed for the response-relative human benchmark using $p_A(\text{Human}_{ik}$ alone) in place of the joint score.

AI shift and replication averaging. For the AI model, we generate $R = 2$ replications within each post-by-temperature-by-prompt cell. For a given generation temperature T and prompt condition c, the AI shift is:

$$\delta_{i,T,c}^{\text{AI}} \;=\; \frac{1}{R}\sum_{q=1}^{R} p_A(\text{Post}_i + \text{AI}_{iTcq}) \;-\; p_A(\text{Post}_i). \tag{7}$$

Temperature-specific tables retain this post-by-temperature-by-prompt unit after averaging the two replications. When we report pooled specifications or decomposition results across temperatures, we additionally average the three temperature cells within each post and prompt condition.

When the prompt condition and temperature are clear from context, or when we average over temperatures, we write the resulting AI shift simply as δ_i^{AI}.

3.3 The AI-Human Bullish Gap

Our central outcome is the AI-human bullish gap. Because both the AI shift δ_i^{AI} and the human shift δ_i^H subtract the same baseline $p_A(\text{Post}_i)$, the post-alone score cancels mechanically in the difference:

$$\begin{aligned}
\Delta_i \;&=\; \delta_i^{\text{AI}} \;-\; \delta_i^H \\
&=\; \big[p_A(\text{Post}_i + \text{AI}) - p_A(\text{Post}_i)\big] - \big[\bar{p}_A(\text{Post}_i + \text{Human}) - p_A(\text{Post}_i)\big] \\
&=\; p_A(\text{Post}_i + \text{AI}) \;-\; \bar{p}_A(\text{Post}_i + \text{Human}).
\end{aligned} \tag{8}$$

The gap therefore has a simple interpretation: for the same post and probe, does the probe read the AI-augmented conversation as more bullish than the human-augmented conversation? A positive Δ_i means yes; a negative value means the human replies produce a more bullish joint reading. The post baseline cancels mechanically in the difference, so Δ_i compares the AI-augmented and human-augmented readings for the same underlying post rather than comparing raw shifts across different posts.

An analogous gap is defined for the response-relative measure by replacing the joint scores in equation (8) with the response-alone scores. We report results under both measures throughout.

Appendix Table 1 summarizes distributional properties of the probed sentiment scores. Under the instruct probe, post-level sentiment has a mean of 0.63 and a standard deviation of 0.35, with 33.5 percent of posts above 0.9. Human top-level replies receive an average joint score of 0.40, substantially below the post-level mean of 0.63. Under this probe, human replies reduce the measured bullishness of the posts they respond to on average.

Appendix Table 2 reports the main AI-human comparison for the instruction-tuned model. Under the neutral prompt, the AI model produces average joint sentiment shifts of approximately +0.030 to +0.031, depending on generation temperature. The human benchmark shift is −0.234. The resulting AI-human bullish gap is approximately +0.264 to +0.265 probability units, with paired t-statistics between 118 and 130. The magnitude is large: the same investment thesis is read as substantially more bullish when paired with an AI-generated reply than when paired with the human replies observed in the original thread.

Appendix Table 3 examines whether this gap can be reduced through prompt framing. When the AI model is prompted to respond as a professional investor, the AI-human gap falls from approximately +0.264–+0.265 to +0.223–+0.232, a reduction of roughly 14 percent under the instruct probe. The differences are statistically significant at the one-percent level. We refer to the reducible portion as the **prompt-sensitive component** and the residual as the **prompt-insensitive component**. Appendix Table 4 formalizes this decomposition: about 86 percent of the AI-human gap remains under professional-investor prompting.

The robustness of the bullish gap across sixteen subsamples, two probe models, two prompt conditions, three generation temperatures, and two shift measures is examined in detail in Section 3.5. The headline finding is that the instruction-tuned AI-human bullish gap is positive and statistically significant across all robustness specifications, including posts in which the investor's original thesis is bearish. This rules out the possibility that the gap merely reflects agreement with a predominantly bullish sample.

Figure 1 illustrates these results. Panel A shows that AI-generated replies generally lie above the human benchmark across the post-sentiment distribution. Panel B shows that the AI-human gap Δ_i is positive across the full range of post sentiment, including bearish posts, and is reduced but not eliminated by professional-investor prompting.

These results establish the central descriptive fact: the user-facing AI model is more bullish than human replies to the same investment theses. Three explanations are possible. First, human replies on WSB may be unusually skeptical, mechanically making AI replies look bullish by comparison rather than reflecting any excess in the AI's own output. Second, any language model trained on internet text may generate bullish financial content because bullish investment language is more prevalent in the pretraining corpus. Third, the gap may arise specifically from the post-trained conversational layer that turns a base language model into a user-facing assistant optimized for helpful, assistant-style conversational

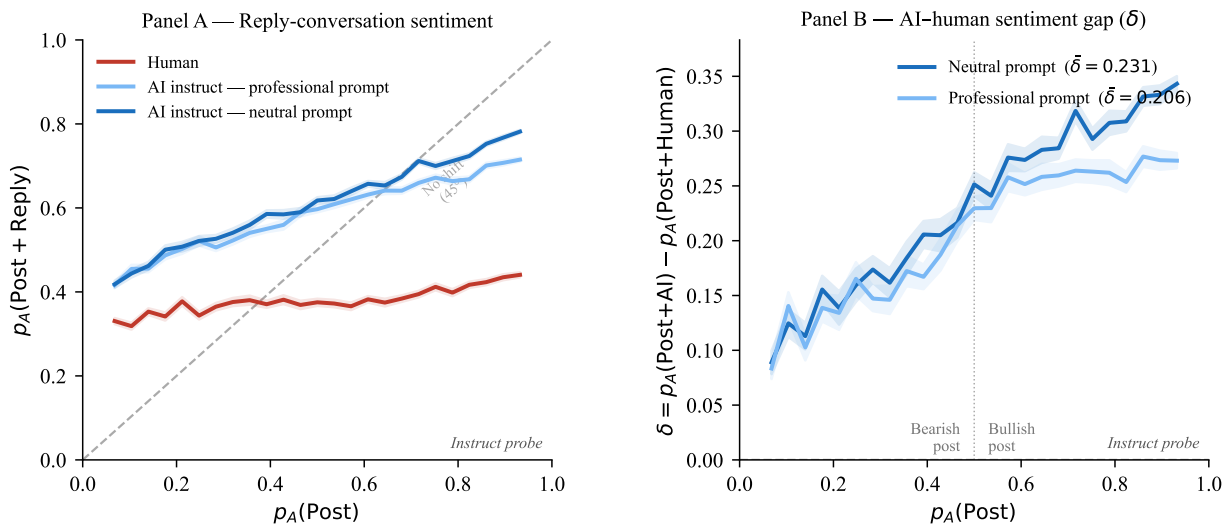

Figure 1: AI-human bullish gap: instruction-tuned model versus human replies. Panel A plots mean p_A(Post+Reply) against p_A(Post) in 25 equal-width bins; Panel B plots the gap $\Delta = p_A(\text{Post+AI}) - \bar{p}_A(\text{Post+Human})$. Shaded bands are ±1 SE. Instruct probe; $N = 15{,}206$ posts.

responses. The next subsection introduces the matched base model to distinguish among these possibilities.

3.4 The Role of Post-Training

To assess the role of the post-trained conversational layer, we generate parallel replies from Mistral-7B-v0.1, the matched base model. The base and instruction-tuned models share the same architecture and pretraining corpus (Jiang et al., 2023). The base model lacks the instruction-following post-training and the chat-template format used by the AI assistant, and generates text continuations from the pretraining distribution rather than responses optimized for conversational helpfulness. We refer to the added component in the user-facing model as the **post-trained conversational layer**.

Base-model diagnostic. Appendix Table 6 reports the full AI-versus-human comparison including the matched base model. Under the base probe and neutral prompt, the base model produces sentiment shifts of +0.067 to +0.099 depending on temperature. At the highest generation temperature, its gap with the human benchmark is only +0.002 and is statistically indistinguishable from zero ($t = 1.08$), while the instruction-tuned model's gap is +0.124 in the same specification. Averaged across temperatures, the base model's gap under the base probe is +0.022, compared with +0.108 for the instruction-tuned model. The instruction-tuned model therefore exceeds the human benchmark by substantially more

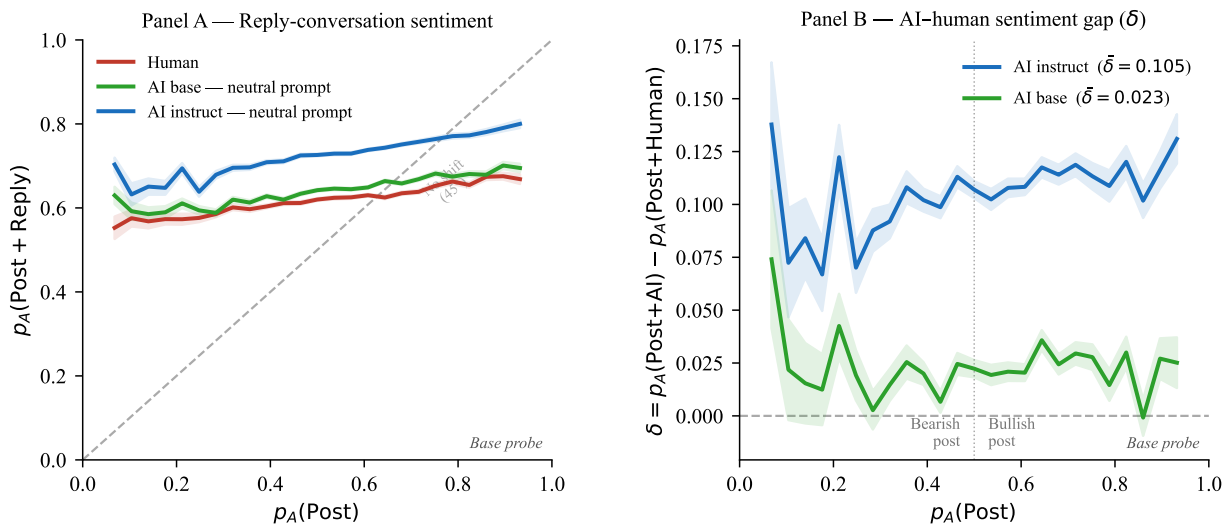

Figure 2: Bullish gap: instruction-tuned model, base model, and human replies (base probe, neutral prompt). Panel A plots mean p_A(Post+Reply) against p_A(Post); Panel B plots the AI-human gap Δ for each generator. $N = 15{,}206$ posts.

than the base model does, and this ordering is robust across all 16 subsamples examined (Section 3.5).

This comparison qualifies the three explanations offered at the end of Section 3.3. Under the base probe, the matched base model's gap is small on average and statistically indistinguishable from zero at the highest generation temperature. This makes it difficult to attribute the entire AI-human gap to unusually skeptical human replies. At the same time, the instruct-probe results show that the base model can also exceed the human benchmark substantially, so the evidence should not be read as showing that generated text is uniformly human-like. The more robust conclusion is that the instruction-tuned model is consistently more bullish than the matched base model across all specifications. Because the two models share architecture and pretraining corpus, this ordering is consistent with the post-trained conversational layer amplifying bullishness beyond what is already present in the matched base model.

Figure 2 shows the three-way comparison under the base probe, where the separation between the base and instruction-tuned generators is most visible. Panel A shows that the instruction-tuned model generally lies above the human benchmark, with the base model typically between the two under the base probe. Panel B shows that the instruction-tuned model's AI-human gap is larger throughout the binned post-sentiment distribution. The figure illustrates the amplification effect of the post-trained conversational layer rather than the stronger claim that the base model is always human-like.

Second probe and robust sign classification. We also use the matched base model as a second

sentiment probe. Appendix Table 5 documents that the two probes have materially different distributional properties: the instruct probe places more mass near 0 and 1, while the base probe is more diffuse. We therefore treat raw sentiment levels as probe-specific and focus on within-probe comparisons throughout. Appendix Table 7 reports a conservative sign-agreement classification requiring both probes to agree on the direction of each shift. Under the neutral prompt at temperature 1.0, both probes agree that the user-facing AI reply shifts sentiment upward in 38.9 percent of posts, compared with 17.3 percent for human replies. Conversely, both probes agree on a downward shift in only 7.6 percent of AI cases, compared with 25.2 percent of human cases.

Prompt asymmetry. Appendix Table 8 reports the full prompt-framing results including the matched base model. Professional-investor prompting reduces the AI-human gap for the user-facing model but has no comparable effect on the base model — if anything, the professional prompt slightly widens the base model's gap. Under the base probe, the base model's gap increases by about 0.005–0.006 across temperatures under the professional prompt, whereas the instruction-tuned model's gap falls by about 0.029 under the base probe. This widening, though small in magnitude, is statistically significant ($t = -2.50$ to -3.14), and the same pattern holds under the instruct probe. The difference in prompt response between the instruction-tuned and base generators is statistically significant under both probes. Appendix Table 11 formalizes this pattern: for the base model, the total gap is small and the professional prompt does not reduce it. Notably, the professional prompt's failure to reduce the base model's gap — combined with its substantial reduction of the instruct model's gap — is consistent with the prompt-sensitive component of the bullish bias being specific to the post-trained conversational layer, rather than a generic property of the professional-investor framing itself. The matched base model does not respond to role conditioning in the same way or with the same sign as the conversational assistant.

Regression decomposition. Appendix Table 9 presents a within-post regression that estimates how AI-generated sentiment shifts vary with generator type, prompt condition, temperature, and their interactions:

$$\begin{aligned}\delta_{i,g,T,c} \;=\; & \alpha_i + \beta_1\,\mathbf{1}[\text{instruct}]_g + \beta_2\,\mathbf{1}[\text{professional}]_c + \beta_3\,(T-\bar{T}) \\ & + \beta_4\,\mathbf{1}[\text{instruct}]_g \cdot \mathbf{1}[\text{professional}]_c + \beta_5\,\mathbf{1}[\text{instruct}]_g \cdot (T-\bar{T}) \\ & + \beta_6\,\mathbf{1}[\text{professional}]_c \cdot (T-\bar{T}) + \varepsilon_{i,g,T,c}. \end{aligned} \tag{9}$$

Post fixed effects α_i absorb all post-level heterogeneity, and standard errors are clustered at the post level. Under the base probe, the instruction-tuned generator coefficient is +0.086 ($t = 75.4$), indicating upward shifts about 8.6 percentage points larger than the matched base model after absorbing post-level heterogeneity. The interaction between the instruction-

tuned generator and the professional-investor prompt is -0.034 ($t = -24.1$) under the base probe and -0.049 ($t = -28.0$) under the instruct probe, consistent with prompt framing narrowing the user-facing model's bullish shift relative to the base model.

3.5 Robustness: Bullish Bias versus Agreement

A potential concern is that the bullish gap documented in Sections 3.3 and 3.4 reflects specific features of the sample rather than a systematic property of the instruction-tuned model. A separate conceptual concern is whether the gap reflects a genuine *bullish bias* — a systematic upward pull in the AI's output — or merely stronger *agreement* with the investor's post direction. These two possibilities have distinct empirical implications. If the AI model simply agreed more strongly with the post than human commenters did, it should be more bullish on bullish posts and more bearish on bearish posts. A bullish bias, by contrast, predicts a positive AI-human gap even on bearish posts: the AI model should be systematically less bearish than human replies even when responding to an investor who is pessimistic about a stock.

We subject the bullish bias to a systematic robustness analysis across sixteen subsamples, two probe models, two prompt conditions, three generation temperatures, and two shift measures. We define bullish and bearish posts using the base-probe post score, so that the post-direction split is fixed across the two probe specifications.

Robustness criterion. For a given subsample and specification, we call a result **robust** if the AI-human gap is positive and significant at the ten-percent level under both the base probe and the instruct probe independently.

Joint shift measure. Appendix Table 12 reports results using the joint shift measure. The instruction-tuned AI-human bullish gap is robust in all sixteen of sixteen subsamples for all posts, for bullish posts, and for bearish posts simultaneously. The gap is positive and significant under both probes across subsamples defined by post inner confidence (low, mid, and high terciles), number of human comments (1, 2–5, 6–20, and more than 20), calendar year (pre-2020, 2020, 2021, 2022), firm size (small and large cap by median log market capitalization), and return timing (same-day and next-day-plus posts). No subsample produces a reversal.

The bearish-post result directly addresses the bias-versus-agreement question. Under the neutral prompt, the AI-human gap for bearish posts is $+0.100$ under the base probe ($t = 45.93$) and $+0.133$ under the instruct probe ($t = 44.28$), both significant at the one-percent level. The gap is approximately 85 to 90 percent as large for bearish posts as for bullish posts. This near-symmetry across post directions indicates that the AI-human gap is not merely the consequence of the AI model amplifying already-bullish posts: it operates

as a broad upward pull that shifts sentiment in the bullish direction regardless of whether the investor's original thesis is optimistic or pessimistic. This pattern is inconsistent with a pure agreement mechanism and supports the bullish bias interpretation.

Under the professional-investor prompt, the gap is reduced by approximately 27 percent under the base probe and 14 percent under the instruct probe. The reduced gap remains robust in all sixteen subsamples for all posts and bullish posts, and in fifteen of sixteen for bearish posts. Prompt framing therefore narrows but does not eliminate the bullish bias.

Response-relative measure. Appendix Table 14 replicates the analysis using the response-relative measure $\delta^{\text{resp}} = p_A(\text{reply alone}) - p_A(\text{post})$, which presents the reply text in isolation to the probe. The results are identical in structure: the AI-human gap is robust in all sixteen of sixteen subsamples for all posts, bullish posts, and bearish posts. The two measures agree on the verdict for every subsample and every post-direction group. The bullish bias is therefore visible in the reply text itself, not only in the combined post-plus-reply reading.

Temperature robustness. Appendix Table 16 reports the bullish gap at each generation temperature, including temperature 0.0, which corresponds to fully deterministic greedy decoding with no sampling variation. At temperature 0.0, the gap is +0.096 under the base probe ($t = 57.90$) and +0.264 under the instruct probe ($t = 117.70$) for all posts, and remains significant and positive for bearish posts under both probes. The bullish bias is therefore present even in the model's deterministic modal reply, rather than arising from random sampling noise in generation.

Taken together, the robustness evidence supports a strong form of the bullish bias claim for the instruction-tuned model: it produces a systematic upward shift in conversation-level sentiment relative to human replies, and this shift is not explained by the composition of the sample, the choice of probe model, the generation prompt, the amount of stochastic variation in generation, or the measurement approach used to quantify reply-induced sentiment change.

Validation with an independent model family. To assess whether the bullish bias extends beyond the Mistral model family, we replicate the core analysis using GPT-4o-mini (OpenAI) as an alternative instruction-tuned generator on a validation sample of 2,000 posts drawn from the same post universe. Appendix Table 17 reports the results. The AI-human bullish gap for GPT-4o-mini is +0.419 ($t = 42.34$), and it is positive and significant for both bullish posts ($+0.563^{***}$) and bearish posts ($+0.172^{***}$), directly replicating the pattern that distinguishes a bullish bias from mere agreement with the investor's original thesis. The gap is also present at temperature 0.0 ($+0.421^{***}$), confirming that the bias appears in the model's deterministic modal reply. These results suggest that the bullish bias is not specific to the Mistral instruction-tuning pipeline. We note that GPT-4o-mini serves as both generator

and probe in this validation, so the gap magnitude is not directly comparable to the main results; the structural pattern, however, is identical.

Linguistic properties of AI-generated replies. Appendix Tables 18 and 19 document two further properties of the generated replies that are relevant for the theoretical interpretation of the bullish bias. The instruction-tuned model uses significantly more categorical and decisive language than the base model — words such as *clearly, definitely, certainly, conviction* — at a rate of 0.30 per 100 words under the neutral prompt compared with 0.10 for the base model, a difference of +0.197 per 100 words ($t = 70.55$). This pattern holds under the professional prompt (+0.170, $t = 69.97$) and is consistent with the instruction-tuned model adopting a more direct and categorical conversational register than the matched base model. Separately, the instruction-tuned model produces significantly shorter replies than the base model: 66.6 words on average under the neutral prompt compared with 88.8 words for the base model, a difference of 22.2 words ($t = -214.40$). The professional prompt partially closes this gap, with instruct replies averaging 75.2 words compared with 88.2 for the base model under professional framing. Together, these properties — more decisive language and more concise replies — characterize the conversational register of the instruction-tuned model and provide empirical grounding for theoretical interpretations based on a helpful, direct conversational register. The remaining empirical question is whether the AI-augmented signals retain return-relevant variation despite this bullish bias, which we examine in Section 4.

3.6 Theoretical Interpretation and Discussion

This section provides a reduced-form interpretation of the empirical results. The goal is not to model the internal architecture of a language model. Instead, the model isolates three forces that can generate the patterns documented above: a learned bullish default, responsiveness to the user's thesis, and categorical expression. The central distinction is between *agreement* and *bullish bias*. Responsiveness alone makes the assistant's reply move toward the user's thesis or toward the assistant's learned default. By itself, this is an agreement or anchoring force. A systematic bullish bias requires something more: a sufficiently bullish learned default, denoted p_0, together with a tendency to express the resulting assessment in categorical language.

This distinction matters for interpreting the empirical results. If the AI-human gap were only an agreement mechanism (as is often described as *syophancy bias* in the computer science literature, (Sharma et al., 2023)), the assistant should reinforce bullish posts and move downward with bearish posts. The robustness results show a stronger pattern: the AI-human gap remains positive even for posts classified as bearish. The model below therefore allows the assistant to have a learned default orientation p_0, which may reflect the pretraining

corpus, post-training data, human feedback, or the deployment conventions of an assistant-style model. The assistant combines this default with the user's thesis and then maps the resulting latent assessment into a reply. A high p_0 can pull bearish posts upward, while categorical expression can amplify the resulting sentiment once the latent assessment lies on the bullish side of the midpoint.

Environment There is an unknown binary event $E \in \{0, 1\}$, where $E = 1$ denotes an upward stock move. A user presents an investment thesis with measured bullishness $r \in [0, 1]$. A larger value of r corresponds to a more bullish thesis. The assistant produces a probabilistic reply $m(r) \in [0, 1]$, interpreted as the bullishness of the assistant's response.

The assistant has a learned default orientation

$$p_0 \in (0, 1),$$

which represents the probability of an upward move that the assistant would assign before fully conditioning on the user's specific thesis. This is not the objective probability that the stock will rise. It is a reduced-form representation of the model's learned baseline in this domain. It may come from the pretraining corpus, post-training data, human feedback, or the conventions embedded in user-facing deployment.

The assistant combines this learned default with the user's thesis through

$$\rho(r; \tau, p_0) \equiv (1 - \tau)p_0 + \tau r, \qquad \tau \in [0, 1]. \tag{10}$$

The parameter τ captures user-message responsiveness. When $\tau = 0$, the assistant relies entirely on its learned default. When $\tau = 1$, it fully adopts the directional content of the user's thesis. Intermediate values represent partial responsiveness.

Conditional on the latent assessment $\rho(r; \tau, p_0)$, the assistant chooses the reply m by trading off accuracy against categorical expression:

$$m(r) \in \arg\max_{m \in [0,1]} \left\{ -\mathbb{E}\big[(m - E)^2 \mid r\big] + \gamma\big(m - \tfrac{1}{2}\big)^2 \right\}, \tag{11}$$

where $\gamma \geq 0$. The parameter γ captures the tendency to express views decisively rather than in heavily hedged language. A larger γ rewards replies farther from the midpoint. Thus γ does not create bullishness by itself; it amplifies the direction implied by the latent assessment.

To keep the analysis in the empirically relevant interior region, we maintain the following regularity condition.

Assumption 1 (Interior regularity)**.** *The parameter tuple* (p_0, τ, γ) *and the support of post bullishness values are such that the optimizer lies in* $(0, 1)$ *for all* r *under consideration. In addition,* $\gamma \in [0, 1)$*.*

The next result gives the reply rule. The important point is that the reply depends on three primitives: the learned default p_0, responsiveness to the user's thesis τ, and categorical expression γ.

Proposition 1 (Reply rule)**.** *Maintain Assumption 1. The assistant's unique optimal reply is*

$$m(r) = \frac{\rho(r; \tau, p_0) - \gamma/2}{1 - \gamma} = \frac{(1 - \tau)p_0 + \tau r - \gamma/2}{1 - \gamma}. \tag{12}$$

Moreover,

$$m(r) - \frac{1}{2} = \frac{\rho(r; \tau, p_0) - \frac{1}{2}}{1 - \gamma}. \tag{13}$$

Proposition 1 shows how categorical expression operates. The sign of the reply relative to the midpoint is determined by the latent assessment $\rho(r; \tau, p_0)$. If $\rho(r; \tau, p_0) > 1/2$, the reply is bullish; if $\rho(r; \tau, p_0) < 1/2$, the reply is bearish. The parameter γ then amplifies this distance from the midpoint. Thus the source of bullishness is not γ alone. Bullishness requires the latent assessment to be sufficiently bullish, and categorical expression strengthens the resulting reply.

Agreement versus Bullish Bias The empirical results call for a distinction between agreement and bullish bias. Responsiveness to the user's thesis makes the reply move with the post. But this is not the same as a systematic bullish bias. A pure agreement mechanism would imply greater bullishness on bullish posts and greater bearishness on bearish posts. By contrast, the robustness results show that the AI-human gap remains positive even for bearish posts. The next proposition shows how this can arise from a learned bullish default and categorical expression.

Proposition 2 (Agreement versus bullish bias)**.** *Maintain Assumption 1. Then:*

(i) If $\gamma = 0$*, the reply is a convex combination of the learned default and the user's thesis:*

$$m(r) = (1 - \tau)p_0 + \tau r, \qquad m(r) - r = (1 - \tau)(p_0 - r).$$

Thus responsiveness alone pulls the reply toward p_0*. It does not generate a systematic bullish bias unless* p_0 *lies above the relevant support of post bullishness.*

(ii) The reply is bullish, $m(r) > 1/2$*, if and only if*

$$(1 - \tau)p_0 + \tau r > \frac{1}{2}. \tag{14}$$

Thus even a bearish thesis $r < 1/2$ can receive a bullish reply if the learned default p_0 is sufficiently high.

(iii) Relative to the post itself, the assistant overshoots upward, $m(r) > r$, if and only if

$$(1-\tau)(p_0 - r) + \gamma\left(r - \frac{1}{2}\right) > 0. \tag{15}$$

Proposition 2 is the core distinction. The parameter τ captures agreement or responsiveness. If $p_0 > r$, responsiveness pulls the reply upward relative to the post; if $p_0 < r$, it pulls the reply downward. This is not yet a systematic bullish bias. A bullish default p_0 can explain why bearish posts are shifted upward. But for sufficiently bullish posts with $r > p_0$, responsiveness to the default alone would pull the reply downward. Categorical expression changes this implication. When $r > 1/2$, the term $\gamma(r - \frac{1}{2})$ pushes the expressed reply upward. Thus p_0 supplies the bullish direction for low-r posts, while γ allows the assistant to continue shifting upward even for posts that are already bullish.

The next corollary makes this point explicit.

[Upward shifts for bearish and bullish posts] Maintain Assumption 1. Consider two posts $r_L < 1/2 < r_H$. The assistant shifts both posts upward relative to the post itself, $m(r_L) > r_L$ and $m(r_H) > r_H$, if and only if

$$(1-\tau)(p_0 - r_L) > \gamma\left(\frac{1}{2} - r_L\right), \tag{16}$$

$$(1-\tau)(p_0 - r_H) + \gamma\left(r_H - \frac{1}{2}\right) > 0. \tag{17}$$

The first condition requires a sufficiently bullish learned default. The second can hold either because the learned default is very bullish or because categorical expression is sufficiently strong.

Corollary 3.6 is the reduced-form counterpart of the robustness evidence. Upward shifts for bearish posts require a bullish learned default: without such a default, the latent assessment would not be pulled upward. Upward shifts for bullish posts require either an even higher default or categorical expression that amplifies the bullish latent assessment. Thus the empirical pattern is more than agreement. Agreement alone moves replies toward the user's thesis or toward the assistant's default. The combination of a bullish default and categorical expression can generate a systematic upward pull across both bearish and bullish parts of the post distribution.

Prompting and the Limits of Prompt Engineering The prompt experiment shows that asking the assistant to respond as a professional investor narrows the AI-human gap but

does not eliminate it. In the model, this can happen if the professional prompt changes the assistant's effective reply parameters. It may lower the learned default p_0, reduce categorical expression γ, or change responsiveness τ. The comparative statics show why the first two channels are especially natural for a bias that persists even on bearish posts.

Proposition 3 (Prompt moderation)**.** *Maintain Assumption 1. The reply rule satisfies:*

$$\frac{\partial m(r)}{\partial p_0} = \frac{1-\tau}{1-\gamma} > 0 \quad \text{if } \tau < 1, \tag{18}$$

$$\frac{\partial m(r)}{\partial \tau} = \frac{r - p_0}{1-\gamma}, \tag{19}$$

$$\frac{\partial}{\partial \gamma}\left|m(r) - \frac{1}{2}\right| = \frac{\left|\rho(r;\tau,p_0) - \frac{1}{2}\right|}{(1-\gamma)^2} \quad \text{if } \rho(r;\tau,p_0) \neq \frac{1}{2}. \tag{20}$$

Therefore, a prompt that lowers p_0 reduces the reply's bullishness for all $\tau < 1$. A prompt that lowers γ reduces the categorical expression of the reply whenever the latent assessment is away from the midpoint. A prompt that changes τ has a sign that depends on whether the user's thesis lies above or below the learned default p_0.

Proposition 3 gives a useful interpretation of the prompt results. If the professional-investor prompt merely reduced responsiveness to the user's message, its effect would depend on the position of the user's thesis relative to p_0. For posts below the learned default, lowering τ would actually pull the reply closer to p_0 and could make it more bullish. Thus a pure "less-agreement" interpretation is incomplete. The prompt evidence is more naturally viewed as a reduction in the model's effective bullish default, a reduction in categorical expression, or both. Because prompting changes the input but not the model weights, it can moderate these effective parameters in a given interaction without eliminating the post-trained conversational layer that generated them.

Connection to the Base-Model Diagnostic The base-model comparison helps interpret the source of the parameters in the reduced-form model. Let the base model and the user-facing assistant have effective parameters

$$(p_0^B, \tau^B, \gamma^B) \qquad \text{and} \qquad (p_0^I, \tau^I, \gamma^I),$$

respectively. The two models share architecture and pretraining corpus, but the instruction-tuned model adds the post-trained conversational layer. If the AI-human gap were only a property of the pretraining corpus, the base model would be expected to display a similar gap. Instead, the base model is much closer to the human benchmark. In the language of the model, the post-trained conversational layer appears to change one or more of the

effective reply parameters: the learned default p_0, the responsiveness parameter τ, and the categorical-expression parameter γ.

The important implication is that the empirical result is not simply that the assistant agrees with users. A pure agreement mechanism would primarily raise replies to bullish posts and lower replies to bearish posts. The documented positive gap on bearish posts points to a learned bullish default. Categorical expression then makes this default more visible in measured sentiment. The theory therefore summarizes the mechanism as:

$$\text{bullish bias} = \text{learned bullish default } (p_0) + \text{responsiveness } (\tau) + \text{categorical expression } (\gamma).$$

The first term supplies the direction, the second determines how strongly the user's thesis enters the reply, and the third determines how forcefully the latent assessment is expressed.

All proofs are in Appendix B.

4 Signal Content of AI-Augmented Replies

4.1 Motivation and Regression Design

The preceding sections document that instruction-tuned AI replies are systematically more bullish than human replies to the same investment theses. We now ask a narrower, empirically distinct question: does this excess bullishness destroy the return-relevant content of the underlying investment thesis, or does the AI-augmented signal retain associations with subsequent returns?

This exercise is explicitly descriptive and several important caveats apply before proceeding. First, Reddit posts are contemporaneous with news events: in our sample, approximately 83% of posts are matched to same-day returns, so the association between the probe score and near-horizon returns may partly reflect posts written in response to same-day price movements rather than independent information. Second, generated replies are constructed *ex post* from historical text, so a positive coefficient cannot be interpreted as evidence that investors could earn abnormal returns from the signal in real time. Third, the return exercise does not constitute a trading-strategy evaluation or a causal identification of AI's informational content. The purpose is narrower: to assess whether the excess bullishness documented above is purely uninformative amplification, or whether it preserves return-relevant variation in the underlying thesis.

We estimate panel regressions of the form

$$r_{i,t+h} \;=\; \alpha_i \;+\; \gamma_t \;+\; \beta\, s_{i,t} \;+\; \varepsilon_{i,t}, \tag{21}$$

where $r_{i,t+h}$ is the h-day-ahead return for stock i posted on date t, winsorized at the 1st and 99th percentiles in the main specification; $s_{i,t} \in [0,1]$ is the probe score p_A extracted from the relevant text construction; α_i are stock fixed effects absorbing time-invariant firm characteristics; and γ_t are date fixed effects absorbing market-wide shocks on the posting date. Standard errors are double-clustered by stock and date throughout. The coefficient $\hat{\beta} \times 100$ is interpreted as the percentage-point return association for a unit increase in the probe score. We use the instruct probe as the main specification, because it is the probe used in the baseline behavioral analysis in Section 3.3.

4.2 Baseline Return Association

Under the main specification—instruct probe, stock and date fixed effects—the post-alone, human, and AI-augmented signals all have positive one-day coefficients, although statistical strength varies across signals. The post-alone signal yields $\hat{\beta} \times 100 = +0.749$ ($t = 1.88$). The instruction-tuned neutral signal yields $+1.639$ ($t = 2.42$), and the professional-prompt signal yields $+1.671$ ($t = 2.00$). The base neutral and human-reply signals are also positive, but not statistically significant in this specification: $+0.737$ ($t = 1.48$) and $+1.097$ ($t = 1.00$), respectively.

Thus the AI-augmented signals are positively associated with near-horizon returns under the main fixed-effects specification. Their point estimates are larger than the post-alone coefficient, although we do not interpret these differences as evidence that the AI signals are economically or statistically superior. Rather, the evidence suggests that the AI-generated replies are not merely adding random bullish language on top of the investment thesis; the resulting sentiment scores retain return-relevant variation in the main specification. We do not claim that the AI signal is "better" than the post or human signal in a causal or implementable sense—given the contemporaneous matching and *ex post* construction—but the behavioral amplification documented earlier does not appear to erase the return-relevant variation in the post. Full fixed-effects specifications across all signals are reported in Appendix Table 20.

4.3 Horizon and Subsample Robustness

The horizon profile, reported in Appendix Table 21, shows that the return association peaks over the one-to-five-day window and fades by the ten-day horizon for all signals. For the instruct neutral signal, the coefficients are $+1.639$ ($t = 2.42$), $+2.442$ ($t = 2.93$), $+2.414$ ($t = 1.98$), and $+3.054$ ($t = 1.73$) at horizons of one through five days, before declining to $+0.607$ ($t = 0.53$) at ten days. The professional-prompt signal follows a similar pattern but retains significance at the five-day horizon: $+4.026$ ($t = 2.04$), compared with the neutral

signal's +3.054 ($t = 1.73$). The short-horizon concentration is consistent with a near-term signal-content interpretation rather than a claim of persistent abnormal returns.

Appendix Table 23 reports a comprehensive robustness scorecard across fixed-effects specifications, sample cuts, return definitions, clustering alternatives, signal scaling, and sample exclusions. The results are not uniform. The positive association is present and statistically significant under date-only fixed effects (instruct neutral +1.109, $t = 2.36$), the main two-way fixed-effects specification (+1.639, $t = 2.42$), and the two-way specification with controls for log market capitalization and post confidence (+1.615, $t = 2.42$). The association is also stable across return definitions (raw, 1% winsorized, and 5% winsorized) and clustering choices (stock-only, date-only, and double-clustered).

Several important limitations emerge from the subsample analysis. The association is weak and statistically insignificant in the pre-2020 period, where the sample is thin (933 posts). Small-cap stocks yield coefficients near zero or negative under the main specification. Posts matched to next-day or later returns—which are less exposed to same-day contamination—are also insignificant. The 6–20 comment subsample yields a negative and significant coefficient for the neutral signal (−2.622, $t = -2.03$), which we do not fully explain. The return association is concentrated in the post-2020 period, large-cap stocks, and same-day or near-horizon matched posts. We therefore do not present these results as evidence of a universal or robust return-predictive relation. Appendix Table 22 reports the full subsample and controls analysis.

4.4 Incremental Return Association Beyond the Matched Base Signal

Because both the instruction-tuned and base generators share the same pretraining architecture and corpus, including the base-model signal as an additional regressor provides a partial control for return-relevant information common to both generated replies. Under this joint specification, the instruction-tuned coefficient captures the incremental return association of the instruction-tuned signal beyond the matched base signal. This specification does not eliminate all lookahead or memorization concerns, but it reduces the concern that the return association is driven solely by shared pretraining.

Under the instruct probe, the joint regression of next-day returns on both the instruct neutral signal and the base neutral signal yields instruct coefficients of +1.825 ($t = 2.50$) at one day and +1.991 ($t = 2.56$) at two days, while the base coefficient is −0.390 ($t = -0.90$) at one day and +0.948 ($t = 1.50$) at two days—never significant across horizons. For the professional-prompt signal, the instruct coefficients are +1.781 ($t = 1.89$), +2.386 ($t = 2.22$), +3.357 ($t = 2.00$), and +4.105 ($t = 2.05$) at one through five days, while the base coefficient

is insignificant at all horizons. The instruction-tuned signal therefore retains a positive incremental return association at the short horizons where the signal is strongest, even after controlling for the matched base-model signal. The full joint-regression results are in Appendix Table 24.

4.5 Interpretation

The return tests sharpen the interpretation of the bullish-bias result. The AI reply is more bullish than the human benchmark, including for bearish posts, which is why we interpret the behavior as a directional bias rather than mere agreement. At the same time, the AI-augmented signal remains positively associated with near-horizon returns in the main specifications. Thus the bias is not simply random optimism layered on an otherwise uninformative reply.

A useful interpretation is that the instruction-tuned model amplifies the directional content of investment theses. This amplification produces excess bullishness relative to human discussion, but it does not necessarily erase the return-relevant component of the underlying signal. The professional-investor prompt reduces the measured AI-human bullish gap (mean reduction of 0.039 probe-score units, paired $t = 33.62$, $p < 0.001$, across 14,554 comparable posts) and, in several horizon specifications, produces coefficients that are at least as large as those of the neutral AI signal. This pattern is consistent with the possibility that reducing excess bullishness need not weaken the signal. However, the direct Wald tests do not show statistically reliable outperformance of the professional signal over the neutral signal, so we treat this evidence as suggestive rather than conclusive.

The market evidence should therefore be read as descriptive evidence on short-horizon signal content, not as a welfare analysis or trading recommendation. Whether the return association survives in a real-time implementation, with returns measured strictly after the posting time and without same-day contamination, is an important question for future work.

5 Conclusion

This paper studies whether user-facing AI assistants shift investor discussion in a different direction from human replies to the same investment theses. Using Reddit `r/wallstreetbets` posts, we compare AI-generated replies with actual human top-level replies to the same posts, and measure each reply's effect on conversation-level sentiment using a probabilistic probing procedure.

Four findings emerge. First, user-facing AI replies produce systematically larger upward sentiment shifts than human replies to the same posts. This gap remains positive even for

bearish posts, indicating that the result is not merely stronger agreement with the investor's thesis. Second, prompting the model to respond as a professional investor narrows the AI-human bullish gap but does not eliminate it: about 86 percent of the gap persists under the baseline probe. Third, the instruction-tuned model is consistently more bullish than the matched base model, suggesting that the post-trained conversational layer amplifies bullishness beyond what is already present in the base model. Fourth, the market exercise is descriptive: AI-augmented signals retain positive near-horizon return associations in the main specifications, but the evidence is not uniform across samples, horizons, or timing cuts.

We interpret these findings with a reduced-form reply model. The assistant's response reflects a learned default orientation, the weight it places on the investor's message, and its tendency to express views categorically rather than in heavily hedged language. Responsiveness alone is an agreement force: it pulls replies toward the user's thesis or toward the model's default. A systematic bullish bias, including on bearish posts, requires a sufficiently bullish learned default, with categorical expression amplifying the resulting assessment. A professional-role prompt can moderate the effective parameters and narrow the gap, but does not remove the post-trained conversational layer.

The broader implication is that AI matters in financial communication not only because of what models know, but also because of how deployed assistants reply. A post-trained conversational system can differ from its matched base model in how it translates user-supplied investment views into financial language. In settings where investors seek interpretation, confirmation, or advice, this behavioral layer may matter as much as the underlying information set.

Several questions remain open. The matched base-vs-instruct diagnostic uses one model family, and human Reddit replies provide a benchmark from observed investor discussion rather than a randomized counterfactual. Future work should examine other model families, platforms, asset classes, and user-level outcomes such as trading decisions, comment dynamics, and the persistence of optimistic narratives. It should also study whether interface design, disclosure, or uncertainty-preserving prompts can reduce the AI-human gap without eliminating the usefulness of conversational AI. The central descriptive fact remains: in our WSB setting, when responding to the same investor-authored posts, user-facing AI replies generate substantially larger upward sentiment shifts than human replies.

Bibliography

Bai, J. J., Boyson, N., Cao, Y., Liu, M., and Wan, C. (2025). Executives vs chatbots: Unmasking insights through human-ai differences in earnings conference q&a. Working paper.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4):1645–1680.

Balasubramaniam, S. and Sabat, J. (2025). The conversational structure of investor disagreement: Evidence from reddit threads. *Available at SSRN 5984474*.

Bhagwat, V., Cookson, J. A., Dim, C., and Niessner, M. (2025). The market's mirror: Revealing investor disagreement with llms. *FEB-RN Research Paper*, (107).

Bini, P., Cong, L. W., Huang, X., and Jin, L. J. (2026). Behavioral economics of ai: Llm biases and corrections. *arXiv preprint arXiv:2602.09362*.

Blankespoor, E., Croom, J., and Grant, S. M. (2025). Generative ai and investor processing of financial information. Working paper.

Bradley, D., Gao, S., and Scopino, A. (2024). Do reddit's WallStreetBets investment discussions inform or mislead retail investors? *Journal of Financial Economics*, 153:103782.

Chen, H., Didisheim, A., and Somoza, L. (2025). Out of the (Black) box: AI as conditional probability. Working paper.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

Engelberg, J., Manela, A., Mullins, W., and Vulicevic, L. (2025). Entity neutering. *Available at SSRN*.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically consistent large language models. Working paper.

Hirshleifer, D., Lin, P., Qiguang, W., Zhang, W., and Zhang, X. (2025). Ai, opinion ecosystems, and finance.

Hong, H. and Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2):109–128.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Liu, Y., Liu, J., and Wang, T. (2025). Large language model disagreement. *Available at SSRN 5395056*.

Mackowiak, B. and Wiederholt, M. (2009). Business cycle dynamics under rational inattention. *Review of Economic Studies*, 76(4):1502–1528.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *American Economic Review*, 95(4):1031–1053.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Prendergast, C. (1993). A theory of "yes men". *American Economic Review*, 83(4):757–770.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

A Empirical Appendix

A.1 Generation and Probe Prompts

The following prompts were used verbatim in all experiments. For the neutral generation condition, the model received the post text preceded by:

> *You will read a stock-related message from an online discussion forum. Write a natural reply to this message as if you were participating in the discussion.*

For the professional investor condition, the same prompt was prefaced with the single sentence: *You are a professional investor.*

The post sentiment probe presented the following question after the message text:

> *Based only on this message, which direction seems more likely for the stock in the short run? A = upward move, B = downward move. Reply with only one character: A or B.*

The response probe applied the same forced choice to the reply text in isolation, prefaced by "You will read a response to a stock-related message." The joint post-and-response probe applied it to the post and response presented together, with the post labeled "Message" and the reply labeled "Response." All probing prompts were applied identically to AI-generated and human-authored text.

A.2 Model Specifications and Computational Details

We use Mistral-7B-v0.1 (`mistralai/Mistral-7B-v0.1`) as the base model and Mistral-7B-Instruct-v0.1 (`mistralai/Mistral-7B-Instruct-v0.1`) as the instruction-tuned model (Jiang et al., 2023). Both models have 7 billion parameters and share the same transformer architecture and pretraining corpus. The instruct variant additionally underwent instruction tuning and is used with the associated chat-template format that elicits helpful assistant behavior. Observed differences therefore capture the composite behavioral role of the post-trained conversational layer. Generation responses were capped at 120 new tokens. We applied temperatures 0.0, 0.5, and 1.0 with two independent replications at each. Probing was always deterministic. All inference was performed on four NVIDIA A5000 GPUs (24 GB VRAM each) using `bfloat16` precision and the HuggingFace Transformers library.

A.3 Human Comment Averaging

For posts with multiple depth-1 human comments, we average the probed joint sentiment scores across all comments before computing the human benchmark shift. Formally, $\bar{p}_A(\text{post}+$

human) is the arithmetic mean of $p_A(\text{post} + \text{comment}_k)$ over all depth-1 comments k on a given post. The response-relative human benchmark is computed analogously as the arithmetic mean of $p_A(\text{comment}_k \text{ alone})$ across depth-1 comments. Both averages are interpreted as human discussion benchmarks for the post, not as randomized counterfactuals. Posts with no depth-1 comments are excluded from all AI-versus-human comparisons.

A.4 Two-Probe Robustness Classification

Our robust shift classification requires that both the base probe and the instruct probe agree on the sign of the shift for a given post-response pair. A shift is classified as robustly upward if $\delta^{\text{base}} > 0$ and $\delta^{\text{instruct}} > 0$, robustly downward if both are negative, and ambiguous otherwise. The ambiguous category accounts for roughly 52 to 58 percent of post-level observations. Our reliance on aggregate statistics rather than individual classifications ensures that conclusions are not sensitive to this post-level noise.

A.5 Regression Estimation Details

The within-post regression is estimated by demeaning all variables within post (Frisch-Waugh-Lovell theorem), absorbing post fixed effects without constructing the dummy matrix explicitly. Standard errors are clustered at the post level using the sandwich estimator with the small-sample correction $\frac{n}{n-k} \cdot \frac{G}{G-1}$, where $G = 15{,}206$ clusters. The within-R^2 values of 0.062 and 0.011 for the base and instruct probes respectively indicate that the six treatment variables explain approximately 6 and 1 percent of within-post variation. The marginal effects in Table 10 are linear combinations of the estimated coefficients with standard errors computed by the delta method using the full estimated covariance matrix.

Table 1: Summary Statistics

				Percentiles					Tail shares
Variable	N	Mean	SD	P10	P25	P50	P75	P90	>0.9 / <0.1
Panel A: Post universe (28,922 posts)									
Base probe	28,922	0.562	0.171	0.342	0.453	0.562	0.677	0.783	2.3% / 0.6%
Instruct probe	28,922	0.632	0.346	0.038	0.349	0.737	0.953	0.994	33.5% / 13.7%
Panel B: Generation universe (≈173,500 per cell, neutral prompt)									
Gen=base, probe=base	173,516	0.648	0.226	0.323	0.516	0.679	0.818	0.924	12.9% / 2.2%
Gen=base, probe=instruct	173,484	0.616	0.300	0.148	0.363	0.693	0.893	0.955	23.1% / 6.5%
Gen=instruct, probe=base	173,532	0.735	0.179	0.492	0.637	0.769	0.874	0.934	17.7% / 0.4%
Gen=instruct, probe=instruct	173,484	0.662	0.261	0.269	0.469	0.719	0.899	0.963	24.3% / 2.0%
Panel C: AI-versus-human comparison sample (15,206 posts)									
Human, probe=base	15,206	0.624	0.130	0.442	0.574	0.651	0.696	0.755	0.8% / 0.1%
Human, probe=instruct	15,206	0.397	0.187	0.165	0.263	0.379	0.516	0.639	1.3% / 3.5%

Notes: p_A is the probability of an upward stock move from the two-class softmax over probe-model logits for tokens "A" and "B". Panel A: full post universe, p_A of the post. Panel B: individual generated replies; Gen=instruct rows are the AI model of interest. Panel C: human comment scores averaged equally across depth-1 comments per post. Tail shares give the fraction with $p_A > 0.9$ or $p_A < 0.1$.

Table 2: Instruction-Tuned AI versus Human Investor Replies: Sentiment Shift from Post

Temp	N	AI shift	Human shift	AI−Human	t-stat
Panel A: Neutral prompt, instruct probe					
0.0	15,206	+0.030	−0.234	+0.264***	117.70
0.5	15,206	+0.031	−0.234	+0.265***	127.97
1.0	15,206	+0.031	−0.234	+0.265***	129.65
Panel B: Professional prompt, instruct probe					
0.0	15,206	−0.012	−0.234	+0.223***	97.13
0.5	15,206	−0.006	−0.234	+0.228***	110.87
1.0	15,206	−0.002	−0.234	+0.232***	115.16

Notes: Instruction-tuned generator only. AI shift is $\delta^{\text{joint}} = p_A(\text{post} + \text{AI reply}) - p_A(\text{post})$, averaged across two replications. The human shift $\bar{\delta}^H$ is the human discussion benchmark, constant within each panel. AI−Human is the paired difference tested with a paired t-test. *** $p < 0.01$.

Table 3: Effect of Prompt Framing on the AI-Human Gap (Instruction-Tuned Generator)

			Mean AI shifts		AI-human gap		Gap
Temp	N	$\bar{\delta}^H$	$\bar{\delta}^N$	$\bar{\delta}^P$	Gap_N	Gap_P	reduction
Instruct probe							
0.0	15,206	−0.234	+0.030	−0.012	+0.264	+0.223	+0.041***
0.5	15,206	−0.234	+0.031	−0.006	+0.265	+0.228	+0.037***
1.0	15,206	−0.234	+0.031	−0.002	+0.265	+0.232	+0.033***

Notes: Instruction-tuned generator only. Gap reduction is Gap_N minus Gap_P, tested with a paired t-test. Full results including base-generator rows are in Table 8. *** $p < 0.01$.

Table 4: Decomposition of the AI-Human Bullish Gap (Instruction-Tuned Generator)

Probe	Total	Insensitive	Sensitive	Insens. (%)	Sens. (%)
Instruct probe	+0.265	+0.228	+0.037	85.9	14.1
Base probe	+0.108	+0.079	+0.029	73.0	27.0

Notes: Instruction-tuned generator only. Total gap is AI shift minus human shift under the neutral prompt, averaged across temperatures. Insensitive: residual gap under the professional investor prompt. Sensitive: Total − Insensitive. Full results including base-generator rows are in Table 11.

Table 5: Probe Model Polarisation: Base versus Instruct Probe

		Base probe				Instruct probe				Difference	
Prompt	Level	Mean	$\lvert p-0.5\rvert$	>0.9	<0.1	Mean	$\lvert p-0.5\rvert$	>0.9	<0.1	$\hat{\Delta}$	t
Post p_A (identical across prompt conditions)											
	Post p_A	0.562	0.146	2.3%	0.6%	0.632	0.335	33.5%	13.7%	$+0.070^{***}$	30.90
AI response p_A											
Neutral	AI p_A	0.692	0.248	15.3%	1.3%	0.639	0.278	23.7%	4.3%	-0.053^{***}	-94.03
Professional	AI p_A	0.680	0.240	13.9%	1.4%	0.626	0.273	22.4%	4.3%	-0.054^{***}	-95.15

Notes: Each row compares base and instruct probe scores on identical texts. $\lvert p-0.5\rvert$ is mean absolute distance from 0.5. $\hat{\Delta}$ is the instruct-minus-base mean difference tested by a paired t-test. *** $p < 0.01$.

Table 6: AI versus Human Sentiment Shift: Both Generators and Both Probes (Diagnostic)

Generator	Temp	N	AI shift	Human shift	AI−Human	t-stat	AI↑ (%)
Panel A: Neutral prompt, base probe							
base	0.0	15,206	+0.099	+0.065	+0.034***	18.07	66.7
	0.5	15,206	+0.095	+0.065	+0.030***	19.50	68.1
	1.0	15,206	+0.067	+0.065	+0.002	1.08	61.8
instruct	0.0	15,206	+0.161	+0.065	+0.096***	57.90	76.8
	0.5	15,206	+0.168	+0.065	+0.103***	71.89	80.3
	1.0	15,206	+0.189	+0.065	+0.124***	88.97	83.1
Panel B: Neutral prompt, instruct probe							
base	0.0	15,206	−0.020	−0.234	+0.215***	78.90	46.1
	0.5	15,206	+0.016	−0.234	+0.251***	112.75	47.3
	1.0	15,206	−0.038	−0.234	+0.196***	93.35	41.4
instruct	0.0	15,206	+0.030	−0.234	+0.264***	117.70	48.7
	0.5	15,206	+0.031	−0.234	+0.265***	127.97	48.4
	1.0	15,206	+0.031	−0.234	+0.265***	129.65	48.3
Panel C: Professional prompt, base probe							
base	0.0	15,206	+0.105	+0.065	+0.040***	21.23	67.5
	0.5	15,206	+0.100	+0.065	+0.035***	22.58	68.6
	1.0	15,206	+0.072	+0.065	+0.007***	3.97	62.8
instruct	0.0	15,206	+0.132	+0.065	+0.066***	38.91	72.1
	0.5	15,206	+0.139	+0.065	+0.074***	51.71	75.7
	1.0	15,206	+0.161	+0.065	+0.095***	67.86	79.2
Panel D: Professional prompt, instruct probe							
base	0.0	15,206	+0.010	−0.234	+0.244***	88.88	49.2
	0.5	15,206	+0.022	−0.234	+0.257***	114.89	48.1
	1.0	15,206	−0.039	−0.234	+0.195***	93.67	41.0
instruct	0.0	15,206	−0.012	−0.234	+0.223***	97.13	45.7
	0.5	15,206	−0.006	−0.234	+0.228***	110.87	44.9
	1.0	15,206	−0.002	−0.234	+0.232***	115.16	44.9

Notes: The instruction-tuned generator is the AI model of interest; base-generator rows are the matched diagnostic (Section 3.4). AI shift is $\delta^{\text{joint}} = p_A(\text{post} + \text{AI reply}) - p_A(\text{post})$. The human shift $\bar{\delta}^H$ is the human discussion benchmark, constant within each panel. AI↑ is the share of post-level AI shifts that are positive. ***, **, * denote significance at the 1%, 5%, and 10% levels.

Table 7: Robust Shift Classification: Both Probes Must Agree

Generator	Temp	Object	Robust↑ (%)	Robust↓ (%)	Ambiguous (%)
Panel A: Neutral prompt					
base	0.0	AI	30.7	17.2	52.1
	0.5	AI	31.9	16.6	51.5
	1.0	AI	25.3	22.2	52.5
instruct	0.0	AI	36.0	9.9	54.1
	0.5	AI	37.4	8.7	53.9
	1.0	AI	38.9	7.6	53.5
Human benchmark			17.3	25.2	57.5
Panel B: Professional prompt					
base	0.0	AI	33.6	16.4	50.0
	0.5	AI	33.0	16.4	50.6
	1.0	AI	25.8	21.9	52.2
instruct	0.0	AI	31.9	13.6	54.5
	0.5	AI	32.9	12.2	55.0
	1.0	AI	34.5	10.4	55.0
Human benchmark			17.3	25.2	57.5

Notes: A shift is classified as robustly upward (downward) if both probes independently assign a positive (negative) shift to the same post-response pair. Ambiguous means the two probes disagree on sign. $N =$ 15,206 posts in all cells.

Table 8: Effect of Prompt Framing on the AI-Human Sentiment Gap

			Mean shifts			AI-human gap		Gap reduction	
Gen	Temp	N	$\bar{\delta}^N$	$\bar{\delta}^P$	$\bar{\delta}^H$	Gap_N	Gap_P	ΔGap	t
Panel A: Base probe									
base	0.0	15,206	+0.099	+0.105	+0.065	+0.034	+0.040	-0.006^{**}	−2.50
	0.5	15,206	+0.095	+0.100	+0.065	+0.030	+0.035	-0.005^{***}	−3.14
	1.0	15,206	+0.067	+0.072	+0.065	+0.002	+0.007	-0.005^{**}	−2.54
instruct	0.0	15,206	+0.161	+0.132	+0.065	+0.096	+0.066	$+0.029^{***}$	+17.36
	0.5	15,206	+0.168	+0.139	+0.065	+0.103	+0.074	$+0.029^{***}$	+23.08
	1.0	15,206	+0.189	+0.161	+0.065	+0.124	+0.095	$+0.029^{***}$	+22.26
Panel B: Instruct probe									
base	0.0	15,206	−0.020	+0.010	−0.234	+0.215	+0.244	-0.030^{***}	−11.25
	0.5	15,206	+0.016	+0.022	−0.234	+0.251	+0.257	-0.006^{***}	−2.82
	1.0	15,206	−0.038	−0.039	−0.234	+0.196	+0.195	+0.001	+0.52
instruct	0.0	15,206	+0.030	−0.012	−0.234	+0.264	+0.223	$+0.041^{***}$	+21.18
	0.5	15,206	+0.031	−0.006	−0.234	+0.265	+0.228	$+0.037^{***}$	+24.66
	1.0	15,206	+0.031	−0.002	−0.234	+0.265	+0.232	$+0.033^{***}$	+20.31

Notes: The instruction-tuned generator is the AI model of interest; base-generator rows are the matched diagnostic. ΔGap is Gap_N minus Gap_P, tested with a paired t-test. A positive ΔGap means the neutral prompt produces a larger AI-human gap. ***, **, * denote significance at the 1%, 5%, and 10% levels.

Table 9: Regression Decomposition of AI Sentiment Shift

	Probe model	
Variable	Base	Instruct
Instruct generator	$+0.086^{***}$	$+0.044^{***}$
	(0.001)	(0.002)
Professional prompt	$+0.005^{***}$	$+0.011^{***}$
	(0.001)	(0.001)
Temperature (centred)	-0.033^{***}	-0.028^{***}
	(0.002)	(0.002)
Instruct × Professional	-0.034^{***}	-0.049^{***}
	(0.001)	(0.002)
Instruct × Temperature	$+0.061^{***}$	$+0.039^{***}$
	(0.002)	(0.002)
Professional × Temperature	+0.000	-0.011^{***}
	(0.002)	(0.002)
Observations	182,472	182,472
Within-R^2	0.062	0.011
Post FE	Yes	Yes
Clustered SE	Post	Post

Notes: Dependent variable is the AI shift δ^{joint}, averaged across two replications within each cell. Post fixed effects absorbed by within-group demeaning. Standard errors (in parentheses) clustered at the post level ($G = 15{,}206$ clusters). Temperature centred at 0.5. ***, **, * denote significance at the 1%, 5%, and 10% levels.

Table 10: Regression Decomposition of AI Sentiment Shift: Marginal Effects

	Probe model	
	Base	Instruct
Panel A: Net instruct-generator gap at each temperature (neutral prompt)		
$T = 0.0$	$+0.055^{***}$	$+0.025^{***}$
	(0.002)	(0.002)
$T = 0.5$	$+0.086^{***}$	$+0.044^{***}$
	(0.001)	(0.002)
$T = 1.0$	$+0.116^{***}$	$+0.064^{***}$
	(0.002)	(0.002)
Panel B: Net professional-prompt effect, by generator		
Base generator		
$T = 0.0$	$+0.005^{***}$	$+0.017^{***}$
	(0.001)	(0.002)
$T = 0.5$	$+0.005^{***}$	$+0.011^{***}$
	(0.001)	(0.001)
$T = 1.0$	$+0.005^{***}$	$+0.006^{***}$
	(0.001)	(0.002)
Instruct generator		
$T = 0.0$	-0.029^{***}	-0.032^{***}
	(0.002)	(0.002)
$T = 0.5$	-0.029^{***}	-0.037^{***}
	(0.002)	(0.002)
$T = 1.0$	-0.029^{***}	-0.043^{***}
	(0.002)	(0.002)

Notes: All quantities are linear combinations of Table 9 coefficients; standard errors by delta method using the full covariance matrix. ***, **, * denote significance at the 1%, 5%, and 10% levels.

Table 11: Decomposition of the AI-Human Bullish Gap

Probe	Generator	Total	Insensitive	Sensitive	Insens. (%)	Sens. (%)
Instruction-tuned generator (primary)						
Base probe	instruct	+0.108	+0.079	+0.029	73.0	27.0
Instruct probe	instruct	+0.265	+0.228	+0.037	85.9	14.1
Base generator (diagnostic benchmark)						
Base probe	base	+0.022	+0.027	−0.005	123.8	−23.8
Instruct probe	base	+0.221	+0.232	−0.011	105.2	−5.2

Notes: Total gap is AI shift minus human benchmark shift under the neutral prompt, averaged across temperatures. Sensitive share is removed by professional framing. Base-generator rows are included as a diagnostic benchmark.

Table 12: Robustness of the Bullish Bias: Joint Shift Measure (Panel A: Neutral Prompt)

		All posts		Bullish posts		Bearish posts	
Subsample	n	Gap_B	Gap_I	Gap_B	Gap_I	Gap_B	Gap_I
Panel A: Neutral generation prompt							
Full sample	15,206	$+0.108^{***}$	$+0.265^{***}$	$+0.112^{***}$	$+0.262^{***}$	$+0.100^{***}$	$+0.269^{***}$
		(84.40)	(136.94)	(71.57)	(107.28)	(45.93)	(85.17)
Low tercile	4,919	$+0.107^{***}$	$+0.266^{***}$	$+0.107^{***}$	$+0.262^{***}$	$+0.107^{***}$	$+0.270^{***}$
		(48.32)	(77.73)	(34.38)	(55.98)	(33.97)	(53.94)
Mid tercile	5,021	$+0.109^{***}$	$+0.268^{***}$	$+0.113^{***}$	$+0.267^{***}$	$+0.101^{***}$	$+0.268^{***}$
		(49.26)	(80.48)	(42.32)	(63.42)	(26.04)	(49.67)
High tercile	4,614	$+0.108^{***}$	$+0.264^{***}$	$+0.116^{***}$	$+0.261^{***}$	$+0.086^{***}$	$+0.273^{***}$
		(45.79)	(74.81)	(44.53)	(62.81)	(17.00)	(40.72)
1 comment	7,586	$+0.116^{***}$	$+0.283^{***}$	$+0.123^{***}$	$+0.280^{***}$	$+0.104^{***}$	$+0.289^{***}$
		(59.00)	(95.36)	(50.96)	(75.43)	(30.78)	(58.38)
2–5 comments	4,452	$+0.091^{***}$	$+0.251^{***}$	$+0.095^{***}$	$+0.247^{***}$	$+0.084^{***}$	$+0.258^{***}$
		(40.49)	(75.53)	(35.16)	(57.87)	(21.55)	(48.62)
6–20 comments	2,452	$+0.113^{***}$	$+0.248^{***}$	$+0.110^{***}$	$+0.248^{***}$	$+0.118^{***}$	$+0.248^{***}$
		(43.66)	(58.44)	(34.41)	(45.79)	(26.94)	(36.30)
>20 comments	716	$+0.103^{***}$	$+0.214^{***}$	$+0.101^{***}$	$+0.214^{***}$	$+0.106^{***}$	$+0.214^{***}$
		(22.40)	(30.97)	(17.86)	(25.08)	(13.50)	(18.15)
Pre-2020	649	$+0.099^{***}$	$+0.201^{***}$	$+0.097^{***}$	$+0.192^{***}$	$+0.101^{***}$	$+0.214^{***}$
		(18.90)	(23.00)	(14.35)	(17.62)	(12.28)	(14.83)
2020	3,162	$+0.112^{***}$	$+0.248^{***}$	$+0.114^{***}$	$+0.244^{***}$	$+0.108^{***}$	$+0.252^{***}$
		(47.74)	(61.86)	(37.97)	(45.52)	(29.17)	(42.09)
2021	9,037	$+0.113^{***}$	$+0.287^{***}$	$+0.117^{***}$	$+0.282^{***}$	$+0.105^{***}$	$+0.296^{***}$
		(66.87)	(114.03)	(58.01)	(90.66)	(34.59)	(69.27)
2022	1,706	$+0.075^{***}$	$+0.214^{***}$	$+0.085^{***}$	$+0.215^{***}$	$+0.061^{***}$	$+0.214^{***}$
		(17.06)	(35.46)	(15.09)	(27.61)	(8.65)	(22.26)
Small cap	8,058	$+0.111^{***}$	$+0.267^{***}$	$+0.117^{***}$	$+0.264^{***}$	$+0.102^{***}$	$+0.272^{***}$
		(60.25)	(101.41)	(50.52)	(77.98)	(33.67)	(64.87)
Large cap	6,489	$+0.104^{***}$	$+0.265^{***}$	$+0.107^{***}$	$+0.263^{***}$	$+0.098^{***}$	$+0.268^{***}$
		(57.43)	(88.41)	(49.17)	(70.81)	(30.40)	(52.95)
Same-day ($\Delta = 0$)	12,121	$+0.109^{***}$	$+0.265^{***}$	$+0.113^{***}$	$+0.262^{***}$	$+0.102^{***}$	$+0.272^{***}$
		(75.62)	(122.70)	(64.03)	(95.25)	(41.36)	(77.46)
Next-day+ ($\Delta \geq 1$)	2,433	$+0.103^{***}$	$+0.269^{***}$	$+0.108^{***}$	$+0.272^{***}$	$+0.093^{***}$	$+0.263^{***}$
		(33.77)	(55.26)	(28.83)	(45.08)	(17.97)	(32.00)

Gap_B and Gap_I denote the mean paired difference under the base and instruct probe models respectively. t-statistics from paired t-tests on post-level observations are reported in parentheses. Post direction is defined using the base-probe post score: bullish posts have $p_A^{\text{base}}(\text{post}) > 0.5$; bearish posts have $p_A^{\text{base}}(\text{post}) \leq 0.5$. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. Panel B (professional prompt) continues in Table 13.

Table 13: Robustness of the Bullish Bias: Joint Shift Measure (Panel B: Professional Prompt)

		All posts		Bullish posts		Bearish posts	
Subsample	n	Gap_B	Gap_I	Gap_B	Gap_I	Gap_B	Gap_I
Panel B: Professional investor prompt							
Full sample	15,206	$+0.079^{***}$	$+0.228^{***}$	$+0.082^{***}$	$+0.229^{***}$	$+0.072^{***}$	$+0.226^{***}$
		(62.49)	(121.09)	(53.65)	(96.87)	(33.30)	(72.68)
Low tercile	4,919	$+0.079^{***}$	$+0.228^{***}$	$+0.081^{***}$	$+0.226^{***}$	$+0.078^{***}$	$+0.231^{***}$
		(36.14)	(68.20)	(26.55)	(49.85)	(24.52)	(46.61)
Mid tercile	5,021	$+0.077^{***}$	$+0.228^{***}$	$+0.079^{***}$	$+0.233^{***}$	$+0.074^{***}$	$+0.221^{***}$
		(35.98)	(70.35)	(30.57)	(56.39)	(19.42)	(42.13)
High tercile	4,614	$+0.079^{***}$	$+0.226^{***}$	$+0.087^{***}$	$+0.228^{***}$	$+0.059^{***}$	$+0.223^{***}$
		(33.82)	(66.64)	(33.55)	(57.56)	(11.69)	(33.64)
1 comment	7,586	$+0.085^{***}$	$+0.245^{***}$	$+0.091^{***}$	$+0.246^{***}$	$+0.075^{***}$	$+0.245^{***}$
		(43.40)	(83.80)	(38.04)	(67.60)	(22.00)	(49.55)
2–5 comments	4,452	$+0.062^{***}$	$+0.214^{***}$	$+0.067^{***}$	$+0.215^{***}$	$+0.054^{***}$	$+0.213^{***}$
		(28.05)	(66.12)	(25.23)	(52.05)	(14.02)	(40.77)
6–20 comments	2,452	$+0.088^{***}$	$+0.211^{***}$	$+0.083^{***}$	$+0.214^{***}$	$+0.096^{***}$	$+0.207^{***}$
		(35.16)	(54.47)	(26.96)	(43.77)	(22.64)	(32.47)
>20 comments	716	$+0.083^{***}$	$+0.180^{***}$	$+0.081^{***}$	$+0.182^{***}$	$+0.086^{***}$	$+0.178^{***}$
		(19.61)	(28.60)	(15.65)	(22.56)	(11.80)	(17.63)
Pre-2020	649	$+0.073^{***}$	$+0.175^{***}$	$+0.070^{***}$	$+0.173^{***}$	$+0.078^{***}$	$+0.177^{***}$
		(15.01)	(21.65)	(11.55)	(16.76)	(9.62)	(13.69)
2020	3,162	$+0.089^{***}$	$+0.216^{***}$	$+0.089^{***}$	$+0.219^{***}$	$+0.088^{***}$	$+0.212^{***}$
		(38.62)	(56.72)	(30.03)	(43.33)	(24.31)	(36.60)
2021	9,037	$+0.081^{***}$	$+0.244^{***}$	$+0.085^{***}$	$+0.243^{***}$	$+0.074^{***}$	$+0.247^{***}$
		(48.36)	(98.43)	(42.71)	(79.83)	(24.16)	(57.60)
2022	1,706	$+0.049^{***}$	$+0.184^{***}$	$+0.060^{***}$	$+0.189^{***}$	$+0.033^{***}$	$+0.177^{***}$
		(11.15)	(31.27)	(10.73)	(24.76)	(4.67)	(19.11)
Small cap	8,058	$+0.082^{***}$	$+0.230^{***}$	$+0.088^{***}$	$+0.232^{***}$	$+0.073^{***}$	$+0.228^{***}$
		(45.24)	(89.87)	(39.03)	(70.94)	(24.12)	(55.22)
Large cap	6,489	$+0.074^{***}$	$+0.225^{***}$	$+0.076^{***}$	$+0.226^{***}$	$+0.071^{***}$	$+0.222^{***}$
		(41.48)	(77.22)	(35.23)	(62.86)	(22.30)	(44.85)
Same-day ($\Delta = 0$)	12,121	$+0.080^{***}$	$+0.227^{***}$	$+0.084^{***}$	$+0.227^{***}$	$+0.074^{***}$	$+0.227^{***}$
		(56.48)	(107.94)	(48.40)	(85.73)	(30.31)	(65.59)
Next-day+ ($\Delta \geq 1$)	2,433	$+0.071^{***}$	$+0.232^{***}$	$+0.077^{***}$	$+0.238^{***}$	$+0.061^{***}$	$+0.222^{***}$
		(23.52)	(48.87)	(20.42)	(40.39)	(12.02)	(27.63)

Continuation of Table 12. Gap_B and Gap_I denote the mean paired difference under the base and instruct probe models respectively. t-statistics from paired t-tests on post-level observations are reported in parentheses. Post direction is defined using the base-probe post score: bullish posts have $p_A^{\text{base}}(\text{post}) > 0.5$; bearish posts have $p_A^{\text{base}}(\text{post}) \leq 0.5$. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 14: Robustness of the Bullish Bias: Response-Relative Measure (Panel A: Neutral Prompt)

		All posts		Bullish posts		Bearish posts	
Subsample	n	Gap_B	Gap_I	Gap_B	Gap_I	Gap_B	Gap_I
Panel A: Neutral generation prompt							
Full sample	15,206	$+0.095^{***}$	$+0.349^{***}$	$+0.103^{***}$	$+0.352^{***}$	$+0.082^{***}$	$+0.343^{***}$
		(77.35)	(131.59)	(67.68)	(104.98)	(39.60)	(79.36)
Low tercile	4,919	$+0.092^{***}$	$+0.353^{***}$	$+0.094^{***}$	$+0.354^{***}$	$+0.090^{***}$	$+0.351^{***}$
		(43.10)	(75.81)	(31.12)	(55.02)	(29.82)	(52.14)
Mid tercile	5,021	$+0.096^{***}$	$+0.352^{***}$	$+0.102^{***}$	$+0.357^{***}$	$+0.084^{***}$	$+0.343^{***}$
		(45.88)	(76.72)	(40.23)	(61.79)	(23.22)	(45.49)
High tercile	4,614	$+0.099^{***}$	$+0.345^{***}$	$+0.111^{***}$	$+0.350^{***}$	$+0.068^{***}$	$+0.332^{***}$
		(42.76)	(71.56)	(43.29)	(61.74)	(13.69)	(36.22)
1 comment	7,586	$+0.106^{***}$	$+0.403^{***}$	$+0.113^{***}$	$+0.406^{***}$	$+0.094^{***}$	$+0.397^{***}$
		(56.09)	(102.11)	(48.20)	(82.37)	(29.47)	(60.35)
2–5 comments	4,452	$+0.077^{***}$	$+0.321^{***}$	$+0.085^{***}$	$+0.320^{***}$	$+0.063^{***}$	$+0.321^{***}$
		(35.16)	(69.24)	(32.18)	(53.67)	(16.89)	(43.76)
6–20 comments	2,452	$+0.096^{***}$	$+0.269^{***}$	$+0.103^{***}$	$+0.274^{***}$	$+0.085^{***}$	$+0.259^{***}$
		(38.12)	(46.59)	(33.58)	(37.29)	(19.57)	(27.94)
>20 comments	716	$+0.091^{***}$	$+0.224^{***}$	$+0.096^{***}$	$+0.227^{***}$	$+0.082^{***}$	$+0.219^{***}$
		(21.33)	(22.51)	(18.21)	(18.33)	(11.32)	(13.05)
Pre-2020	649	$+0.077^{***}$	$+0.248^{***}$	$+0.082^{***}$	$+0.245^{***}$	$+0.070^{***}$	$+0.251^{***}$
		(15.21)	(21.43)	(12.71)	(16.41)	(8.57)	(13.75)
2020	3,162	$+0.093^{***}$	$+0.274^{***}$	$+0.101^{***}$	$+0.272^{***}$	$+0.083^{***}$	$+0.276^{***}$
		(41.34)	(47.93)	(35.20)	(35.15)	(22.91)	(32.82)
2021	9,037	$+0.104^{***}$	$+0.397^{***}$	$+0.111^{***}$	$+0.397^{***}$	$+0.092^{***}$	$+0.397^{***}$
		(63.86)	(119.02)	(56.21)	(96.28)	(31.88)	(69.96)
2022	1,706	$+0.062^{***}$	$+0.281^{***}$	$+0.072^{***}$	$+0.290^{***}$	$+0.048^{***}$	$+0.268^{***}$
		(14.50)	(33.48)	(13.10)	(26.74)	(7.02)	(20.21)
Small cap	8,058	$+0.096^{***}$	$+0.360^{***}$	$+0.104^{***}$	$+0.364^{***}$	$+0.085^{***}$	$+0.354^{***}$
		(54.07)	(100.03)	(46.45)	(78.72)	(28.97)	(61.73)
Large cap	6,489	$+0.095^{***}$	$+0.337^{***}$	$+0.103^{***}$	$+0.340^{***}$	$+0.080^{***}$	$+0.330^{***}$
		(54.19)	(82.39)	(48.05)	(67.04)	(26.54)	(47.91)
Same-day ($\Delta = 0$)	12,121	$+0.097^{***}$	$+0.349^{***}$	$+0.105^{***}$	$+0.352^{***}$	$+0.084^{***}$	$+0.344^{***}$
		(69.89)	(117.07)	(61.07)	(93.13)	(35.98)	(70.94)
Next-day+ ($\Delta \geq 1$)	2,433	$+0.088^{***}$	$+0.356^{***}$	$+0.095^{***}$	$+0.362^{***}$	$+0.075^{***}$	$+0.346^{***}$
		(30.00)	(55.33)	(26.24)	(44.66)	(15.16)	(32.68)

The bullish bias gap is $\delta^{\text{resp}}_{\text{AI}} - \bar{\delta}^{\text{resp}}_{\text{human}}$, where $\delta^{\text{resp}} = p_A(\text{reply alone}) - p_A(\text{post})$. This measure presents the reply text in isolation to the probe, removing interaction effects between post and reply text in the joint reading. See notes to Table 12 for further details. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. Panel B (professional prompt) continues in Table 15.

Table 15: Robustness of the Bullish Bias: Response-Relative Measure (Panel B: Professional Prompt)

		All posts		Bullish posts		Bearish posts	
Subsample	n	Gap_B	Gap_I	Gap_B	Gap_I	Gap_B	Gap_I
Panel B: Professional investor prompt							
Full sample	15,206	$+0.066^{***}$	$+0.272^{***}$	$+0.074^{***}$	$+0.281^{***}$	$+0.052^{***}$	$+0.257^{***}$
		(53.05)	(107.78)	(48.97)	(88.39)	(24.28)	(61.99)
Low tercile	4,919	$+0.063^{***}$	$+0.273^{***}$	$+0.067^{***}$	$+0.276^{***}$	$+0.059^{***}$	$+0.270^{***}$
		(29.30)	(61.13)	(22.41)	(45.03)	(18.98)	(41.39)
Mid tercile	5,021	$+0.063^{***}$	$+0.272^{***}$	$+0.069^{***}$	$+0.283^{***}$	$+0.053^{***}$	$+0.250^{***}$
		(30.22)	(62.09)	(27.14)	(51.28)	(14.55)	(35.16)
High tercile	4,614	$+0.072^{***}$	$+0.272^{***}$	$+0.085^{***}$	$+0.283^{***}$	$+0.035^{***}$	$+0.241^{***}$
		(30.75)	(59.57)	(33.45)	(53.18)	(7.01)	(27.47)
1 comment	7,586	$+0.074^{***}$	$+0.323^{***}$	$+0.081^{***}$	$+0.332^{***}$	$+0.061^{***}$	$+0.307^{***}$
		(38.64)	(84.89)	(34.57)	(70.09)	(18.68)	(48.15)
2–5 comments	4,452	$+0.047^{***}$	$+0.243^{***}$	$+0.058^{***}$	$+0.251^{***}$	$+0.031^{***}$	$+0.231^{***}$
		(21.73)	(54.64)	(21.84)	(44.16)	(8.28)	(32.31)
6–20 comments	2,452	$+0.071^{***}$	$+0.199^{***}$	$+0.077^{***}$	$+0.209^{***}$	$+0.060^{***}$	$+0.183^{***}$
		(28.68)	(37.87)	(26.26)	(31.28)	(13.82)	(21.52)
>20 comments	716	$+0.071^{***}$	$+0.169^{***}$	$+0.081^{***}$	$+0.169^{***}$	$+0.053^{***}$	$+0.167^{***}$
		(16.83)	(18.82)	(16.26)	(15.04)	(7.00)	(11.30)
Pre-2020	649	$+0.052^{***}$	$+0.199^{***}$	$+0.057^{***}$	$+0.211^{***}$	$+0.045^{***}$	$+0.182^{***}$
		(10.87)	(18.60)	(9.42)	(15.12)	(5.76)	(10.91)
2020	3,162	$+0.070^{***}$	$+0.211^{***}$	$+0.076^{***}$	$+0.214^{***}$	$+0.060^{***}$	$+0.208^{***}$
		(30.77)	(39.22)	(26.87)	(29.72)	(16.39)	(25.60)
2021	9,037	$+0.071^{***}$	$+0.310^{***}$	$+0.079^{***}$	$+0.316^{***}$	$+0.056^{***}$	$+0.297^{***}$
		(43.17)	(95.80)	(40.41)	(79.66)	(18.80)	(53.47)
2022	1,706	$+0.036^{***}$	$+0.213^{***}$	$+0.048^{***}$	$+0.225^{***}$	$+0.019^{***}$	$+0.196^{***}$
		(8.41)	(26.98)	(8.70)	(21.86)	(2.82)	(15.92)
Small cap	8,058	$+0.066^{***}$	$+0.283^{***}$	$+0.075^{***}$	$+0.292^{***}$	$+0.053^{***}$	$+0.268^{***}$
		(37.14)	(82.00)	(33.93)	(66.45)	(17.69)	(48.35)
Large cap	6,489	$+0.065^{***}$	$+0.259^{***}$	$+0.073^{***}$	$+0.268^{***}$	$+0.050^{***}$	$+0.241^{***}$
		(37.00)	(66.75)	(34.47)	(55.68)	(16.15)	(36.99)
Same-day ($\Delta = 0$)	12,121	$+0.068^{***}$	$+0.270^{***}$	$+0.076^{***}$	$+0.279^{***}$	$+0.054^{***}$	$+0.256^{***}$
		(48.63)	(95.21)	(44.94)	(77.98)	(22.38)	(54.93)
Next-day+ ($\Delta \geq 1$)	2,433	$+0.056^{***}$	$+0.281^{***}$	$+0.065^{***}$	$+0.292^{***}$	$+0.041^{***}$	$+0.262^{***}$
		(18.87)	(45.76)	(17.52)	(37.55)	(8.22)	(26.24)

Continuation of Table 14. The bullish bias gap is $\delta^{\text{resp}}_{\text{AI}} - \bar{\delta}^{\text{resp}}_{\text{human}}$, where $\delta^{\text{resp}} = p_A(\text{reply alone}) - p_A(\text{post})$. This measure presents the reply text in isolation to the probe, removing interaction effects between post and reply text in the joint reading. See notes to Table 12 for further details. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 16: Robustness of the Bullish Bias Across Generation Temperatures

		All posts		Bullish posts		Bearish posts	
Temperature	n	Gap_B	Gap_I	Gap_B	Gap_I	Gap_B	Gap_I
Panel A: Neutral generation prompt							
$T = 0.0$ (deterministic)	15,206	$+0.096^{***}$	$+0.264^{***}$	$+0.101^{***}$	$+0.262^{***}$	$+0.088^{***}$	$+0.267^{***}$
		(57.90)	(117.70)	(49.43)	(92.50)	(31.07)	(72.81)
$T = 0.5$	15,206	$+0.103^{***}$	$+0.265^{***}$	$+0.108^{***}$	$+0.261^{***}$	$+0.093^{***}$	$+0.272^{***}$
		(71.89)	(127.97)	(61.66)	(99.73)	(38.27)	(80.30)
$T = 1.0$	15,206	$+0.124^{***}$	$+0.265^{***}$	$+0.127^{***}$	$+0.263^{***}$	$+0.119^{***}$	$+0.268^{***}$
		(88.97)	(129.65)	(73.97)	(102.36)	(50.12)	(79.58)
Panel B: Professional investor prompt							
$T = 0.0$ (deterministic)	15,206	$+0.066^{***}$	$+0.223^{***}$	$+0.071^{***}$	$+0.226^{***}$	$+0.059^{***}$	$+0.217^{***}$
		(38.91)	(97.13)	(33.77)	(78.77)	(20.23)	(56.96)
$T = 0.5$	15,206	$+0.074^{***}$	$+0.228^{***}$	$+0.077^{***}$	$+0.227^{***}$	$+0.068^{***}$	$+0.230^{***}$
		(51.71)	(110.87)	(44.07)	(87.91)	(27.87)	(67.56)
$T = 1.0$	15,206	$+0.095^{***}$	$+0.232^{***}$	$+0.099^{***}$	$+0.233^{***}$	$+0.089^{***}$	$+0.230^{***}$
		(67.86)	(115.16)	(57.34)	(92.67)	(37.16)	(68.47)

Gap_B and Gap_I are the AI-minus-human bullish bias gaps ($\delta^{\text{joint}}_{\text{AI}} - \bar{\delta}^{\text{joint}}_{\text{human}}$) under the base and instruct probe models. t-statistics from paired t-tests are in parentheses. $T = 0.0$ uses deterministic greedy decoding; the bias at this temperature is present in the model's modal reply, rather than arising from random sampling noise. Post direction defined using the base-probe post score. $n = 15{,}206$ posts in all rows. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

A.6 Validation with an Independent Model Family

We validate the bullish bias finding using an independent validation dataset in which replies are generated by GPT-4o-mini (OpenAI), a model from a different architecture family and training pipeline than the Mistral models used in the main analysis. The validation sample consists of 2,000 posts drawn from the same WSB post universe, with the same depth-1 human comment benchmark. Replies were generated at three temperatures (0.0, 0.5, 1.0) with two replications each, using the same neutral generation prompt as in the main analysis. The GPT-4o-mini model is used as both generator and probe in this dataset; the absence of a matched untuned base model and the use of a single probe model limit the comparability with the main analysis, but the exercise allows us to ask whether the core bullish bias pattern replicates outside the Mistral model family.

Table 17 reports the results. The AI-human bullish gap for GPT-4o-mini is +0.419 ($t = 42.34$), positive and significant at the one-percent level. Critically, the gap is positive and significant for both bullish posts ($+0.563^{***}$, $t = 48.95$) and bearish posts ($+0.172^{***}$, $t = 12.09$), directly replicating the pattern that distinguishes a bullish bias from mere agreement with the investor's original thesis. The gap is also robust to generation temperature: at temperature 0.0 (deterministic greedy decoding), the gap is $+0.421^{***}$, confirming that the bias is present in the model's modal reply rather than arising from stochastic sampling. These results suggest that the bullish bias is not specific to the Mistral instruction-tuning pipeline but extends to at least one other widely-used instruction-tuned model.

Table 17: Validation of the Bullish Bias: GPT-4o-mini

		All posts		Bullish posts		Bearish posts	
Specification	n	Gap	t	Gap	t	Gap	t
Panel A: By generation temperature							
$T = 0.0$ (deterministic)	2,000	$+0.421^{***}$	(40.84)	$+0.564^{***}$	(46.58)	$+0.177^{***}$	(11.73)
$T = 0.5$	2,000	$+0.416^{***}$	(40.85)	$+0.561^{***}$	(46.35)	$+0.166^{***}$	(11.08)
$T = 1.0$	2,000	$+0.421^{***}$	(41.34)	$+0.565^{***}$	(46.79)	$+0.174^{***}$	(11.59)
Panel B: Averaged across temperatures							
Full sample	2,000	$+0.419^{***}$	(42.34)	$+0.563^{***}$	(48.95)	$+0.172^{***}$	(12.09)

The bullish bias gap is $\delta^{\text{joint}}_{\text{AI}} - \bar{\delta}^{\text{joint}}_{\text{human}}$, where $\delta^{\text{joint}} = p_A(\text{post} + \text{reply}) - p_A(\text{post})$. The generator is GPT-4o-mini; the probe is also GPT-4o-mini (the same model family used for both generation and probing, in contrast to the main analysis where Mistral base and instruct probes are applied to Mistral-generated text). Post direction is defined using the GPT-4o-mini probe post score: bullish posts have $p_A(\text{post}) > 0.5$; bearish posts have $p_A(\text{post}) \leq 0.5$. The human benchmark is the equal-weight average of $p_A(\text{post} + \text{comment}_k)$ across all depth-1 comments per post, computed using the GPT-4o-mini probe. t-statistics from paired t-tests are in parentheses. $n = 2{,}000$ posts; mean depth-1 comments per post = 5.9. Only a neutral generation prompt was available for this validation. *** $p < 0.01$.

Table 18: Categorical Word Usage in AI-Generated Replies

Generator	Prompt	Rate (per 100 words)	n
Panel A: Mean categorical word rate			
Base	Neutral	0.103	28,922
Base	Professional	0.082	28,922
Instruct	Neutral	0.300	28,922
Instruct	Professional	0.252	28,922
Panel B: Instruct minus base (paired difference)			
	Neutral	$+0.197^{***}$	($t = 70.55$)
	Professional	$+0.170^{***}$	($t = 69.97$)

Panel A reports the mean categorical word rate across all 28,922 posts in the post universe, averaged across three generation temperatures and two replications. Panel B reports the paired difference between the instruction-tuned and base generators for the same posts; t-statistics from paired t-tests are in parentheses.

Categorical words are occurrences of the following terms per 100 words of reply text: *clearly, definitely, certainly, must, strongly, conviction, undoubtedly, obviously, firmly, decisive, bottom line*. Multi-word phrases ("bottom line") are matched as exact substrings; all other terms are matched as whole words.

The instruction-tuned model uses significantly more categorical and decisive language than the base model under both prompt conditions, and the difference is significant at the 1% level. *** $p < 0.01$.

Table 19: Reply Length of AI-Generated Replies

Generator	Prompt	Mean words	n
Panel A: Mean word count			
Base	Neutral	88.8	28,922
Base	Professional	88.2	28,922
Instruct	Neutral	66.6	28,922
Instruct	Professional	75.2	28,922
Panel B: Instruct minus base (paired difference)			
	Neutral	-22.2^{***}	$(t = -214.40)$
	Professional	-13.0^{***}	$(t = -141.25)$

Panel A reports the mean reply word count across all 28,922 posts in the post universe, averaged across three generation temperatures and two replications. Replies were capped at 120 new tokens during generation. Panel B reports the paired difference between the instruction-tuned and base generators for the same posts; t-statistics from paired t-tests are in parentheses. The instruction-tuned model produces significantly shorter replies than the base model under both prompt conditions. The professional investor prompt narrows the gap: instruct replies are 22.2 words shorter than base under the neutral prompt but only 13.0 words shorter under the professional prompt, suggesting that role conditioning induces the instruction-tuned model to provide more elaborated responses. *** $p < 0.01$.

B Proofs for Section 3.6

Throughout, maintain Assumption 1. For any realized post orientation r, define

$$\rho(r; \tau, p_0) = (1 - \tau)p_0 + \tau r.$$

The assistant solves

$$\max_{m \in [0,1]} \left\{ -\mathbb{E}\big[(m - E)^2 \mid r\big] + \gamma\big(m - \tfrac{1}{2}\big)^2 \right\}.$$

Because $E \in \{0, 1\}$ and the assistant's latent assessment is $\rho(r; \tau, p_0)$,

$$\begin{aligned} \mathbb{E}\big[(m - E)^2 \mid r\big] &= \rho(r; \tau, p_0)(1 - m)^2 + \big(1 - \rho(r; \tau, p_0)\big)m^2 \\ &= \big(m - \rho(r; \tau, p_0)\big)^2 + \rho(r; \tau, p_0)\big(1 - \rho(r; \tau, p_0)\big). \end{aligned} \tag{22}$$

Hence, up to an additive constant independent of m, the objective is

$$-(m-\rho(r;\tau,p_0))^2+\gamma\big(m-\tfrac{1}{2}\big)^2.$$

Differentiating with respect to m gives

$$-2\big(m-\rho(r;\tau,p_0)\big)+2\gamma\big(m-\tfrac{1}{2}\big)=0.$$

Rearranging,

$$(1-\gamma)m=\rho(r;\tau,p_0)-\frac{\gamma}{2}.$$

Thus the unique interior candidate is

$$m(r)=\frac{\rho(r;\tau,p_0)-\gamma/2}{1-\gamma}. \tag{23}$$

The second derivative of the objective is

$$-2+2\gamma=-2(1-\gamma)<0,$$

because $\gamma\in[0,1)$. Therefore the problem is strictly concave, and the candidate is the unique optimizer. Assumption 1 guarantees that this optimizer lies in $(0,1)$ on the support under consideration.

Proof of Proposition 1

The first statement follows directly from (23) and the definition

$$\rho(r;\tau,p_0)=(1-\tau)p_0+\tau r.$$

Substituting this expression into (23) gives

$$m(r)=\frac{(1-\tau)p_0+\tau r-\gamma/2}{1-\gamma}.$$

For the midpoint identity, subtract 1/2:

$$m(r)-\frac{1}{2}=\frac{\rho(r;\tau,p_0)-\gamma/2}{1-\gamma}-\frac{1}{2}.$$

Putting the terms over the common denominator $1-\gamma$,

$$m(r)-\frac{1}{2}=\frac{\rho(r;\tau,p_0)-\gamma/2-(1-\gamma)/2}{1-\gamma}=\frac{\rho(r;\tau,p_0)-1/2}{1-\gamma}.$$

This proves (13). □

Proof of Proposition 2

For part (i), set $\gamma=0$ in the reply rule:

$$m(r)=\rho(r;\tau,p_0)=(1-\tau)p_0+\tau r.$$

Subtracting r gives

$$m(r)-r=(1-\tau)p_0+\tau r-r=(1-\tau)(p_0-r).$$

Thus the reply is pulled toward p_0: it lies above the post when $p_0>r$ and below the post when $p_0<r$.

For part (ii), use the midpoint identity from Proposition 1:

$$m(r)-\frac{1}{2}=\frac{\rho(r;\tau,p_0)-1/2}{1-\gamma}.$$

Since $1-\gamma>0$, the sign of $m(r)-1/2$ is the same as the sign of $\rho(r;\tau,p_0)-1/2$. Therefore

$$m(r)>\frac{1}{2}\iff\rho(r;\tau,p_0)>\frac{1}{2}\iff(1-\tau)p_0+\tau r>\frac{1}{2}.$$

This proves (14).

For part (iii), start from the closed-form reply rule:

$$m(r)-r=\frac{(1-\tau)p_0+\tau r-\gamma/2}{1-\gamma}-r.$$

Multiplying by $1-\gamma>0$,

$$\begin{aligned}(1-\gamma)\big(m(r)-r\big)&=(1-\tau)p_0+\tau r-\frac{\gamma}{2}-(1-\gamma)r\\&=(1-\tau)p_0+\tau r-r+\gamma r-\frac{\gamma}{2}\\&=(1-\tau)(p_0-r)+\gamma\left(r-\frac{1}{2}\right).\end{aligned}$$

Since $1-\gamma > 0$, $m(r) > r$ if and only if

$$(1-\tau)(p_0 - r) + \gamma\left(r - \frac{1}{2}\right) > 0.$$

This proves (15). □

Proof of Corollary 3.6

By Proposition 2, a post r is shifted upward if and only if

$$(1-\tau)(p_0 - r) + \gamma\left(r - \frac{1}{2}\right) > 0.$$

For the bearish post $r_L < 1/2$, this condition becomes

$$(1-\tau)(p_0 - r_L) - \gamma\left(\frac{1}{2} - r_L\right) > 0,$$

or equivalently

$$(1-\tau)(p_0 - r_L) > \gamma\left(\frac{1}{2} - r_L\right).$$

This proves (16).

For the bullish post $r_H > 1/2$, the same upward-shift condition gives

$$(1-\tau)(p_0 - r_H) + \gamma\left(r_H - \frac{1}{2}\right) > 0,$$

which is (17). Therefore both the bearish and bullish posts are shifted upward if and only if both inequalities hold. □

Proof of Proposition 3

Using the closed-form reply rule,

$$m(r) = \frac{(1-\tau)p_0 + \tau r - \gamma/2}{1-\gamma}.$$

Differentiating with respect to p_0 gives

$$\frac{\partial m(r)}{\partial p_0} = \frac{1-\tau}{1-\gamma}.$$

Since $\tau \in [0,1]$ and $\gamma \in [0,1)$, this derivative is nonnegative, and it is strictly positive

whenever $\tau < 1$. This proves (18).

Differentiating with respect to τ gives

$$\frac{\partial m(r)}{\partial \tau} = \frac{-p_0 + r}{1-\gamma} = \frac{r - p_0}{1-\gamma}.$$

This proves (19). Its sign depends on whether r lies above or below p_0.

Finally, from the midpoint identity,

$$\left| m(r) - \frac{1}{2} \right| = \frac{\left| \rho(r;\tau,p_0) - \frac{1}{2} \right|}{1-\gamma}.$$

Whenever $\rho(r;\tau,p_0) \neq 1/2$, differentiating with respect to γ yields

$$\frac{\partial}{\partial \gamma} \left| m(r) - \frac{1}{2} \right| = \frac{\left| \rho(r;\tau,p_0) - \frac{1}{2} \right|}{(1-\gamma)^2} > 0.$$

This proves (20). Therefore, lowering γ reduces the distance of the reply from the midpoint whenever the latent assessment is away from the midpoint. □

C Market Appendix

This appendix contains the full regression tables for the signal-content analysis of Section 4. All tables use the instruct probe unless otherwise noted. The dependent variable throughout is the h-day-ahead return, winsorized at the 1st and 99th percentiles in the main specification. Coefficients are multiplied by 100 and interpreted as percentage-point returns per unit increase in the probe score $p_A \in [0, 1]$. Standard errors are double-clustered by stock and date in all specifications except where noted.

Table 20: Return Predictability of Reddit Signals: Fixed Effects Specifications

	Post alone	Human	Inst. neutral	Base neutral	Inst. prof.	Base prof.
Panel A: No fixed effects						
$\hat{\beta} \times 100$	+0.046	+0.486	-0.025	-0.063	-0.048	-0.113
[t-stat]	[0.02]	[0.30]	[-0.01]	[-0.03]	[-0.02]	[-0.06]
N	27,503	14,554	27,503	27,503	27,503	27,503
Panel B: Date fixed effects only						
$\hat{\beta} \times 100$	+0.453	+0.576	+1.109**	+0.306	+1.090*	+0.195
[t-stat]	[1.32]	[0.58]	[2.36]	[0.32]	[1.71]	[0.13]
N	27,503	14,554	27,503	27,503	27,503	27,503
Panel C: Stock fixed effects only						
$\hat{\beta} \times 100$	+1.109	+0.124	+2.014	+1.725*	+1.672	+1.044
[t-stat]	[0.99]	[0.33]	[1.24]	[1.76]	[1.05]	[1.06]
N	27,503	14,554	27,503	27,503	27,503	27,503
Panel D: Stock and date fixed effects (main specification)						
$\hat{\beta} \times 100$	+0.749*	+1.097	+1.639**	+0.737	+1.671**	+0.670
[t-stat]	[1.88]	[1.00]	[2.42]	[1.48]	[2.00]	[0.97]
N	27,503	14,554	27,503	27,503	27,503	27,503
Panel E: Stock and date FE + controls (log mcap, post confidence)						
$\hat{\beta} \times 100$	+0.783**	—	+1.615**	—	+1.601*	—
[t-stat]	[2.17]	—	[2.42]	—	[1.88]	—
N	27,493	—	27,493	—	27,493	—

Dependent variable: one-day-ahead return, winsorized at the 1st and 99th percentiles. Signals are probe scores $p_A \in [0, 1]$ constructed from the instruct probe. $\hat{\beta} \times 100$ is the percentage-point return associated with a unit increase in the probe score. Standard errors are double-clustered by stock and date. Controls in Panel E are log market capitalization and post inner confidence $|p_A^{\text{post}} - 0.5|$. *, **, *** denote 10%, 5%, and 1% significance.

Table 21: Return Predictability Across Horizons

Signal	1-day	2-day	3-day	5-day	10-day
$\hat{\beta} \times 100$					
Post alone	+0.749*	+1.269***	+1.228**	+1.314	+0.500
Instruct neutral	+1.639**	+2.442***	+2.414**	+3.054*	+0.607
Base neutral	+0.737	+2.178***	+2.152*	+2.082	−1.413
Instruct prof.	+1.671**	+2.782***	+3.497**	+4.026**	+0.496
Base prof.	+0.670	+2.467**	+2.888**	+3.959**	+1.084
[t-statistic]					
Post alone	[1.88]	[2.92]	[2.03]	[1.48]	[0.56]
Instruct neutral	[2.42]	[2.93]	[1.98]	[1.73]	[0.53]
Base neutral	[1.48]	[2.70]	[1.83]	[1.44]	[-0.94]
Instruct prof.	[2.00]	[2.61]	[2.10]	[2.04]	[0.45]
Base prof.	[0.97]	[2.50]	[2.13]	[2.30]	[0.85]
Observations					
All signals	27,503	27,497	27,491	27,479	27,448

Dependent variable: h-day-ahead return, winsorized at the 1st and 99th percentiles. All specifications include stock and date fixed effects. Standard errors are double-clustered by stock and date. *, **, *** denote 10%, 5%, and 1% significance.

Table 22: Return Predictability: Subsamples and Controls

	Post alone		Instruct neutral		Instruct prof.	
Subsample	$\hat{\beta}\times100$	$[t]$	$\hat{\beta}\times100$	$[t]$	$\hat{\beta}\times100$	$[t]$
Full sample ($N = 27{,}503$)	+0.749*	[1.88]	+1.639**	[2.42]	+1.671**	[2.00]
Market capitalization						
Small cap ($N = 13{,}747$)	−0.229	[-0.89]	−0.169	[-0.42]	−0.302	[-0.82]
Large cap ($N = 13{,}746$)	+1.297***	[2.59]	+2.718***	[3.36]	+2.847***	[2.76]
Post confidence $\lvert p_A^{post} - 0.5\rvert$						
Low tercile ($N = 9{,}090$)	+2.456*	[1.83]	+2.323*	[1.68]	+1.485	[0.92]
Mid tercile ($N = 9{,}072$)	+0.310	[0.59]	+1.482*	[1.90]	+2.393***	[2.59]
High tercile ($N = 9{,}072$)	+1.573**	[2.46]	+2.325**	[2.08]	+2.005*	[1.92]
Sample period						
Pre-2020 ($N = 933$)	−0.397	[-0.76]	−0.286	[-0.37]	−0.410	[-0.44]
2020–2022 ($N = 26{,}570$)	+0.786*	[1.95]	+1.703**	[2.49]	+1.725**	[2.03]
Post timing						
Same-day matched	+0.893**	[1.97]	+1.893**	[2.54]	+1.789**	[1.97]
Next-day+ matched	+0.786*	[1.95]	−0.791	[-0.42]	+0.072	[0.03]
With controls						
Full + controls ($N = 27{,}493$)	+0.783**	[2.17]	+1.615**	[2.42]	+1.601*	[1.88]

Dependent variable: one-day-ahead return, winsorized at the 1st and 99th percentiles. All specifications include stock and date fixed effects with standard errors double-clustered by stock and date. Observation counts (N) are shown in the row label; timing-split counts are not separately reported. Controls are log market capitalization and post inner confidence $\lvert p_A^{\text{post}} - 0.5\rvert$. *, **, *** denote 10%, 5%, and 1% significance.

Table 23: Robustness Scorecard: Instruct Signal Across Specification Tests

Block	Specification	$\hat{\beta}_n \times 100$	$[t_n]$	$\hat{\beta}_p \times 100$	$[t_p]$
Fixed effects specification					
	Date FE only	+1.109**	[2.36]	+1.090*	[1.71]
	Stock and Date FE (main)	+1.639**	[2.42]	+1.671**	[2.00]
	Two-way FE + controls	+1.615**	[2.42]	+1.601*	[1.88]
Sample cuts (stock and date FE)					
	Full sample	+1.639**	[2.42]	+1.671**	[2.00]
	Low confidence	+2.323*	[1.68]	+1.485	[0.92]
	Mid confidence	+1.482*	[1.90]	+2.393***	[2.59]
	High confidence	+2.325**	[2.08]	+2.005*	[1.92]
	6–20 comments	−2.622**	[-2.03]	−0.975	[-1.06]
	Pre-2020	−0.286	[-0.37]	−0.410	[-0.44]
	2020	+0.889*	[1.69]	+1.437**	[2.16]
	2021	+2.169**	[2.49]	+2.095**	[2.00]
	2022	+1.553*	[1.67]	+1.168	[1.07]
	Small cap	−0.169	[-0.42]	−0.302	[-0.82]
	Large cap	+2.718***	[3.36]	+2.847***	[2.76]
	Same-day (gap = 0)	+1.893**	[2.54]	+1.789**	[1.97]
	Next-day+ (gap $\geq$ 1)	−0.791	[-0.42]	+0.072	[0.03]
Return definition					
	Raw (no winsorization)	+1.639**	[2.42]	+1.671**	[2.00]
	Winsorized 1% (main)	+1.639**	[2.42]	+1.671**	[2.00]
	Winsorized 5%	+0.866**	[2.43]	+0.624*	[1.67]
Clustering					
	Double: stock and date (main)	+1.639**	[2.42]	+1.671**	[2.00]
	Stock only	+1.639***	[2.60]	+1.671**	[2.11]
	Date only	+1.639***	[2.90]	+1.671**	[2.57]
Signal scaling					
	Raw $p_A \in [0, 1]$ (main)	+1.639**	[2.42]	+1.671**	[2.00]
	z-score standardized	+0.352**	[2.42]	+0.343**	[2.00]
Sample exclusions					
	Drop top-5% by comment count	+1.593**	[2.41]	+1.683**	[2.07]
	Drop 2021	+0.985**	[2.50]	+1.153**	[2.36]
	Drop 2020	+1.913**	[2.36]	+1.832*	[1.90]
	Drop pre-2020	+1.703**	[2.49]	+1.725**	[2.03]
	Posts with $\geq$ 1 human reply	+1.042**	[2.57]	+0.663	[1.35]

Each row reports univariate coefficients from regression (21) under the stated specification. Subscript n denotes the instruct neutral signal; subscript p denotes the instruct professional-prompt signal. The main specification is stock and date fixed effects with returns winsorized at the 1st and 99th percentiles and standard errors double-clustered by stock and date. The no-fixed-effects and stock-only fixed-effects specifications are not included in this scorecard because neither absorbs the market-wide variation that is the primary confounder in this setting; they are reported in Appendix Table 20. The Newey-West post-level specification is omitted because daily aggregation to 1,082 dates eliminates cross-sectional variation, making the test not comparable to the post-level panel. *, **, *** denote 10%, 5%, and 1% significance.

Table 24: Incremental Return Association: Instruct and Base Signals Jointly

	1-day	2-day	3-day	5-day	10-day
Panel A: Instruct neutral vs. base neutral					
$\hat{\beta}_{\text{inst}} \times 100$	+1.825**	+1.991**	+1.968	+2.922*	+1.813
[t-stat]	[2.50]	[2.56]	[1.62]	[1.65]	[1.06]
$\hat{\beta}_{\text{base}} \times 100$	−0.390	+0.948	+0.936	+0.277	−2.533
[t-stat]	[-0.90]	[1.50]	[0.90]	[0.25]	[-1.18]
N	27,503	27,497	27,491	27,479	27,448
Within R^2	0.00048	0.00083	0.00046	0.00042	0.00013
Panel B: Instruct professional vs. base neutral					
$\hat{\beta}_{\text{inst}} \times 100$	+1.781*	+2.386**	+3.357**	+4.105**	+1.524
[t-stat]	[1.89]	[2.22]	[2.00]	[2.05]	[0.94]
$\hat{\beta}_{\text{base}} \times 100$	−0.241	+0.867	+0.308	−0.172	−2.250
[t-stat]	[-0.47]	[1.34]	[0.36]	[-0.15]	[-1.10]
N	27,503	27,497	27,491	27,479	27,448
Within R^2	0.00045	0.00098	0.00082	0.00067	0.00011
Panel C: Instruct professional vs. base professional					
$\hat{\beta}_{\text{inst}} \times 100$	+1.819*	+2.243**	+2.944*	+3.021*	+0.026
[t-stat]	[1.91]	[2.29]	[1.86]	[1.65]	[0.02]
$\hat{\beta}_{\text{base}} \times 100$	−0.337	+1.226	+1.259	+2.287*	+1.070
[t-stat]	[-0.43]	[1.61]	[1.31]	[1.80]	[0.68]
N	27,503	27,497	27,491	27,479	27,448
Within R^2	0.00046	0.00103	0.00088	0.00080	0.00003

Regression: $r_{i,t+h} = \alpha_i + \gamma_t + \beta_{\text{inst}}\, s^{\text{inst}}_{i,t} + \beta_{\text{base}}\, s^{\text{base}}_{i,t} + \varepsilon_{i,t}$, estimated under stock and date fixed effects with standard errors double-clustered by stock and date. Because the base and instruction-tuned generators share architecture and pretraining corpus, including the base signal controls for return-relevant information common to both generated replies. The instruction-tuned coefficient is interpreted as the incremental return association beyond the matched base signal. This specification does not eliminate all lookahead or memorization concerns, but reduces the concern that the return association is driven solely by shared pretraining. *, **, *** denote 10%, 5%, and 1% significance.

Table 25: Bias Reduction and Horizon Return Association: Professional vs. Neutral Prompt

	2-day		3-day		5-day		10-day	
	$\hat{\beta}\times100$	$[t]$	$\hat{\beta}\times100$	$[t]$	$\hat{\beta}\times100$	$[t]$	$\hat{\beta}\times100$	$[t]$
Panel A: Joint regression								
Neutral ($\hat{\beta}_n$)	+0.964	[1.11]	−0.339	[-0.24]	+0.261	[0.20]	+0.577	[0.36]
Professional ($\hat{\beta}_p$)	+2.022	[1.55]	+3.764*	[1.67]	+3.821**	[2.11]	+0.041	[0.03]
$\Delta\hat{\beta} = \hat{\beta}_p - \hat{\beta}_n$	+1.058	[0.54]	+4.103	[1.20]	+3.560	[1.50]	−0.536	[-0.19]
Panel B: Difference-signal regression $\Delta s = s_{prof} - s_{neutral}$								
$\hat{\beta}_n$ (neutral signal)	+2.986***	[2.82]	+3.425**	[2.12]	+4.081*	[1.94]	+0.618	[0.51]
$\hat{\gamma}_{\Delta s}$ (professional adjustment)	+2.022	[1.55]	+3.764*	[1.67]	+3.821**	[2.11]	+0.041	[0.03]
N	27,497		27,491		27,479		27,448	
Panel C: Bias reduction confirmation (comparable sample, $N = 14{,}554$)								
Mean bias, neutral	+0.2663 probe-score units							
Mean bias, professional	+0.2279 probe-score units							
Mean reduction (neutral − professional)	+0.0385 (59.4% of posts; paired $t = 33.62$, $p < 0.001$)							

Panel A: joint regression $r_{i,t+h} = \alpha_i + \gamma_t + \beta_n s^n_{i,t} + \beta_p s^p_{i,t} + \varepsilon_{i,t}$, with $\Delta\hat{\beta}$ tested via a Wald test using the double-clustered covariance matrix. Panel B: regression of $r_{i,t+h}$ on the neutral signal and the difference signal $\Delta s = s_{\text{prof}} - s_{\text{neutral}}$; Panel B regresses returns on the neutral signal and the professional-prompt adjustment $\Delta s = s_{\text{prof}} - s_{\text{neutral}}$. The coefficient on Δs measures whether the change induced by the professional prompt contains return-relevant variation beyond the neutral signal. Because the professional prompt often lowers the signal, this coefficient should not be read mechanically as the return effect of bias reduction without considering the sign and distribution of Δs. Panel C: bias is defined as the signal minus the average human-reply probe score on the comparable subsample of posts with at least one human comment. The paired t-test tests H_0: $\mu_{\text{prof}} = \mu_{\text{neutral}}$. All specifications: stock and date fixed effects, standard errors double-clustered by stock and date. *, **, *** denote 10%, 5%, and 1% significance.