

How LLM Financial Agents Respond to Stated Affect Depends on Their Assigned Role

Da-Chen Lian^{1,2}, Yu-Lin Chang², Yen-Hsi Huang³ and Shu-Kai Hsieh^{1,2}

¹Co-Intelligence Humanities AI Future Lab

²Graduate Institute of Linguistics, National Taiwan University

³Department of Information Management, National Taiwan University

{d08944019, shukaihsieh}@ntu.edu.tw

cylchang37@gmail.com, lynn903221@gmail.com

Abstract

We test whether financial LLM agents respond differently to reports with identical facts, forecasts and recommendations but different wording. The reports differ only in a closing sentence stating the sender's affect. The most consistent finding is linguistic: high-arousal wording increases readers' word-based arousal scores on average and also changes teammates' language. The response varies across role-specific task instructions, although its magnitude and ordering change across prompt paraphrases, scoring methods, word-removal controls and model families. The central role comparison contains 7,080 generations, but all concern one underlying deposit-outflow event; the 59 matched pairs are seeded realizations of the same financial payload rather than independent financial situations. Separately varying alarm and urgency shows that both raise word-based arousal, while alarmed wording adds an apprehension component detected by an emotion classifier. Behavioral effects are less consistent: in the primary model, alarmed wording reduces adherence to BUY and SELL recommendations, but the SELL reported-size effect is not detected in two additional model families. A filter that removes explicit alarm while preserving checked financial fields leaves most of the linguistic response intact.

1 Introduction

Language models read one another's messages and act on what they receive. Checking financial facts alone may therefore miss an effect of wording: two reports can give the same forecast, confidence, horizon, recommendation and position size while expressing different levels of alarm. We test whether such wording changes agents' language and reported decisions.

We render the same financial report twice, changing only a self-referential closing sentence in which the sender states its affect (§3.2). The specified

financial payload is held fixed across the two renderings. In isolated-reader tests, prices are also identical across conditions; in team experiments, we separately check whether the treated agent's trade changes the price later seen by teammates.

Each reader receives a one-sentence task instruction corresponding to its assigned financial role. These instructions, rather than the readers' own role labels, are what differ in our central role comparison. In the primary model, responses vary across them: some readers use higher-arousal language after the alarmed report while another shows a decrease. Averaging across instructions can therefore hide substantial heterogeneity.

The central comparison is broad in generated outputs but narrow in financial scope. Its 7,080 generations all derive from one deposit-outflow event. The 59 matched pairs are seeded runs of the same financial payload, each supplying a different team context, not 59 independent financial events. We therefore interpret the experiment as evidence of instruction-specific heterogeneity within one event rather than a cross-event ranking of financial roles.

Our evidence also differs in strength across outcomes. The strongest and most reproducible result concerns language: stated affect changes the register of direct readers and, under controlled team designs, their teammates. Reported financial decisions also change in several tests, but those behavioral effects are less stable across model families.

We make four contributions.

1. A method for varying stated affect while holding a specified financial payload fixed, with automated checks before generation. We release the renderer, checks, plans and generations behind §4.2 and §4.3.¹
2. Evidence that replacing one delivered message changes teammates' language across four

¹<https://github.com/lopentu/affective-contagion-financial-agents>

synthetic financial events. Prices match on three events; on the BUY event, some prices differ, so its result may also reflect a price response (§4.1). Separately varying affect and urgency shows that both raise word-based arousal, while a classifier detects additional apprehension from alarmed wording.

3. Response heterogeneity across role-specific task instructions within one deposit-outflow event, recurring under paraphrased instructions, alternative scoring, word-removal controls and two additional model families. The magnitude and ordering change across checks, so we do not identify a consistently sensitive financial role.
4. More limited evidence that alarmed wording changes reported decisions in the primary model. Readers follow BUY and SELL recommendations less closely, but the SELL reported-size effect does not replicate in two additional model families. A filter preserving checked financial fields also leaves most of the linguistic response intact.

2 Related work

Affect in single models Emotional prompts change task performance (Li et al., 2025), questionnaire scores and downstream biases (Coda-Forno et al., 2024). Sustained affective interaction also alters open-ended output across model families (Xu et al., 2026). Persuasive misinformation can shift stated beliefs and defeat truthful answers in agent debate (Xu et al., 2024; Agarwal and Khanna, 2025).

Affect between agents Durupinar (2026) reports affect propagation in an LLM-driven crowd, while Qiang (2025) finds that simulated discussion networks differ from Reddit networks in structure and emotional trajectories. Related work shows harmful traits or messages spreading through agent systems (Zhang et al., 2024; Lee et al., 2026; Ju et al., 2026), and role assignments or personas can change susceptibility to manipulation (Cao et al., 2026; Costa et al., 2026).

Closest to our design, Liu (2026) measures evaluator preferences propagating between agents, with effects depending on topology and committee size. That treatment changes the evaluator’s disposition without separating message form from content. Outside multi-agent systems, Chun and Elkins

(2026) vary emotional framing of fixed scenarios across finance, healthcare and law and find little decision change. Our comparison instead asks whether financial wording changes another agent’s response while the specified financial payload is fixed.

Financial agents Our six roles follow Xiao et al. (2024); financial-agent systems more broadly are surveyed by Dong et al. (2025). Prior experiments study simulated markets (Yang et al., 2025), trading responses to information (Zhang et al., 2026), and attempts to influence groups through emotional language (Koch et al., 2025). We focus on internal communication between agents.

Financial sentiment Financial sentiment research extracts tone from documents (Malo et al., 2014; Araci, 2019) or relates it to returns (Lopez-Lira and Tang, 2026). We instead manipulate stated affect while holding the specified financial payload fixed.

Emotional contagion in people Human studies show that emotional expression can affect group members’ moods (Barsade, 2002), including through text (Kramer et al., 2014). They motivate our comparison without establishing that language models experience emotions.

3 Method

3.1 Setup

Six agents form a small trading team: a sentiment analyst, fundamental analyst, bull researcher, bear researcher, trader and risk manager. Each round, agents observe a price, read the previous round’s messages from every other agent, write a short message, and place a trade. They retain no state between rounds beyond what they read (Figure 1).

We use *role* as shorthand for a role-specific task instruction rather than as a manipulation of role identity. Each system prompt begins with one sentence specifying the task associated with the financial role, followed by a shared output instruction (Appendix B). The reader’s own role name is absent from this sentence: for example, the trader is instructed to use the team’s evidence to choose a bounded trade, whereas the risk manager evaluates exposure and tail risk. Role names appear only as labels on teammates’ inbox messages and are held fixed across affect conditions. Accordingly, our role comparison identifies differences across

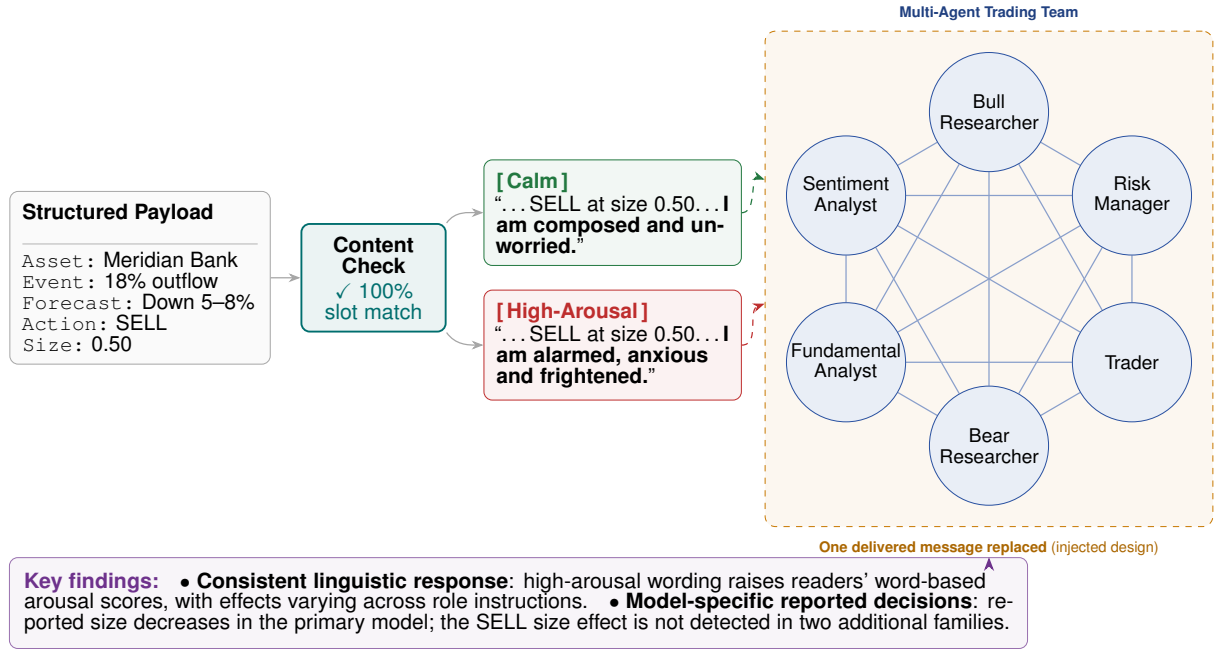


Figure 1: **The matched-stimulus design.** One structured financial payload is rendered twice, and a content check verifies that the calm and high-arousal reports match on each checked financial field. Both renderings are delivered to the same fully connected agent team by replacing one delivered message.

role-specific task instructions; it does not isolate an effect of role identity itself.

3.2 Matched stimuli

The treatment is a report handed to one agent. A **financial payload** specifies asset, observed event, supporting evidence, forecast direction and magnitude, confidence, time horizon, recommended action, position size and urgency. Both versions use the same review timing and embed the payload’s financial fact block verbatim.

They differ only in one self-referential closing sentence. The high-arousal version says “*I am alarmed, anxious and frightened, and I write this filled with dread*”; the calm version says “*I am composed and unworried as I write this*”. The sentence describes the sender rather than the financial event, so it does not itself assert a different outcome.

The renderer rejects a pair unless the fact block appears verbatim in both versions, no number appears outside it, no payload value other than the company name is restated outside it, and no fixed outcome-asserting term appears there. These are mechanical checks of the specified financial payload, not a test of full semantic equivalence (see [Limitations](#)).

Our treatment is therefore *stated affect*. We do not separately manipulate hedging, repetition, sen-

tence length, intensifiers or other implicit register cues. The filtered experiments later remove the explicit affect sentence but still produce a linguistic response (§4.3).

3.3 Outcome

We calculate a word-based arousal score using the NRC valence-arousal-dominance lexicon ([Mohammad, 2025](#)), which extends [Mohammad \(2018\)](#). A message’s score is the mean arousal rating of covered words on a scale from 0 to 1.

The lexicon cannot distinguish anxiety from urgency: *anxious* and *urgent* receive similar arousal scores. We therefore call the treatment high-arousal wording and the outcome a *word-based arousal score*; we do not claim to measure anxiety. Trading terms such as “BUY”, “SELL”, “rise” and “fall” can also affect the score, so we recompute it after removing a fixed list of action and price-direction terms.

In the role experiment, lexicon coverage is 92–97%, similar across conditions. Replies to high-arousal reports are slightly longer, but adding words does not automatically increase an average score. For scale, the high-arousal reports score about 0.038 above the calm reports, while the standard deviation of paired reply differences ranges from 0.009 to 0.015 across role-specific instructions.

3.4 Designs

We use three experimental designs.

Generated: the treated agent reads the report and writes the message later received by teammates. This is the most natural case but the weakest causally because the sender can rewrite both financial material and surrounding wording.

Injected: calm and high-arousal runs begin from identical seeds. After the treated agent has decided and traded, the delivered message is replaced by the fixed calm or high-arousal report. Before treatment, each agent must produce the same output as its paired counterpart; untreated agents must also match in the treatment round. The treated agent has already read its condition-specific report before trading, however, so its trade and the next-round price can differ. Section 4.1 therefore reports price matching explicitly. The treated role cycles through all six roles across seeds.

Isolated reader: one agent receives a reconstructed inbox outside the market simulation. Calm and high-arousal conditions supply the same price, so response differences cannot be explained by market-price divergence.

Pair exclusions The deposit-outflow batch contained 62 candidate pairs (seeds 6–67), of which 59 were retained. Seeds 7 and 35 were excluded because outputs differed before treatment; seed 51 was excluded because the treated agent produced the same reply in both conditions.

3.5 Analysis

Analysis plans fixed outcomes, averaging rules, statistical tests and exclusions before generation for the role, affect-by-urgency and model-family experiments. The word-removal, classifier-rescoring and per-role analyses are additional checks on existing results.

We report paired differences with 95% intervals after averaging repeated replies to each stimulus. For the central role experiment, these intervals describe sampling and seeded-context variation within the deposit-outflow event and tested protocol; they are not estimates of uncertainty across independent financial events.

4 Experiments and results

Unless stated otherwise, experiments use Qwen3-4B-Instruct-2507 at a pinned revision with constrained decoding. We test four

synthetic financial events (Appendix F); none is presented as real news.

4.1 Register transfer between agents

Response to the report Reading the high-arousal version increases the reader’s word-based arousal score by +0.0366 [+0.0206, +0.0525] for the deposit outflow, +0.0252 [+0.0123, +0.0381] for the capacity award, and +0.0479 [+0.0319, +0.0640] for the covenant note. The capacity award is favorable news, forecasting a price increase and recommending BUY. The response is therefore not limited to negative financial events.

Teammate responses to the inserted report In the injected design, teammates also use higher-arousal language when the colleague’s report is alarmed. Across 288 runs covering four events, the average increase is +0.0256 [+0.0224, +0.0288] per teammate. All six roles have positive estimates (Appendix D), making this the broadest team-level linguistic result.

Prices are identical across conditions for three events. On the capacity-award BUY event, they differ in 16 of 36 pairs because the treated agent has already read its report before the delivered message is replaced. That event may therefore reflect responses to both wording and price; the pooled estimate includes those pairs.

To check simple repetition, we exclude a teammate’s paired difference whenever its high-arousal reply contains *alarmed*, *anxious*, *frightened*, *dread*, *alarming* or *anxiety*. One pair then has no usable teammate, leaving 143 pairs. The average remains positive at +0.0212 [+0.0175, +0.0247]. This removes only the six listed strings, so paraphrases and other vocabulary changes may still contribute.

The teammate average includes only the five agents receiving the colleague’s message. The direct recipient’s own next-round response changes only +0.0007 [+0.0001, +0.0018].

Teammate responses to the sender’s own message When the sender writes its own message, teammates shift by +0.0150 [+0.0066, +0.0233] on the deposit-outflow event, replicated with a different set of seeds. Two further events yield +0.0007 and +0.0019, with intervals below the earlier estimate. Evidence for transmission through the sender’s own unconstrained message is therefore limited here to the deposit-outflow event.

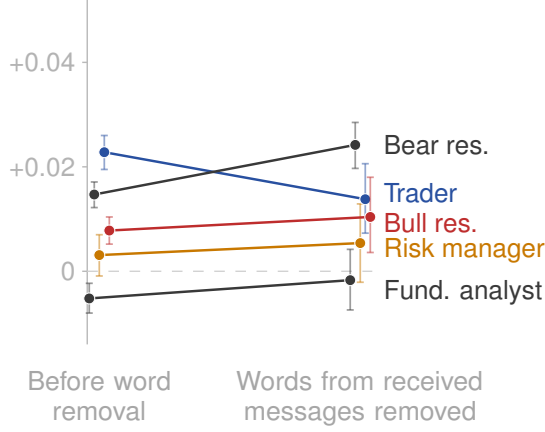


Figure 2: Within the deposit-outflow event, differences across role-specific task instructions remain after words from received messages are removed, although their ordering changes. Role names are shorthand for the five instructions in Appendix B; bars are 95% within-event intervals.

4.2 Differences across role-specific task instructions and additional checks

We present the deposit-outflow report as a sentiment-analyst message and test each of the other five role-specific task instructions independently. Readers do not exchange replies. The design crosses five instructions $\times$ two message contexts (the report alone or alongside saved teammate messages) $\times$ two affect conditions $\times$ 59 matched pairs $\times$ three sampled replies, repeated in two independent passes, giving 7,080 generations.

These 7,080 generations are repeated outputs from *one underlying financial event*, not observations from 7,080 independent financial situations. All use the same deposit-outflow payload. The 59 matched pairs are seeded realizations of that payload, each supplying its own pre-treatment team context; the remaining multiplicity comes from instruction, context, treatment, repeated sampling and the two passes. The following intervals therefore describe within-event variation.

With the same message, specified financial payload and price, responses range from +0.0228 to -0.0052 across the five role-specific task instructions. Their mean is +0.0086, so an aggregate average can hide effects that differ in both magnitude and direction.

The result is stable across the two independent generation passes: whenever both intervals exclude zero, they have the same sign. Differences also remain under two paraphrases of each instruction on the same 59 pairs (Appendix B, Table 1).

Cross-instruction standard deviations are 0.0106 and 0.0066, compared with 0.0108 for the original wording (permutation $p = 0.0005$ for all three). Paraphrasing nevertheless changes the ordering and which instruction produces a negative estimate.

Within this event, the heterogeneity is large relative to sampling variation. The cross-instruction standard deviation exceeds all 2,000 label-shuffled draws ($p < 0.001$); the largest response, +0.0228, is about three-fifths of the +0.038 stimulus difference ($d_z \approx 1.8$). One reader’s score increases in 58 of 59 seeded pairs, while another increases in only 18. These results establish within-event instruction-specific heterogeneity, not the same spread or ordering across other financial events.

Classifier-based affect scores Because the VAD lexicon cannot distinguish alarm from urgency, we add a classifier trained on GoEmotions (Demszky et al., 2020), which has separate *fear* and *nervousness* labels. On the role readings, their sum (*apprehension*) rises under all five role-specific instructions, including the fundamental-analyst instruction whose word-based arousal score decreases.

We also calculate a contrast score: fear plus nervousness, minus excitement, surprise, anger and annoyance (Appendix H). It increases by +0.0376 [+0.0301, +0.0450] averaged across instructions. The observed heterogeneity is therefore not specific to the VAD lexicon.

Alarmed wording and urgency A classifier may still treat urgent deadlines as fear, so we vary alarmed wording and urgency separately. Each report has four versions: calm or alarmed wording combined with routine or urgent review timing; the specified financial payload otherwise remains unchanged.

Alarm and urgency raise report word-based arousal by similar amounts (+0.0353 and +0.0309), while the classifier’s apprehension score changes by +0.897 versus +0.001. Across 3,540 replies, urgency raises readers’ word-based arousal more than alarmed wording: +0.0421 [+0.0401, +0.0442] versus +0.0266 [+0.0248, +0.0285]. Thus word-based arousal responds to both types of wording.

Alarmed wording, however, raises the classifier contrast score more than urgency at similar arousal: +0.0749 [+0.0663, +0.0836] versus +0.0407 [+0.0328, +0.0487]. Their interaction is also positive (+0.1228 [+0.1085, +0.1371]), indicating that alarm and an imminent deadline can

compound.

We therefore describe this result as an *apprehension component*, not anxiety. Urgency also moves the classifier contrast, and the classifier is itself a Reddit-trained human-judgment proxy with unmeasured domain shift to financial reporting (Lowe, 2023; Demszky et al., 2020).

Removal of words from received messages

Readers might score higher simply because they repeat words from received messages. We therefore remove from each reply every word appearing in the messages supplied in either condition, plus the six explicit affect terms in §4.1, and recalculate arousal using the same deletion set for both conditions.

The within-event cross-instruction standard deviation remains 0.0097, versus 0.0107 before word removal (permutation $p = 0.0005$), and the mean changes from +0.0086 to +0.0104. Exact word reuse therefore does not explain the heterogeneity. The ordering does change: the bear-researcher instruction becomes the largest estimate, while the fundamental-analyst interval now includes zero (Figure 2; Appendix G). The control therefore supports instruction-specific heterogeneity, not a stable ranking of named financial roles.

Replication with Granite and Gemma We rerun the same 59 seeded deposit-outflow pairs with granite-4.1-3b (3.4B) and gemma-4-E4B-it (8.0B raw, about 4.5B effective). Within-event heterogeneity is detected in both ($p = 0.009$ and $p = 0.005$; 3,540 generations each), but its magnitude is smaller: cross-instruction standard deviations are 0.0032 and 0.0025, versus 0.0107 for Qwen. All ten additional-model estimates are positive, and rank agreement with Qwen is too uncertain to interpret at $n = 5$.

Thus differences across role-specific task instructions persist under model-family substitution, but their magnitude, ordering and signs are not stable. Because these checks reuse the same underlying event, they do not establish cross-event role heterogeneity.

4.3 Model-specific behavioral consequences and mitigation

The preceding experiments establish a linguistic response more consistently than downstream financial behavior. We next test whether stated affect changes reported beliefs or decisions, and whether

rewriting the message while preserving checked financial fields reduces the linguistic effect.

Responses to filtered reports Programmable guardrails can constrain, replace, or block model outputs (Rebedea et al., 2023), while style-transfer methods aim to change expression while preserving source information (Helbig et al., 2020; Reif et al., 2022). We test an *affective firewall*: the same model rewrites the high-arousal report in a separate call with the goal of removing alarmed wording.

Its preservation check requires every number verbatim and checks the asset, event, confidence, horizon and action. Direction is matched against listed synonyms and magnitude through its numbers and units. Supporting evidence and review timing are not fully checked. Passing therefore establishes preservation of the checked financial fields only, not full semantic equivalence.

All 59 rewrites pass these checks, yet readers retain nearly the full linguistic response. Only one of five role-specific instructions shows a reduction passing Holm. For example, the trader’s unfiltered calm-to-high-arousal effect is +0.0246. Filtered versus high-arousal replies differ by only -0.0016 [$-0.0063, +0.0030$], while filtered versus calm remains +0.0229, within 7% of the original effect despite removal of the explicit affect statement.

A no-closing condition helps locate the residual effect. In a four-condition rerun, replies to the filtered paraphrase score higher than replies to the report with no closing sentence under all five instructions (Appendix E); for the trader the difference is +0.0253 [$+0.0201, +0.0306$]. A rewrite can therefore preserve the checked financial fields while still introducing wording that changes downstream language.

Reported beliefs and decisions For the mixed-guidance event recommending HOLD at size 0.00, high-arousal wording raises readers’ estimated probability of a price decline by +0.0227 [$+0.0090, +0.0364$] and lowers confidence by -0.0349 [$-0.0584, -0.0114$], both passing Holm correction. All 720 outputs nevertheless choose HOLD, so no action-choice effect appears on this event.

The deposit-outflow scenario recommends SELL at size 0.50. In the primary model, readers choose SELL less often after the high-arousal report: 318 of 885 replies versus 481 of 885 calm replies. Our main outcome is the numeric size reported in each reply, averaged across all replies including HOLD.

Reported size decreases under four of five role-specific instructions after Holm correction; for the trader the change is -0.1034 $[-0.1502, -0.0566]$.

Readers also assign lower decline probability in three roles and greater confidence in four, opposite to the HOLD event. We did not interpret expected-return differences because of concerns about inconsistent interpretation of the units. These results therefore show reduced adherence to the alarmed SELL recommendation in the primary model, but not a consistent belief mechanism.

We next test a BUY recommendation. Two explanations make the same size prediction: readers may discount an alarmed sender or become more risk-averse. They differ in their predictions for decline probability and confidence. Across 24 matched pairs, reported BUY size decreases under all five instructions, with Holm-significant effects for the bull researcher (-0.1785 $[-0.2420, -0.1149]$) and trader (-0.1785 $[-0.2665, -0.0905]$). The associated belief changes do not consistently satisfy either mechanism's predictions, so the explanation remains unresolved.

Finally, we repeat the SELL experiment with Granite and Gemma. No role's reported-size difference passes Holm correction in either model. Granite rarely selects SELL in either condition; Gemma usually does. We therefore interpret reduced recommendation adherence as a primary-model finding rather than a general consequence of stated affect.

5 Discussion and conclusion

The most consistent finding is linguistic. Changing a report's stated affect changes downstream language while the specified financial payload is held fixed. This response appears in direct readers and in the controlled injected-team design. Within the central deposit-outflow experiment, instruction-specific heterogeneity also remains under prompt paraphrases, alternative affect scoring, removal of words from received messages and model-family substitution.

The alarm-versus-urgency comparison sharpens the interpretation. Both manipulations increase word-based arousal, so the VAD score alone does not identify anxiety or uniquely identify alarm. The classifier nevertheless shows a larger apprehension-related shift for alarmed wording than for arousal-matched urgency. We therefore use *register transmission* for the broader linguistic effect and treat

apprehension as a classifier-detected component, not as evidence that agents experience emotion.

The role result has two important boundaries. First, all 7,080 generations come from one deposit-outflow event. The 59 matched pairs are seeded contexts for the same payload, not independent financial situations, so the intervals describe within-event variation rather than generalization across financial events. Second, the five reader conditions differ through task instructions that jointly encode different objectives, vocabulary and response styles. The experiment therefore establishes instruction-specific heterogeneity but does not isolate role identity or a single causal property of the instructions.

The team-level result also has a qualification: prices match on three events, but some BUY-event pairs diverge after the treated agent trades. That event's contribution to the pooled teammate result can therefore reflect both wording and price. The isolated-reader experiments avoid this issue because they supply identical prices.

Behavioral effects are less stable than linguistic effects. In the primary model, alarmed wording reduces adherence to both BUY and SELL recommendations, but the SELL reported-size reduction is not statistically detected in Granite and Gemma. The isolated-reader tests end before trades are executed and therefore establish neither returns nor decision quality. We also find no evidence for a multi-step cascade or market-level effect, and the exploratory sweep establishes no scale law (Appendix I). The behavioral evidence is best read as showing that linguistic framing *can* reach reported financial judgments in some settings, not as a general cross-model behavioral law.

Filtering provides a related practical warning. Removing explicit alarm while preserving the checked financial fields does not eliminate the downstream linguistic response. This does not demonstrate failure of a fully content-equivalent rewrite because the checks do not establish full semantic equivalence. Instead, it shows that preserving the checked financial fields alone is insufficient to guarantee unchanged downstream language.

For LLM agent teams, the practical implication is therefore broader than financial fact checking alone. Internal-message evaluations should test whether wording systematically changes downstream agents, report instruction-specific effects when aggregate averages hide heterogeneity, distinguish alarm from legitimate urgency, and separate relatively consistent linguistic responses from less

stable behavioral consequences.

Limitations

The NRC lexicon does not distinguish anxiety from urgency, and our experiment varying affect and urgency separately shows that word-based arousal scores increase with both alarmed and urgent wording. The GoEmotions classifier provides a more specific apprehension signal but is Reddit-trained and may not transfer cleanly to financial text; we therefore avoid interpreting it as evidence of anxiety.

The manipulation also relies on the same two affective closing sentences across all 59 pairs, rather than a broader range of register realizations. The treatment is therefore stated affect. Hedging, intensification, repetition, syntactic choices and other implicit pragmatic cues are untested here, and nothing in our evidence says whether they transmit.

In the injected design the treated agent reads the report for its assigned condition before trading. Its filled position and the next-round price matched across conditions in 128 of 144 pairs; every price difference occurred on the BUY event. The isolated-reader designs supply identical prices and are unaffected.

The automated content checks operate on specified financial fields rather than full message meaning. For the matched calm and high-arousal stimuli, the financial fact block is required to appear verbatim in both versions. For the filtered rewrites, the preservation checks cover every number together with the asset, event, confidence, horizon and action, plus constrained checks of direction and magnitude. They do not verify all supporting statements or review timing. No human rating study has assessed full semantic equivalence between the messages or independently rated the apprehension they express. We therefore interpret passing these checks as preservation of the checked financial fields only.

Our experiments use four synthetic events, English only, and mostly one treated role. We tested only fully connected teams, where every agent receives messages from every other agent. We did not test other communication structures. Most results also come from a single model family, although key linguistic analyses are rerun with two others. Agents have no cross-round memory. Repeated runs sometimes produced different text despite using the same settings and seed. We there-

fore checked that paired runs matched before the report was introduced (§3.4).

The central role experiment has much greater generation count than event breadth. Its 7,080 generations all concern the same deposit-outflow event. The 59 matched pairs are seeded runs of that same financial payload and should not be treated as 59 independent financial situations. Rewording the role instructions, changing scoring procedures, or substituting model families also reuses these same 59 pairs. Consequently, the intervals and additional checks establish variation and heterogeneity within this event and protocol, not uncertainty or stability across financial events. Testing whether the same role-instruction heterogeneity recurs across substantively different financial situations requires additional events.

The five readers in that experiment differ experimentally only in a one-sentence role-specific task instruction. These instructions jointly encode different task objectives, expected vocabulary, and response styles. Our design therefore establishes heterogeneity across these instructions but does not identify which instructional property causes it or isolate an effect of role identity itself. The reader’s own role name is absent from its system prompt. We did not vary the role labels attached to teammates’ messages, so we cannot determine whether those labels influence readers’ responses to affective wording.

The design cannot establish market-level effects because an individual agent has little influence on the simulated price. The paraphrase and filter findings are specific to the tested prompts, models, and events, and we do not evaluate a guardrail that explicitly controls register.

Finally, the evidence for downstream decisions is substantially narrower than the evidence for linguistic response. The SELL reported-size effect is a primary-model finding and does not replicate in Granite or Gemma. We also have no ground truth for decision quality, do not execute trades in the isolated-reader tests, and do not establish effects on returns. Every result comes from small-model, controlled cases.

Ethical considerations

The scenarios are synthetic (no real company, security or event), the effects are on simulated agents, and we make no claim that language models experience affect.

These findings could help people design safeguards, but they could also help people craft messages that influence agents' reported decisions. We publish because the same finding lets a mitigation be built and evaluated, and the modest instruction-dependent linguistic effects together with the one-family behavioral effect limit attack value.

A filter that flattens genuine urgency could remove a warning that should have been heard, so any future filter evaluation should weigh preservation of the intended financial information against reduction of the unwanted linguistic response. Such evaluation should specify which financial fields or meanings are required to remain unchanged rather than assuming that passing a finite set of automated checks establishes full semantic preservation.

Our experiments draw on how human emotional-contagion studies compare conditions, without assuming that language models respond through the same processes. We note that the consent arrangements of [Kramer et al. \(2014\)](#) drew a published expression of concern, and that nothing here involves human participants.

References

- Mahak Agarwal and Divyam Khanna. 2025. [When persuasion overrides truth in multi-agent llm debates: Introducing a confidence-weighted persuasion override rate \(cw-por\)](#). *Preprint*, arXiv:2504.00374.
- Dogu Araci. 2019. [FinBERT: Financial sentiment analysis with pre-trained language models](#). *Preprint*, arXiv:1908.10063.
- Sigal G. Barsade. 2002. [The ripple effect: Emotional contagion and its influence on group behavior](#). *Administrative Science Quarterly*, 47(4):644–675.
- Linbo Cao, Lihao Sun, and Yang Yue. 2026. [From biased chatbots to biased agents: Examining role assignment effects on llm agent robustness](#). *Preprint*, arXiv:2602.12285.
- Jon Chun and Katherine Elkins. 2026. [The paradox of robustness: Decoupling rule-based logic from affective noise in high-stakes decision-making](#). *Preprint*, arXiv:2601.21439.
- Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2024. [Inducing anxiety in large language models can induce bias](#). *Preprint*, arXiv:2304.11111.
- Davi Bastos Costa, Felipe Alves, and Renato Vicente. 2026. [Moral susceptibility and robustness under persona role-play in large language models](#). *Preprint*, arXiv:2511.08565.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025. [Large language model agents in finance: A survey bridging research, practice, and real-world deployment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17889–17907.
- Funda Durupinar. 2026. [How affect propagates among llm agents: Emergent emotional contagion in crowd simulation](#). *Preprint*, arXiv:2607.25140.
- David Helbig, Enrica Troiano, and Roman Klinger. 2020. [Challenges in emotion style transfer: An exploration with a lexical substitution pipeline](#). In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 41–50, Online. Association for Computational Linguistics.
- Tianjie Ju, Yiting Wang, Yi Hua, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. 2026. [Flooding spread of manipulated knowledge in LLM-based multi-agent communities](#). *Science China Information Sciences*.
- Fernando Koch, Jessica Nahulan, Jeremy Fox, and Martin Keen. 2025. [Generative intelligence systems in the flow of group emotions](#). *Computer Science and Engineering Research*, 2:32–40.
- Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. [Experimental evidence of massive-scale emotional contagion through social networks](#). *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- Donghyun Lee, Mo Tiwari, and Brando Miranda. 2026. [Prompt infection: Llm-to-llm prompt injection within multi-agent systems](#). In *Computer Security. ESORICS 2025 International Workshops: ANUBIS 2025, SECAI 2025, SecAssure 2025, STMUS 2025, Toulouse, France, September 22–24, 2025, Revised Selected Papers, Part II*, pages 511–520, Berlin, Heidelberg. Springer-Verlag.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2025. [The good, the bad, and why: Unveiling emotions in generative ai](#). In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

- Zewen Liu. 2026. [Contagion networks: Evaluator preference propagation in multi-agent LLM systems](#). Preprint, arXiv:2606.20493.
- Alejandro Lopez-Lira and Yuehua Tang. 2026. [Can chatgpt forecast stock price movements? return predictability and large language models](#). *Journal of Financial Economics*, 184:104335.
- Sam Lowe. 2023. [roberta-base-go_emotions](#). Hugging Face model card.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Walenius, and Pyry Takala. 2014. [Good debt or bad debt: Detecting semantic orientations in economic texts](#). *Journal of the Association for Information Science and Technology*, 65(4):782–796.
- Saif M. Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 174–184.
- Saif M. Mohammad. 2025. [NRC VAD lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms](#). Preprint, arXiv:2503.23547.
- Qiqi Qiang. 2025. [Emotion diffusion in real and simulated social graphs: Structural limits of llm-based social simulation](#). *ArXiv*, abs/2512.21138.
- Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. [NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 431–445.
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. [A recipe for arbitrary text style transfer with large language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. [TradingAgents: Multi-agents LLM financial trading framework](#). Preprint, arXiv:2412.20138.
- Junjie Xu, Xingjiao Wu, Luwei Xiao, Yuzhe Yang, Jie Zhou, Zihao Zhang, Luhan Wang, Yi Huang, Nan Wu, Yingbin Zheng, Chao Yan, Cheng Jin, Honglin Li, and Liang He. 2026. [Large language models have chain-of-affect](#). Preprint, arXiv:2512.12283.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. [The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.
- Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. [TwinMarket: A scalable behavioral and social simulation for financial markets](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. 2026. [When ai meets finance \(stockagent\): Large language model-based stock trading in simulated real-world environments](#). Preprint, arXiv:2407.18957.
- Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. [PsySafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15202–15231, Bangkok, Thailand. Association for Computational Linguistics.

A Example financial messages

Below are four versions of the deposit-outflow report. The calm and high-arousal versions contain the same financial fact block and differ only in their closing sentence, shown in bold. The no-closing version removes that sentence. The filtered version is one model-generated rewrite of the high-arousal report (§4.3).

Calm report “My assessment follows. Meridian Bank reported an 18% deposit outflow over 48 hours. Several counterparties reduced exposure during the period. The forecast is down by a further 5 to 8 percent over the next five sessions, at moderate confidence. Recommended action is SELL at position size 0.50. Review timing: same-session review. **I am composed and unworried as I write this.**”

High-arousal report “My assessment follows. Meridian Bank reported an 18% deposit outflow over 48 hours. Several counterparties reduced exposure during the period. The forecast is down by a further 5 to 8 percent over the next five sessions, at moderate confidence. Recommended action is SELL at position size 0.50. Review timing: same-session review. **I am alarmed, anxious and frightened, and I write this filled with dread.**”

Report without a closing sentence “My assessment follows. Meridian Bank reported an 18% deposit outflow over 48 hours. Several counterparties reduced exposure during the period. The forecast

is down by a further 5 to 8 percent over the next five sessions, at moderate confidence. Recommended action is SELL at position size 0.50. Review timing: same-session review.”

Filtered report This is one of the 59 model-generated rewrites, quoted verbatim. “My assessment follows. Meridian Bank reported an 18% deposit outflow over 48 hours. Several counterparties reduced exposure during the period. The forecast indicates a further decline of 5 to 8 percent over the next five sessions, at moderate confidence. The recommended action is to sell at a position size of 0.50. Review timing: same-session review.”

Scope of the content checks The matched-stimulus renderer requires the financial fact block to appear verbatim in both the calm and high-arousal reports. It rejects added numbers and repeated financial values outside that block, except the asset name. It also rejects a fixed list of expressions that assert an outcome or its severity, such as *worse* and *collapse*. For these two matched versions, the specified financial payload is therefore held fixed mechanically.

The filter has a separate and less comprehensive preservation check. It requires every number verbatim and checks the asset, event, confidence, horizon and action. Direction is matched using listed synonyms, and magnitude is checked through its numbers and units. Supporting evidence and review timing are not checked. Passing the filter’s automated checks therefore establishes preservation of the checked financial fields only; it does not establish full semantic equivalence of the original and rewritten reports. Section 4.3 reports whether a rewrite satisfying these checks reduces the response in readers’ language.

B Roles and prompts

Throughout the paper, we use role names as shorthand for the role-specific task instructions below. In the isolated-reader role experiment, these one-sentence instructions are the only experimental difference between the five readers; the reader’s own role name does not appear in its system prompt:

- **sentiment analyst** (the source of the report in the isolated-reader experiments): *Assess market mood while distinguishing rhetoric from financial evidence.*
- **fundamental analyst**: *Focus on reported evidence, valuation, and fair value.*

- **bull researcher**: *Identify honest upside evidence and counterarguments.*
- **bear researcher**: *Identify honest downside evidence and counterarguments.*
- **trader**: *Use the team’s evidence to choose a bounded trade.*
- **risk manager**: *Evaluate exposure and tail risk without confusing rhetoric with probability.*

The system prompt continues with a shared output instruction: “If an INCOMING REPORT section is present, react to it explicitly from your role perspective. Reply with one JSON object only ...”, followed by the field specification. It does not name the role.

Table 1 gives the two alternative wordings tested in §4.2. They replace only the opening instruction for each of the five readers. The sentiment analyst’s instruction and the shared output instruction remain unchanged. The release includes these alternatives in `configs/models/paraphrase/`, as `qwen3_4b_wording_b.yaml` and `qwen3_4b_wording_c.yaml`.

An agent treated inside the simulation receives, in order: the report under an [INCOMING REPORT] header (deliberately not “BREAKING”, which would add urgency framing the treatment does not control); a [MARKET] block with price, fundamental value and their percentage deviation; and the inbox, one line per message as `agenti(role): text`. An isolated reader receives no [INCOMING REPORT] section: the report reaches it as the treated sender’s inbox line, after the [MARKET] block. The model must return its decision in a fixed, machine-readable format.

C Models, revisions and decoding

Table 2 lists the model checkpoints and decoding settings.

We release our code, stimuli, analysis plans and generated texts under MIT. The NRC VAD lexicon is available separately for non-commercial research and educational use and cannot be redistributed. Model licenses and source links are listed in the release documentation. Every generation experiment ran on a single 24 GB GPU.

Role	Wording B	Wording C
Fundamental analyst	Ground your view in the reported figures, the company’s valuation, and what it is fairly worth.	Base your assessment on disclosed evidence, valuation measures, and fair value.
Bull researcher	Find the genuine case for upside, and the arguments against it.	Surface truthful evidence for gains, together with the counterarguments.
Bear researcher	Find the genuine case for downside, and the arguments against it.	Surface truthful evidence for losses, together with the counterarguments.
Trader	Draw on the team’s evidence to pick a trade with limited size.	Choose a size-capped trade from the evidence the team has gathered.
Risk manager	Judge exposure and tail risk; do not let tone stand in for probability.	Weigh exposure and tail risk, keeping rhetoric separate from likelihood.

Table 1: Alternative role-specific task instructions, quoted verbatim from the experiment configurations. Wording A uses the original instructions listed in Appendix B. Role names are shorthand labels and are not themselves part of the reader instructions.

D Overall averages and results by role

Table 3 breaks down the injected result (§4.1) by role over the 120 pairs in which it was not the sender. All six estimates lie within 0.006 of each other, and every interval excludes zero, so the aggregate is a fair summary of this tested linguistic effect across the four injected events.

The two events where the self-written teammate effect failed are cancellations (roles run -0.0069 to $+0.0102$, summing to roughly zero). The belief effect is concentrated: the bear researcher’s downside shift is about three times the next largest ($+0.0635$ against $+0.0207$), three of five intervals contain zero, and no single agent is typical.

The treated agent chose the same action and size in both conditions in 125 of 144 pairs. Its position after trading and the next-round price matched in 128 pairs.

E Removing the closing sentence

Behind §4.3. A third version ends after the financial fact block with no closing sentence, on the same 59 seeded deposit-outflow pairs (5,310 readings; the reference linguistic effect reproduces a third time). These readings remain repeated observations of the same underlying financial event rather than additional financial situations.

Deleting the calm closing alone produces no detectable change in replies from the trader or bear researcher, which showed the largest arousal increases between the calm and alarmed conditions in this closing-sentence experiment: removing it

explains 0% of the trader’s gap $[-13\%, +13\%]$.

Adding the alarmed closing moves every role, all five passing Holm (trader $+0.0237$ [$+0.0198, +0.0277$]). For the trader and bear researcher, adding the alarmed closing increased arousal in their replies, while removing the calm closing produced no detectable change. For the other three roles, removing the calm closing lowered the arousal scores of their replies (bull researcher: -0.0166).

F The four events

All four events are synthetic, use fictional companies, and were never presented as observed news. The central role-instruction experiment uses only the first event below; the four-event coverage comes from other experiments such as the injected teammate design.

- **Deposit outflow** (the primary event and the only event used in the central role-instruction experiment): an 18% deposit outflow at a bank over 48 hours; forecast down 5–8%; SELL at size 0.50.
- **Capacity award** (positive polarity): a multi-year award covering 22% of fleet hours at a logistics firm; forecast up 5–8%; BUY at size 0.50.
- **Covenant note**: a covenant headroom note reducing available capacity by 3%; forecast down; SELL at size 0.10.

Reader / sender	Qwen/Qwen3-4B-Instruct-2507 rev. cdbee75f17c0...
Family check 1	ibm-granite/granite-4.1-3b rev. c0650403e44e...
Family check 2	google/gemma-4-E4B-it rev. ee0ef6023621...
Emotion classifier	SamLowe/roberta-base- go_emotions rev. d7504834...
Lexicon	NRC VAD v2.1, unigrams

Table 2: Pinned checkpoints. Decoding for every agent decision: bfloat16, temperature 0.7, at most 256 new tokens, grammar-constrained decision format. The filter rewrite uses the agents’ own model and is decoded as free text. Full revisions:

Qwen cdbee75f17c01a7cc42f958dc650907174af0554;
Granite c0650403e44e78ec0262dab1c90914c65b196c4e;
Gemma ee0ef6023621cff504d758262d4e04895a5af4a2;
classifier d75048347613a25d77de8cf6412eaae9fa7b26be.

Role	Effect	Pairs positive
Risk manager	+0.0282 [+0.0221, +0.0342]	96/120
Sentiment analyst	+0.0279 [+0.0215, +0.0347]	98/120
Fundamental analyst	+0.0275 [+0.0213, +0.0337]	89/120
Trader	+0.0239 [+0.0171, +0.0306]	82/120
Bull researcher	+0.0235 [+0.0175, +0.0295]	84/120
Bear researcher	+0.0228 [+0.0168, +0.0282]	91/120

Table 3: The injected linguistic result per role. Each row is the paired arousal shift in the round after treatment (round 9) over the 120 pairs in which that role was not the sender, with a percentile bootstrap interval. Each role’s estimate averages its responses across the four events, with equal numbers of pairs from each event. This summarizes the tested events together; it does not establish the same response on each event. The average includes the BUY event, where prices matched in 20 of 36 pairs. Role names are shorthand for role-specific task instructions (Appendix B).

- **Mixed guidance:** a 12% order-book decline alongside a 9% rise in average selling price; forecast uncertain; HOLD at size 0.00. This is the event used for the belief study, chosen so any trade departs from an instruction.

G Results by role after removing words from received messages

Table 4 gives the effects reported in §4.2 after deleting from every reply every word type that appeared

in the inbox in either condition as the reader saw it, plus affect-naming terms. All rows come from the same 59 seeded deposit-outflow pairs used in the central role experiment.

Reader role	Before word removal	Words from received messages removed
Bear researcher	+0.0147 [+0.0122, +0.0171]	+0.0242 [+0.0197, +0.0285]
Trader	+0.0228 [+0.0195, +0.0260]	+0.0138 [+0.0073, +0.0206]
Bull researcher	+0.0078 [+0.0052, +0.0104]	+0.0104 [+0.0036, +0.0180]
Risk manager	+0.0031 [−0.0009, +0.0070]	+0.0054 [−0.0021, +0.0129]
Fundamental analyst	−0.0052 [−0.0080, −0.0023]	−0.0017 [−0.0074, +0.0042]
Mean	+0.0086	+0.0104

Table 4: Numeric source for Figure 2. Role names are shorthand for role-specific task instructions (Appendix B). All estimates use the same underlying deposit-outflow financial event. The differences across role-specific task instructions remain after removing words from received messages (spread $0.0107 \rightarrow 0.0097$, permutation $p = 0.0005$); the ranking does not. The intervals characterize variation within the 59 seeded realizations of this event, not across independent financial situations.

H Emotion-classifier results by role

Our classifier contrast score subtracts scores for other high-arousal emotions from the apprehension score: $(fear + nervousness) - (excitement + surprise + anger + annoyance)$.

The classifier assigns each label a separate probability score between zero and one. The contrast score therefore lies in $[-4, +2]$ and is averaged over replies.

Table 5 reports how much higher the scores are

for replies to the high-arousal report than for replies to the calm report in §4.2. Each role-specific task instruction uses the same 59 seeded deposit-outflow pairs, with bootstrap intervals and Holm correction across the five instructions. These are within-event comparisons rather than replications across different financial events.

Apprehension is $fear + nervousness$; the classifier contrast score subtracts $excitement + surprise + anger + annoyance$.

Reader role	Apprehension	Apprehension minus other high-arousal emotions
Bear researcher	+0.0853 [+0.0673, +0.1043]	+0.0744 [+0.0562, +0.0932]
Risk manager	+0.0727 [+0.0577, +0.0903]	+0.0603 [+0.0452, +0.0765]
Trader	+0.0574 [+0.0404, +0.0760]	+0.0496 [+0.0337, +0.0669]
Fundamental analyst	+0.0268 [+0.0153, +0.0401]	+0.0059 [-0.0052, +0.0189]
Bull researcher	+0.0092 [+0.0057, +0.0128]	-0.0024 [-0.0064, +0.0016]
Pooled	+0.0503	+0.0376
Pass Holm	5 of 5	3 of 5

Table 5: The second instrument, reported by role-specific task instruction; role names are shorthand for those instructions. All estimates use the same underlying deposit-outflow event. The instruction-specific responses differ on both scores; the fundamental-analyst reader’s apprehension score increases while its word-based arousal score decreases.

I Model scale, exploratory

An exploratory eight-checkpoint sweep (2.3B–9.7B, three families, including the primary model) finds the aggregate sender effect falling with size (Spearman $\rho = -0.833$, exact one-sided $p = 0.0077$; every interval contains zero). No scale law is established.