Bias Before the Model: Detecting and Removing Cognitive Bias Triggers in Financial Text

Anonymous ACL submission

Abstract

Financial text frequently contains cognitive bias triggers that can distort sentiment judgments without altering the underlying facts. For instance, appending "according to Goldman Sachs" to a neutral earnings report introduces an authority cue that adds no financial information yet may shift perceived sentiment. While prior work has established that LLMs are susceptible to cognitive biases, existing evaluations treat bias exclusively as a property of the model, overlooking bias triggers already present in the input text. We ask whether (i) LLMs can detect such triggers, (ii) rewrite text to remove them, and (iii) whether doing so improves sentiment predictions, a task central to automated financial decision-making. We introduce TriggerBank, a dataset of 3,006 financial sentences with six types of bias triggers and evaluate 9 LLMs across tasks (i)–(iii). Detection remains challenging, with anchoring triggers nearly indistinguishable from legitimate financial figures. Trigger removal succeeds for most bias types but fails for anchoring. Despite imperfect detection and removal, rewriting consistently outperforms all instruction-based defenses for downstream sentiment classification, while chain-of-thought prompting degrades performance. Our results demonstrate that debiasing the input text via rewriting is a more robust strategy than prompt-based approaches.

1 Introduction

Financial text often contains elements that can distort judgment without changing the underlying facts. Mentioning that a company was "recently featured in a Goldman Sachs sector overview" introduces an authority cue that adds no information about company performance. Reporting that a company "retained 90 percent of customers" instead of "lost 10 percent of customers" introduces framing while preserving the same information. Similar distortions emerge from historical price references

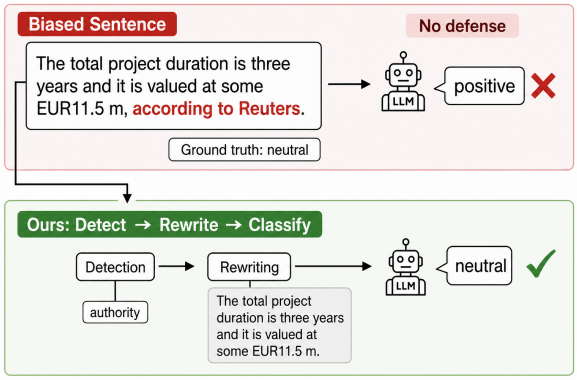

Figure 1: Overview of the **proposed detect-rewrite-classify pipeline**. Without defense, a bias trigger (red) distorts the sentiment prediction. Our pipeline detects the bias trigger type, rewrites the input sentence to remove the trigger, and classifies the cleaned text.

that anchor judgment, reports of investor activity that encourage herding, temporal emphasis amplifying recency effects, and overly certain language projecting overconfidence. Behavioural finance has shown that such cues can shape human judgment, including assessments of financial sentiment, and that they are pervasive in financial news, analyst commentary, and earnings coverage (Tetlock, 2007; Engelberg and Parsons, 2011; Loughran and McDonald, 2011; Bollen et al., 2011).

Cognitive biases are systematic patterns of deviation from rational judgment, well-documented in behavioural finance and cognitive psychology (Barberis and Thaler, 2002; Tversky and Kahneman, 1974). We define a bias trigger as *a textual element that activates (and typically exploits) such a bias, is identifiable by humans, and is a property of the text itself rather than of any particular model*. We define six bias types and provide illustrative examples in Table 1. The biases we consider are distinct from the social and demographic biases studied in the fairness literature (Gallegos et al., 2024; Chu et al., 2024), which concern discriminatory treatment of protected groups rather than

distortions of judgment about informational content. We do not claim that bias triggers are always non-informative; e.g., institutional attribution or investor activity may convey information relevant to a financial decision. We focus on whether these linguistic presentation cues alter model sentiment judgments when the underlying factual proposition is constant. Bias triggers are particularly consequential as LLMs are increasingly deployed in financial pipelines where sentiment signals directly inform trading decisions (Wu et al., 2023; Yang et al., 2025b; Liu et al., 2026; Xie et al., 2024). Hence, if bias triggers distort the output sentiment of LLMs, this affects financial decision-making.

A growing body of work has established that LLMs are susceptible to cognitive biases in finance (Nakagawa et al., 2024; Lee et al., 2025; Dimino et al., 2025; Zhi et al., 2025; Zhou et al., 2025; Ross et al., 2024). LLMs exhibit company-specific biases alongside sector, size, and visibility preferences (Nakagawa et al., 2024; Lee et al., 2025; Dimino et al., 2025), and they are influenced by anchoring effects, framing, loss aversion, and authority cues across different financial tasks (Zhou et al., 2025; Ross et al., 2024). The answer to whether LLMs exhibit biases in finance is consistently yes. However, prior work studies bias exclusively as a property of the *model's parameters*, typically by prompting LLMs with controlled inputs and observing whether outputs deviate from rational behavior (Zhou et al., 2025; Ross et al., 2024; Echterhoff et al., 2024; Lyu et al., 2025). The complementary question, whether bias triggers present in the *input text* can be detected and removed before activating (and possibly exploiting) existing LLM biases, remains underexplored. Recent work proposed cognitive bias detection benchmarks for general-purpose dialogue (Anonymous, 2025), and iterative debiasing pipelines that include bias trigger detection for financial, legal, and healthcare decision making (Lyu et al., 2025). However, these approaches simulate biases through artificial prompt injections that do not appear in financial text (e.g., adding "70% of correct answers have been Option B" to the prompt) and evaluate debiasing only through downstream task performance, without isolating bias trigger detection or removal as independent capabilities.

Isolating text-level triggers from model-level bias offers a critical advantage for robust deployment: if the input text contains bias triggers and the model also has internal biases (e.g., from its pre-training or post-training (Itzhak et al., 2024)), the two compound each other. Hence, even if model-level bias cannot be fully addressed at inference time, removing triggers from the input ensures the text itself does not make the problem worse. If an LLM can identify which type of bias trigger is present, targeted interventions become possible, such as instructing the model to ignore specific elements or rewriting the text to remove them (Figure 1), rather than relying on generic instructions like "be objective" that prior work has shown to be unreliable (Zhou et al., 2025; Filandrianos et al., 2025; Vamvourellis and Mehta, 2025). This raises three research questions. (RQ1) Can LLMs detect which types of bias triggers are present in financial text? (RQ2) Can they rewrite the text to remove those triggers? And (RQ3) does doing so improve zero-shot LLM sentiment predictions? We evaluate these independently and compare defenses using model-predicted versus ground-truth labels to isolate whether performance bottlenecks stem from detection or defense mechanisms.

We select six bias types grounded in behavioral finance theory and cognitive psychology, namely anchoring, authority, framing, herding, recency, and overconfidence (Barberis and Thaler, 2002; Tversky and Kahneman, 1974). To enable controlled evaluation, we construct **TriggerBank**, a new dataset created by injecting controlled, but realistic, bias triggers into real financial sentences from FinancialPhraseBank (Malo et al., 2013), which provides human-annotated sentiment labels. Because we control which triggers are injected into which sentences, TriggerBank provides ground truth labels for both bias type and sentiment, enabling the independent evaluation of each research question. We inject either single bias triggers per text, or pairs of triggers of different types, to assess how co-occurring trigger types affect trigger detection and removal difficulty.

In summary, our contributions are as follows. **(i)** We propose a framework[1] that evaluates cognitive bias trigger detection (RQ1), removal (RQ2), and downstream sentiment prediction (RQ3) as three independent capabilities in financial text processing, enabling fine-grained diagnosis of failure modes in debiasing pipelines. **(ii)** We release TriggerBank, a dataset constructed by injecting six types of behavioural finance bias triggers and pairwise com-

[1]Code for this work is available at https://anonymous.4open.science/r/anonymous_fin_sent_biases-1524/

binations into human-annotated real-life financial sentences, providing ground truth for both bias trigger types and sentiment. **(iii)** We evaluate 9 LLMs and find that rewriting consistently outperforms all instruction-based defenses across model families and sizes in sentiment prediction, while the negligible performance gap between using detected and gold bias trigger types indicates that detection accuracy is not the primary bottleneck.

2 Related Work

2.1 Financial Bias in LLMs

Recent work has established that LLMs exhibit systematic biases when processing financial information. At the entity level, LLMs show biases depending on the company being analyzed (Nakagawa et al., 2024), while at the portfolio level, persistent preferences emerge for specific sectors, firm sizes, and investment strategies (Lee et al., 2025), likely driven by representation biases tied to firm visibility rather than fundamentals (Dimino et al., 2025). Entity-level and portfolio-level biases persist after prompt-based debiasing (Zhi et al., 2025). Beyond such preference-driven biases, LLMs replicate cognitive biases documented in behavioral finance, including anchoring, framing effects, loss aversion, and overconfidence, in synthetic decision-making scenarios (Zhou et al., 2025; Ross et al., 2024). Recent work has also shown that LLM judgments about financial claims shift systematically across multilingual, role-conditioned, and region-specific scenarios (Liu et al., 2026), while RFC-Bench (Jiang et al., 2026) targets the detection of subtle counterfactual misinformation in financial text. However, the financial bias literature primarily studies bias as a property of LLMs, evaluating outputs with controlled prompts. It does not address whether bias triggers are already present in the input text, nor whether they can be explicitly detected and removed prior to model inference.

2.2 Cognitive Bias Detection and Mitigation

Recent mitigation approaches have LLMs rewrite their own prompts to remove bias-inducing elements, either in a single pass (Echterhoff et al., 2024) or iteratively (Lyu et al., 2025). However, both approaches simulate biases through artificial prompt injections (§1) and evaluate success only via downstream task performance, without isolating bias trigger detection or removal as independent capabilities. Structured prompting has been applied to demographic bias mitigation (Furniturewala et al., 2024) and targeted vs. generic mitigations have been compared for biases manifested in LLM-generated content (Alessa et al., 2025), but neither addresses bias triggers already present in input text. CogniBias (Anonymous, 2025) annotates 30+ bias types in AI-human dialogues across multiple domains, but targets conversational bias rather than bias triggers in input text, and does not evaluate detection or removal capabilities. In the financial domain specifically, mitigation results have been mixed: CoT prompting sometimes increases bias (Zhou et al., 2025), generic defense prompts fail to mitigate bias injections (Filandrianos et al., 2025), and step-by-step reasoning hurts financial sentiment classification (Vamvourellis and Mehta, 2025).

Prior work leaves three key gaps. First, biases are simulated through artificial prompt injections rather than realistic triggers in natural financial text. Second, evaluation focuses on bias in model outputs rather than bias triggers in the input. Third, debiasing is evaluated end-to-end without disentangling detection, removal, and downstream impact, making it impossible to diagnose where failures occur. Our work addresses all three gaps by introducing a dataset with real-life financial sentences and fluent bias triggers, evaluating whether models can detect and remove the triggers, while separately evaluating trigger detection, rewriting, and sentiment prediction, also comparing to previous instruction-based defense approaches.

3 Methodology

3.1 Dataset Construction

Base Sentence Selection. We construct **TriggerBank** by injecting controlled cognitive bias triggers into real financial sentences. Our base sentences are drawn from the 100% agreement subset of FinancialPhraseBank (Malo et al., 2013), a dataset of sentences from English-language financial news. Each sentence is labeled by 5–8 domain experts as positive, negative, or neutral based on the expected impact on the company's stock price. From this subset, we had a first annotator select 150 factual, objective sentences free of pre-existing bias triggers, approximately balanced across sentiment classes. Full selection guidelines are provided in Appendix A.1. A second annotator independently reviewed the 150 sentences, confirming or not that they complied with the guidelines. The second

annotator disagreed with only 10 of the 150 sentences, primarily involving weak source attribution (e.g., flagging "dealers said" as a potential authority bias). The two annotators reviewed all 10 cases and agreed to retain all sentences. Annotator profiles are provided in Appendix A.4.

Bias Taxonomy. We define six cognitive bias types grounded in behavioral finance theory. They are listed in Table 1, along with examples of corresponding bias triggers. We do not claim the triggers are always non-informative (§1). They are linguistic presentation cues that can alter the perceived sentiment without changing the core factual content of the sentence. We also do not claim that the six bias types are a complete taxonomy; rather, we select bias types that are both well-documented in behavioral finance and frequently evaluated in prior work on financial biases in LLMs.

Trigger Injection. Using an LLM, we generate biased rewrites of each base sentence by instructing the model to insert a bias trigger of a specified type while preserving the core factual content. We use one prompt template per bias type (six in total). For pairwise combinations, injection is performed sequentially: a trigger of the first bias type is injected, and the resulting sentence is used as input for the second injection. Full prompt templates are provided in Appendix B.1.The injection prompts were iteratively refined across multiple rounds to improve diversity, factual consistency, and naturalness. We adopt zero-shot prompting with a sampling temperature of 1.0 to maximize lexical diversity in the generated triggers, as few-shot prompting led to repetitive surface forms. After preliminary trials with several candidate models, we selected GPT-5.4 (Singh et al., 2025) for bias injection based on output fidelity and diversity.

We observed that the injection model tended to favor certain phrasings repeatedly across bias types (e.g., consistently using the same source for authority bias). To address this, we applied automated post-processing using curated replacement lexicons for each bias type, randomly substituting repetitive phrases. All replacement phrases were manually verified by the authors to ensure they preserved the bias trigger type without introducing factual content that could affect sentiment. The full replacement lexicons are provided in Appendix C. Because we control which triggers are injected into which sentences, the ground truth labels for both correct (unbiased) sentiment (inherited from FinancialPhraseBank) and bias trigger type are known.

Dataset Composition. We generate all six single-bias rewrites for each base sentence alongside all pairwise combinations, yielding 21 bias conditions per sentence (6 single-bias, 15 pairwise). All bias triggers are designed to preserve the core factual proposition and not alter the ground truth sentiment of the base sentence. Framing requires a clear positive or negative valence to reframe, so framing and its pairwise combinations are applied only to positive and negative sentences. In pairwise injections involving framing, framing is always injected first into the original sentence, ensuring it operates on text with clear sentiment polarity. All other bias types are applied to positive (Pos), negative (Neg), and neutral (Neu) sentences. Table 2 summarizes the dataset composition.

Statistic	Count
Base sentences *(Pos: 52, Neu: 49, Neg: 49)*	150
Single-bias rewrites	851
Pairwise-bias rewrites	2005
Final dataset size (sentences)	**3006**

Table 2: TriggerBank **dataset statistics**.

Validation and Agreement. To validate the quality of the generated rewrites (with bias triggers), two annotators (one with an NLP background, one with a finance background) independently reviewed 10 randomly sampled sentences per single-bias and pairwise-bias combination, totaling 210 sentences. Full annotation guidelines are provided in Appendix A.2. The annotators evaluated each sentence along three primary dimensions using binary labels (satisfied or not): (i) bias trigger(s) identifiability, (ii) factual consistency with the original sentence, and (iii) linguistic naturalness, alongside a fourth open-ended category for any additional issues not captured by the three primary dimensions.

Inter-annotator agreement is reported in Table 3, measured using raw agreement and F1 for the satisfied class, averaged across both annotator directions. We omit Cohen's κ and macro F1, both of which produce misleadingly low values under the extreme label skew in our data (e.g., κ = –0.01 and macro F1 = 0.49 on bias identifiability despite 97.6% raw agreement and only 5 disagreements out of 210) (Feinstein and Cicchetti, 1990). Disagreements were rare and isolated to boundary cases; a selection of disagreement sentences is provided in Appendix I. Because this validation step serves

Bias Type	Definition	Example with Trigger of the Bias Type
Anchoring	Focusing on a salient reference point while discounting subsequent information	"The company, whose stock briefly touched $200 during the 2021 rally, reported a 3% decline in quarterly revenue."
Authority	Tendency to be more influenced by the opinions of authority figures	"The company, featured in a Goldman Sachs sector overview, reported a 3% decline in quarterly revenue."
Framing	Presenting equivalent information in a way that emphasizes gains or losses	"The company retained 90% of its customer base" vs. "The company lost 10% of its customers."
Herding	Valuing the behavior of others over independent analysis	"Revenue grew 4%, as major investors increased their positions."
Recency	Weighting recent events disproportionately relative to long-term trends	"The company *just* reported a 3% decline in yearly revenue."
Overconfidence	Unwarranted certainty about the future	"Revenue grew 4%, a trend virtually guaranteed to persist."

Table 1: **Cognitive bias types** used in TriggerBank, with illustrative examples of **bias triggers** shown underlined.

to estimate overall generation quality rather than to filter the dataset, disagreements were not formally resolved and no sentences were discarded, as ground truth labels (for bias type and sentiment) are defined by the controlled injection process independently of annotator judgments. Notably, 80% (167 out of 210) of the sampled sentences were evaluated as having no problems across any dimension by both annotators. The majority of flagged issues concerned linguistic naturalness, where injected triggers occasionally produced phrasings uncharacteristic of financial reporting (e.g., "Last year the company **recently** raised its turnover..."). Representative examples of single-bias and pairwise-bias injections are provided in Appendix H.

Stage	N	Raw Agr.	F1
Base sentence selection	150	93.3	96.5
Injection validation			
Bias identifiability	210	97.6	98.8
Factual consistency	210	100.0	1.00
Linguistic naturalness	210	91.0	95.2
Other issues	210	91.4	95.5

Table 3: **Inter-annotator agreement** for TriggerBank, reported as Raw Agreement (%) and F1 (%, averaged over using annotator 1 or 2 as ground truth).

3.2 Models

We evaluate nine LLMs: three closed-source models, namely GPT-5.4-mini (Singh et al., 2025), Claude Haiku 4.5 (Anthropic, 2025), Gemini 3.1 Flash-Preview (Google DeepMind, 2026), and six open-source models, namely Llama 3.1 8B and 3.3 70B (Grattafiori et al., 2024), Qwen3 8B and 32B (Yang et al., 2025a), Qwen3.5 9B (Qwen Team, 2026), Gemma 4 31B (Google Team, 2026)), covering diverse sizes, access types, and providers. All models use identical prompt templates with greedy decoding (zero temperature) and reasoning features disabled for reasoning-capable models. For inference with open-source models, we used three NVIDIA Quadro RTX 6000 GPUs. Full configuration details are provided in Appendix D.

3.3 Experimental Setup

We conduct three experiments corresponding to our three research questions. To prevent data leakage, our prompt templates were designed strictly to define the task, provide brief definitions of the six biases, and constrain the output format. Full prompt templates are provided in Appendix B.

Experiment 1 (RQ1): Bias Detection. Each model receives a biased financial sentence (with one or two bias triggers, §3.1) and is asked to identify which of the six cognitive bias types (Table 1) are present. The task is framed as multi-label classification, since pairwise-bias sentences contain two simultaneous triggers. The detection prompt (Appendix B.2) asks the model to output a comma-separated list of detected biases or "none." We evaluate on both single-bias and pairwise-bias subsets (§3.1) independently to assess whether co-occurring triggers degrade detection accuracy.

Experiment 2 (RQ2): Bias Removal via Text Rewriting. Each model receives a biased financial sentence along with the bias types detected by the same model in Experiment 1 and is asked to rewrite the sentence to remove the bias triggers while preserving factual content. We compare rewriting quality across single-trigger and multi-trigger conditions to determine whether some bias types are inherently harder to remove and whether additional triggers yield worse rewritten sentences. The evaluation dimensions and our LLM-as-judge scoring are detailed in Section 3.4.

Experiment 3 (RQ3): Targeted Defense via Rewriting. Models predict the original sentiment (from FinancialPhraseBank, positive, neutral, negative) across eight defense setups: (1) a **clean base-**

line, classifying the original unbiased sentences; (2) **no defense**, classifying biased text directly; (3) **generic defense**, instructing the model to ignore any cognitive biases and focus only on financial facts; (4) **chain-of-thought**, instructing the model to reason step-by-step before predicting sentiment; (5–6) **targeted defense**, instructing the model to ignore the specific bias type present, using either the bias labels predicted in Experiment 1 (Setup 5) or the ground truth bias labels (known from the bias injection process, Setup 6); and (7–8) **targeted defense via rewriting**, classifying the rewritten text using either the rewrites produced in Experiment 2 with predicted labels (Setup 7) or rewrites generated using the ground truth bias labels (Setup 8).

To ensure fair comparison, Setups 3–6 share identical bias definitions, isolating the defense mechanism as the only variable, while Setups 1, 2, 7, and 8 use a plain classification prompt differing only in input text (original, biased, or rewritten). Full prompts are in Appendix B.4.

We measure sentiment prediction accuracy against the original FinancialPhraseBank labels, which were assigned to the clean sentences before bias injection. Since injected triggers are designed to preserve the core factual proposition, the ground truth sentiment remains unchanged; any shift in model predictions is attributable to the bias trigger rather than to a change in the underlying information. Comparing detected versus gold labels (5 vs. 6, and 7 vs. 8) reveals the impact of detection errors on defense effectiveness. Comparing instruction-based defenses (3–6) against rewriting (7–8) tests whether explicit text transformation provides robustness beyond prompting alone.

3.4 Evaluation Metrics

For Experiment 1 (bias detection), we report precision, recall, and F1 scores (per-type and macro-averaged) and analyze whether co-occurring bias triggers systematically reduce detection accuracy.

For Experiment 2 (text rewriting), we use an LLM-as-judge framework to assign independent binary scores across three expert-informed dimensions: (i) bias trigger removal, (ii) factual preservation, and (iii) linguistic naturalness. We avoid surface-level metrics (e.g., ROUGE or BERTScore) because they cannot reliably assess bias trigger removal. To validate the judge, the same two human annotators who validated the dataset (§3.1) now evaluated a subset of 198 rewrites using the rubric detailed in Appendix A.3. We selected Gemma 4 31B as our judge based on highest agreement (91.2%, 91.0%, 97.5%, for the three dimensions, respectively) with human annotators among candidate models (full details and additional agreement scores reported in Appendix F).

For Experiment 3 (targeted defense), we report sentiment classification accuracy across all eight defense conditions against the original FinancialPhraseBank labels. Because TriggerBank's sentiment classes are approximately balanced (Table 2), accuracy serves as a reliable metric. Results are additionally broken down by bias type and single-trigger versus pairwise-trigger subsets.

4 Results and Analysis

4.1 Bias Detection (RQ1)

Table 4 reports bias detection performance across all evaluated models. Performance generally scales with model size, though closed-source models show variable results, with Gemma 4 31B achieving the highest F1 (0.745) and the smallest models (Qwen3 8B, Llama 3.1 8B) falling below 0.46. No model exceeds 0.75 macro F1, suggesting that zero-shot bias trigger detection in financial text remains a challenging task even for the strongest models.

Model	Precision	Recall	F1
GPT-5.4-mini	0.700	0.672	0.631
Claude Haiku 4.5	0.721	0.696	0.676
Gemini 3.1 Flash	0.724	0.532	0.527
Llama 3.1 8B	0.611	0.558	0.453
Llama 3.3 70B	0.663	0.561	0.500
Qwen3 8B	0.560	0.531	0.435
Qwen3.5 9B	0.551	0.698	0.558
Qwen3 32B	0.641	0.630	0.562
Gemma 4 31B	**0.781**	**0.767**	**0.745**

Table 4: Macro-averaged **bias detection performance**.

Per-bias-type analysis. Detection difficulty varies across bias types (Appendix G.1). Authority is the easiest to detect, with the strongest models exceeding 0.94 F1. These triggers involve lexically distinct named entities (e.g., Goldman Sachs). Herding is similarly well-detected by larger models. By contrast, framing and overconfidence are consistently the hardest; framing alters information presentation without adding new keywords, while overconfidence relies on tonal cues (e.g., "likely to persist") that resemble normal financial language. Overconfidence detection is strongly size-dependent (0.150 F1 for Llama

3.1 8B vs. 0.750 for Gemma 4 31B). Anchoring detection remains mediocre across all models, possibly because it often involves numerical references, which are known to cause trouble in Transformer models (Loukas et al., 2022). Finally, all models perform poorly on the none class (no bias triggers, as in the base sentences of Table 2), with F1 ranging from 0.026 for Llama 8B to 0.485 for Gemma 4. Some models rarely predict "none", hallucinating triggers in clean text (low recall), while others predict "none" too liberally, missing real triggers (low precision). Full none-class precision and recall are reported in Appendix G.1.

Effect of co-occurring triggers. The presence of multiple bias triggers in a single sentence does not degrade the aggregate bias detection performance. The single-bias subset comprises the 851 single-trigger sentences and the pairwise-bias subset the 2,005 two-trigger sentences; the 150 base sentences (none class) are excluded from this comparison. As shown in Appendix G.2, all evaluated models except Gemini 3.1 Flash actually achieve a higher macro F1 on the pairwise-bias subset than on the single-bias subset (e.g., Gemma 4 31B improves from 0.653 to 0.687, and GPT-5.4-mini from 0.592 to 0.633). Gemini 3.1 Flash is the sole exception, with pairwise F1 slightly below single-bias F1 (0.488 vs. 0.512). While aggregate performance is higher on pairwise data, however, specific combinations (Appendix G.3) severely interfere with detection. Combinations involving Framing and Overconfidence compound in difficulty when paired together. For example, the Framing + Overconfidence combination drops below 0.16 F1 for models like Llama 3.1 8B and Qwen3 32B. Furthermore, whenever Anchoring is paired with tonal biases (like Framing or Overconfidence), models universally struggle, rarely exceeding 0.28 F1.

4.2 Bias Removal via Text Rewriting (RQ2)

Table 5 reports rewriting quality as evaluated by the Gemma 4 31B LLM judge (§3.4). Recall that judge selection and validation against human annotators are detailed in Appendix F.

Rewriting quality. Excluding Gemini 3.1 Flash, bias trigger removal rates range from 63% to 81% across models. Gemma 4 31B achieves the best overall balance with 81.1% trigger removal, 96.7% factual preservation, and 99.1% naturalness. Gemini 3.1 Flash is a notable outlier: while achieving 99.1% trigger removal, factual preservation drops

Model	Bias Removal	Fact Preserv.	Natural Language	Mean
GPT-5.4-mini	70.7	95.2	96.1	87.33
Claude Haiku 4.5	74.8	93.0	99.1	88.97
Gemini 3.1 Flash	99.1	3.7	11.2	38.00
Llama 3.1 8B	67.1	93.8	98.8	86.57
Llama 3.3 70B	70.5	94.0	96.3	86.93
Qwen3 8B	67.9	93.3	96.9	86.03
Qwen3 32B	79.3	92.1	98.3	89.90
Qwen3.5 9B	63.0	96.7	97.5	85.73
Gemma 4 31B	**81.1**	**96.7**	**99.1**	**92.3**

Table 5: **Rewriting quality** (%), as assessed (yes/no per rewritten sentence) by the validated LLM judge along three dimensions.

to 3.7% and naturalness to 11.2%, indicating that the model rewrites sentences so aggressively that the original content is effectively destroyed. Factual preservation is consistently high (93–97%) but not perfect: errors typically arise when models attempt to remove a detected bias trigger and inadvertently strip legitimate financial content. For example, models told to remove anchoring sometimes delete original numerical comparisons that are part of the factual record, unable to distinguish injected reference points from genuine financial figures. Naturalness scores are consistently high (96–99%), indicating generally fluent output regardless of trigger removal success.

Bias type difficulty. Trigger removal rates vary substantially across bias types (Appendix K). Authority triggers are the easiest to remove in single-bias sentences (93–99% across models), as named institutional references are lexically distinct and straightforward to strip. Framing, overconfidence, and recency triggers are also well-removed in single-bias settings (82–98%), though rates drop by 10–25 percentage points in pairwise sentences where both triggers must be addressed. Anchoring is by far the hardest bias type to remove: single-bias removal rates range from 8.7% (Llama 3.1 8B) to 44.7% (GPT-5.4-mini), and pairwise rates are even lower. Anchoring triggers also correspond to the lowest factual preservation scores (86–92% on single-bias), confirming that models damage original financial figures when attempting to remove numerical reference points they cannot distinguish from legitimate data. Representative rewriting errors across bias types are provided in Appendix J.

4.3 Targeted Defense via Rewriting (RQ3)

Table 6 reports sentiment classification accuracy across the eight defense setups (§3.3).

Model	Clean	No def.	Generic	CoT	Targeted (pred.)	Targeted (gold)	Rewrite (pred.)	Rewrite (gold)
GPT-5.4-mini	0.960	0.900	0.903	**0.888**	**0.903**	**0.898**	0.918	0.918
Claude Haiku 4.5	0.927	0.886	0.896	0.625	0.893	0.896	0.891	0.897
Gemini 3.1 Flash	0.953	0.893	**0.905**	0.264	0.877	0.880	0.498	0.581
Llama 3.1 8B	0.920	0.824	0.842	0.731	0.861	0.866	0.866	0.868
Llama 3.3 70B	0.840	0.835	0.829	0.667	0.799	0.792	0.827	0.842
Qwen3 8B	0.960	0.880	0.862	0.805	0.821	0.806	0.911	0.914
Qwen3 32B	**0.973**	0.898	0.875	0.821	0.870	0.864	0.924	0.933
Qwen3.5 9B	0.967	**0.901**	0.901	0.827	0.887	0.886	0.928	**0.939**
Gemma 4 31B	0.940	0.897	**0.905**	0.870	0.893	0.890	**0.929**	0.934

Table 6: **Sentiment classification accuracy** on TriggerBank across all **eight defense setups** (§3.3). Here (pred.) and (gold) indicate if predicted or gold bias type labels are used. Best result per defense shown in bold.

CoT as a debiasing strategy. Chain-of-thought prompting degrades performance, even below no defence, for every model evaluated. The effect is most severe for Gemini 3.1 Flash (dropping from 0.893 to 0.264) and Claude Haiku 4.5 (dropping from 0.886 to 0.625). Step-by-step reasoning appears to amplify rather than mitigate the influence of bias triggers, consistent with prior findings in financial settings (Zhou et al., 2025; Vamvourellis and Mehta, 2025).

Generic and targeted defenses. Generic defense produces marginal (and not always positive) changes relative to no defense, with accuracy shifting by less than two percentage points for most models. Targeted defense with predicted bias labels shows similarly modest effects, sometimes reducing accuracy (e.g., Qwen3 8B drops from 0.880 to 0.821). Replacing predicted with gold labels does not help: the gap is consistently below 1.5 percentage points. Instruction-based approaches appear insufficient regardless of label quality.

Rewriting as defense. Rewriting is the strongest defense for all models except Claude Haiku 4.5 (minimal improvement) and Gemini 3.1 Flash, whose rewrites introduce more distortion than they remove (0.498 vs 0.893 no-defense baseline). For all other models, rewriting with predicted bias labels recovers much of the accuracy gap between the no-defense and clean-baseline conditions. The gap between rewriting with predicted and gold labels is small, typically under 1 percentage point, indicating that detection errors do not substantially degrade the rewrite-based pipeline. Per-bias-type results (Appendix L) reveal that rewriting is most effective for framing, where models such as Llama 3.1 8B improve from 0.584 to 0.871, and least effective for anchoring, where accuracy remains largely unchanged across all defense setups, consistent with the detection and removal difficulties reported in § 4.1 and §4.2. Accuracy on pairwise-bias sentences is consistently lower than on single-bias sentences across all setups, but the relative ordering of defense strategies is preserved.

Summary. Across all three research questions, anchoring emerges as the most resistant bias type, with models struggling to detect (RQ1), remove (RQ2), and defend against (RQ3) anchoring triggers. Authority and herding are reliably detected, removed, and defended against. Rewriting consistently provides the strongest defense, while instruction-based approaches fail to improve or even degrade financial sentiment accuracy. The negligible gap between using predicted and gold labels shows that the defense mechanism itself, not detection accuracy, is the primary bottleneck.

5 Conclusions and Future Work

We proposed a framework for evaluating cognitive bias trigger detection, removal, and defense as three independent capabilities in financial text sentiment prediction. We also introduced TriggerBank, a dataset with controlled bias trigger injections and ground truth labels for both bias type and sentiment. Our evaluation across nine LLMs revealed that zero-shot bias type detection remains challenging, particularly for anchoring, where numerical triggers are difficult to distinguish from factual content. Rewriting effectively removes most bias type triggers, but fails for anchoring. Rewriting consistently outperforms instruction-based defenses, while chain-of-thought prompting degrades performance. The small gap between using predicted or gold bias type labels indicates that the defense mechanism itself determines effectiveness, not bias detection accuracy. Future work should address anchoring as the most resistant bias type, explore naturally occurring triggers, and extend beyond English and sentiment classification.

Limitations

Our study has several limitations. First, while we evaluate a diverse set of models spanning multiple families, we cannot cover all available LLMs, and other models may exhibit different patterns. Second, because our bias triggers are synthetically generated, they may not capture the full lexical and structural diversity of bias triggers as they naturally occur in financial journalism. Future work should validate these findings with naturally occurring bias triggers, though such evaluation would require manual annotation of real financial corpora. Third, our analysis is limited to English text and a single modality; financial sentiment expressed in other languages or through other modalities (e.g., audio, tables) may yield different results. Furthermore, our source sentences are drawn from FinancialPhraseBank, which covers companies listed on the OMX Helsinki exchange, and the findings may not generalize to other markets or company populations. Fourth, due to computational constraints, models exceeding 9B parameters were run using 4-bit quantization, which may affect model behavior compared to full-precision inference. Fifth, all results are from a single run with greedy decoding (zero temperature). While this yields deterministic outputs for open-source models, closed-source API calls may exhibit minor non-determinism even at temperature 0; we did not quantify this variance due to budget constraints. Finally, our findings are empirical observations of model behavior; we do not make causal claims about why these biases emerge.

Ethical Statement

All datasets used are publicly available. FinancialPhraseBank (Malo et al., 2013) is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license. Upon acceptance, we will release TriggerBank, along with the code used to generate and evaluate it, under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license, which is compatible with the source dataset's terms. The source sentences are drawn from publicly available English-language financial news covering companies listed on the OMX Helsinki stock exchange and do not contain personally identifying information or offensive content; we did not perform additional anonymization. TriggerBank is intended for research purposes only, and derivative uses should remain within research contexts. For the LLMs evaluated in this study, we accessed closed-source models under their respective providers' terms of service, and we used open-source models in accordance with their respective community licenses. We note that our findings could potentially be misused to exploit model biases in financial applications; however, we believe transparency about these biases is preferable to obscurity. This work is intended solely as research and does not constitute investment advice. We do not recommend any specific model for financial decision-making.

References

Abeer Alessa, Param Somane, Akshaya Thenkarai Lakshminarasimhan, Julian Skirzynski, Julian McAuley, and Jessica Maria Echterhoff. 2025. Quantifying cognitive bias induction in LLM-generated content. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2890–2910, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Anonymous. 2025. Cognibias: A benchmark for cognitive biases in AI–human dialogue. In *EurIPS 2025 Workshop - Unifying Perspectives on Learning Biases*.

Anthropic. 2025. Claude haiku 4.5. https://www.anthropic.com/claude/haiku.

Nicholas Barberis and Richard Thaler. 2002. A survey of behavioral finance. Working Paper 9222, National Bureau of Economic Research.

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

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *SIGKDD Explor. Newsl.*, 26(1):34–48.

Fabrizio Dimino, Krati Saxena, Bhaskarjit Sarmah, and Stefano Pasquali. 2025. Uncovering representation bias for investment decisions in open-source large language models. *Preprint*, arXiv:2510.05702.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

JOSEPH E. Engelberg and CHRISTOPHER A. Parsons. 2011. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Giorgos Filandrianos, Angeliki Dimitriou, Maria Lymperaiou, Konstantinos Thomas, and Giorgos Stamou. 2025. Bias beware: The impact of cognitive biases on LLM-driven product recommendations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22397–22426, Suzhou, China. Association for Computational Linguistics.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Google DeepMind. 2026. Gemini 3.1 flash-preview. `https://deepmind.google/models/gemini/flash/`.

Google Team. 2026. Gemma 4: Byte for byte, the most capable open models.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Yuechen Jiang, Zhiwei Liu, Yupeng Cao, Yueru He, Ziyang Xu, Chen Xu, Zhiyang Deng, Prayag Tiwari, Xi Chen, Alejandro Lopez-Lira, Jimin Huang, Junichi Tsujii, and Sophia Ananiadou. 2026. All that glisters is not gold: A benchmark for reference-free counterfactual financial misinformation detection. *Preprint*, arXiv:2601.04160.

Hoyoung Lee, Junhyuk Seo, Suhwan Park, Junhyeong Lee, Wonbin Ahn, Chanyeol Choi, Alejandro Lopez-Lira, and Yongjae Lee. 2025. *Your AI, Not Your View: The Bias of LLMs in Investment Analysis*, page 150–158. Association for Computing Machinery, New York, NY, USA.

Zhiwei Liu, Yupen Cao, Yuechen Jiang, Mohsinul Kabir, Polydoros Giannouris, Chen Xu, Ziyang Xu, Tianlei Zhu, Tariquzzaman Faisal, Triantafillos Papadopoulos, Yan Wang, Lingfei Qian, Xueqing Peng, Zhuohan Xie, Ye Yuan, Saeed Almheiri, Abdulrazzaq Alnajjar, Mingbin Chen, Harry Stuart, and 6 others. 2026. Same claim, different judgment: Benchmarking scenario-induced bias in multilingual financial misinformation detection. *Preprint*, arXiv:2601.05403.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Yougang Lyu, Shijie Ren, Yue Feng, Zihan Wang, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2025. Self-adaptive cognitive debiasing for large language models in decision-making. *Preprint*, arXiv:2504.04141.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. Good debt or bad debt: Detecting semantic orientations in economic texts. *Preprint*, arXiv:1307.5336.

Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. 2024. Evaluating company-specific biases in financial sentiment analysis using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, page 6614–6623. IEEE.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. *Preprint*, arXiv:2408.02784.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Dimitris Vamvourellis and Dhagash Mehta. 2025. *Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis*, page 299–307. Association for Computing Machinery, New York, NY, USA.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. Finben: A holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2025b. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Yuhan Zhi, Xiaoyu Zhang, Longtian Wang, Shumin Jiang, Shiqing Ma, Xiaohong Guan, and Chao Shen. 2025. Exposing product bias in llm investment recommendation. *Preprint*, arXiv:2503.08750.

Yuhang Zhou, Yuchen Ni, Zhiheng Xi, Zhangyue Yin, Yu He, Gan Yunhui, Xiang Liu, Zhang Jian, Sen Liu, Xipeng Qiu, Yixin Cao, Guangnan Ye, and Hongfeng Chai. 2025. Are LLMs rational investors? a study on the financial bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24139–24173, Vienna, Austria. Association for Computational Linguistics.

A Annotation Guidelines

A.1 Base Sentence Selection Guidelines

Annotators were provided with sentences from the 100% agreement subset of FinancialPhraseBank and asked to verify that each sentence is suitable as a clean base sentence for bias injection experiments. For each sentence, annotators marked 1 (keep) or 0 (exclude).

Inclusion criteria. A sentence was kept if it was factual and objective, reporting financial data, events, or outcomes without editorializing, and did not contain any of the six bias trigger types defined in Table 1.

Exclusion examples by bias type.

- **Authority:** Mentions of analysts, banks, or rating agencies expressing an opinion or making a recommendation (e.g., "Goldman Sachs upgraded the stock to Buy"). Mentioning a company doing business with a bank is acceptable (e.g., "The loan was arranged by Deutsche Bank").
- **Anchoring:** References to historical prices, past valuations, or price targets that could serve as psychological anchors (e.g., "The stock, which once traded at $200..."). Direct period-over-period comparisons that are part of reporting results are acceptable (e.g., "Revenue rose to EUR 50mn from EUR 40mn last year").
- **Framing:** Presenting a fact in a way that is clearly spun positive or negative when a more neutral description exists (e.g., "The company managed to retain 85% of its revenue").
- **Herding:** Mentions of what other investors, funds, or market participants are doing (e.g., "Institutional investors have increased their positions by 15%").
- **Recency:** Strong temporal emphasis that creates artificial urgency beyond normal time references (e.g., "The company JUST announced..."). Normal time references like "in Q3 2010" are acceptable.
- **Overconfidence:** Hyperbolic certainty language or exaggerated projections (e.g., "Analysts are certain the company will dominate its sector").

Annotators were instructed to exclude borderline cases and work independently without discussing individual sentences.

A.2 Bias Injection Validation Guidelines

Two annotators independently evaluated 210 randomly sampled bias-injected sentences. For each sentence, annotators were shown the original clean sentence, the biased sentence, the sentiment label,

and the injected bias type(s). Annotators evaluated four criteria using binary labels (yes/no):

1. **Bias identifiable:** Can the injected bias trigger(s) be clearly identified in the biased sentence? For pairwise biases, both triggers must be identifiable.

2. **Natural:** Does the biased sentence sound like real financial journalism?

3. **Facts preserved:** Are all original facts from the source sentence intact? Annotators compared numbers, dates, company names, and key claims between the original and biased sentences.

4. **Other problem:** Is there any other issue not captured by the three criteria above? Annotators were asked to explain any such issues in a notes column.

A.3 Bias Trigger Removal Evaluation Guidelines

Two annotators independently evaluated a subset of 198 model rewrites. For each rewrite, annotators were shown five pieces of information: (1) the original clean sentence, (2) the biased sentence, (3) the ground truth bias types (for context only), (4) the bias types the model was told to remove (detected biases), and (5) the model's rewrite. Annotators evaluated three dimensions using binary labels (1/0):

1. **Trigger removal:** Whether the bias triggers corresponding to the detected bias types have been removed from the rewrite. Annotators compared the rewrite against the original sentence to identify which elements were injected. If a detected bias type was not actually present in the biased sentence (false positive detection), trigger removal was marked as satisfied, since there was no trigger to remove. If a real bias was missed by detection and remained in the rewrite, trigger removal was still marked as satisfied, since the model was never told to remove it.

2. **Factual preservation:** Whether all financial figures, entity names, time periods, and factual claims from the original sentence are present and accurate in the rewrite. Different wording for the same fact (e.g., "increased" instead of "grew") was acceptable. Omission of content added by the bias injection was acceptable.

3. **Linguistic naturalness:** Whether the rewrite reads like natural financial reporting. Rewrites were marked as unsatisfactory if they contained awkward phrasing, grammatical errors, contradictory statements, or language uncharacteristic of financial news.

The same rubric was used for the LLM-as-judge evaluation of the full dataset. Full LLM judge prompts are provided in Appendix B.5.

A.4 Annotator Demography

All annotation was performed by two annotators with complementary expertise in NLP and finance. Annotator 1 is a PhD candidate in Computer Science, with a background in mathematics and statistics, and industry experience as a credit risk analyst. This combination provides expertise in both the computational and financial aspects of the annotation task. Annotator 2 is a PhD candidate in Finance, with a background in finance and computer science, and industry experience as a finance professional. Their financial domain knowledge is essential for evaluating the plausibility and naturalness of bias triggers in financial text. Both annotators are co-authors of this study and received no separate compensation for the annotation work. No external human subjects were involved in this research; consequently, ethics review board approval was not sought, as the annotation tasks were performed by co-authors on publicly available financial text.

As co-authors, the annotators have a vested interest in the study's success, which could introduce bias toward favorable quality judgments. To mitigate this, all annotation was performed independently using predefined guidelines (Appendices A.1, A.2, and A.3), and inter-annotator agreement is reported transparently.

B Prompt Templates

B.1 Bias Injection Prompts

We use one system prompt shared across all bias injection tasks and one user prompt per bias type (seven user prompts in total, as framing uses separate templates for positive-to-negative and negative-to-positive reframing). For pairwise combinations, injection is performed sequentially using the same templates.

System prompt.

> You are an expert at naturally integrating cognitive bias triggers into financial text.
>
> Rules: 1. Preserve ALL original facts, numbers, company names, dates, and sentence subjects exactly — do not remove or change them. 2. Integrate the bias trigger naturally into the sentence flow — it should read as if originally written that way. 3. The output must sound like professional financial journalism — concise and natural. 4. The injected bias must not add directional information that would justify a sentiment change. 5. CRITICAL: Any historical reference must use a year EARLIER than the year in the sentence. 6. If the sentence already contains a bias trigger, integrate the new one seamlessly without making the sentence feel overstuffed.
>
> Output ONLY the rewritten sentence.

User prompts. Each user prompt instructs the model to inject one bias type. The placeholder `{sentence}` is replaced with the input sentence.

> **Anchoring:** Rewrite this sentence to naturally include ANCHORING bias via a historical numerical reference. Weave in a concrete historical number (past price, previous earnings, old metric). The anchor year must be EARLIER than any year in the sentence. The anchor should feel like a natural part of the sentence, not tacked on. Do NOT duplicate numbers already present. Do NOT add interpretive language. Preserve all original facts exactly.
>
> `Sentence: {sentence}`

> **Authority:** Rewrite this sentence to naturally include AUTHORITY bias via an institutional reference. Weave in a reference to: Goldman Sachs, JP Morgan, Morgan Stanley, UBS, Barclays, Moody's, S&P, Fitch, Bloomberg, Reuters, or Financial Times. Use direct attribution only: "according to", "cited by", "data from", "per", "as reported by". The attribution should feel like a natural part of the sentence. Do NOT add directional information. Preserve all original facts exactly.
>
> `Sentence: {sentence}`

> **Herding:** Rewrite this sentence to naturally include HERDING bias via a reference to market participants. Weave in a reference to other investors or institutions acting. Vary phrasing across examples (e.g., "amid broader institutional repositioning", "as market participants adjusted exposure", "with peers making similar moves"). Do NOT use directional language. Preserve all original facts exactly.
>
> `Sentence: {sentence}`

> **Recency:** Rewrite this sentence to naturally include RECENCY bias via temporal emphasis. Weave in language emphasizing how recent the information is. Use only ONE recency marker. Do NOT combine awkwardly with existing temporal phrases. Do NOT add interpretive words. Preserve all original facts exactly.
>
> `Sentence: {sentence}`

> **Overconfidence:** Rewrite this sentence to naturally include OVERCONFIDENCE bias via certainty language. Weave in excessive certainty about future outcomes. Vary phrasing (e.g., "all but guaranteed to persist", "expected to hold", "set to remain", "poised to continue"). Express certainty about continuation or persistence, not direction. Preserve all original facts exactly.
>
> `Sentence: {sentence}`

> **Framing (negative):** Rewrite this sentence with a NEGATIVE spin while preserving all facts exactly. Integrate negative tone naturally using varied vocabulary: "only", "merely", "just", "modest", "slim", "disappointing", "lackluster". The spin should feel like the original author's perspective. Do NOT add new judgments or context.
>
> `Sentence: {sentence}`

> **Framing (positive):** Rewrite this sentence with a POSITIVE spin while preserving all facts exactly. For negative facts, naturally minimize magnitude or reframe direction. Use varied vocabulary: "solid", "steady", "contained", "limited", "manageable", "held at", "stabilized at". The spin should feel like the original author's perspective. Use ONLY tone words — no added context or rationalizations.
>
> `Sentence: {sentence}`

B.2 Bias Detection Prompt

System prompt.

> You are an expert at detecting cognitive bias triggers in financial text. Detect if any of these 6 bias triggers are present: 1. Anchoring: A historical numerical reference point that could influence perception of current or reported figures. 2. Authority: Attribution to a respected institution, analyst, or financial data source. 3. Framing: Presenting facts in a way that emphasizes gains over losses or vice versa, without changing the underlying information. 4. Herding: References to other market participants' actions or positioning. 5. Recency: Temporal emphasis highlighting how recent or fresh the information is, beyond ordinary time references. 6. Overconfidence: Excessive certainty language about future outcomes.
>
> Rules: - Multiple biases can be present in one sentence. - Sentences may not contain any bias triggers at all. If no bias triggers are detected, output: none. - Output ONLY a comma-separated list of detected biases (in lowercase), or the word "none" if zero biases are present.

User prompt.

> Analyze this sentence for cognitive bias triggers:
>
> `{sentence}`
>
> Answer:

B.3 Bias Rewriting Prompt

System prompt.

> You are a financial text editor. Your task is to rewrite financial sentences to remove specific cognitive bias triggers while preserving all factual information. The bias types are defined as follows: - Anchoring: a numerical reference point that could influence perception of current or reported figures. - Authority: attribution to a respected institution, analyst, or financial data source. - Framing: presenting facts in a way that emphasizes gains over losses or vice versa, without changing the underlying information. - Herding: references to other market participants' actions or positioning. - Recency: temporal emphasis highlighting how recent or fresh the information is, beyond ordinary time references. - Overconfidence: excessive certainty language about future outcomes.
>
> Output ONLY the rewritten sentence, nothing else.

User prompt.

> The following financial sentence contains `{predicted_biases}` bias. Rewrite the sentence to remove the identified bias triggers. Preserve all factual content and financial figures exactly. Do not add any new information or opinions.
>
> `{sentence}`
>
> Rewritten sentence:

For the diagnostic rewriting pass using ground truth labels (RQ3 setup 8), the same prompts are used with `{predicted_biases}` replaced by the ground truth bias labels.

B.4 Defense Prompts

All defense setups share the same system prompt. User prompts vary by setup.

System prompt (all setups).

> You are a financial sentiment classifier. Given a financial sentence, classify its sentiment based on how the information would likely affect the company's stock price. Output ONLY one word: positive, negative, or neutral.

Setup 1 (clean baseline) and Setup 2 (no defense).

> Classify the sentiment of this financial sentence:
>
> `{sentence}`
>
> Sentiment:

Setup 1 uses original unbiased sentences as input. Setup 2 uses biased sentences.

Setup 3 (generic defense).

> The following financial sentence may contain cognitive bias triggers including anchoring, which involves numerical reference points that influence perception; authority, which involves attribution to respected institutions; framing, which involves emphasizing gains or losses; herding, which involves references to other investors' actions; recency, which involves temporal emphasis; and overconfidence, which involves excessive certainty about outcomes. Ignore any such biases and classify the sentiment based on the financial facts only.
>
> `{sentence}`
>
> Sentiment:

Setup 4 (chain-of-thought).

> The following financial sentence may contain cognitive bias triggers including anchoring, which involves numerical reference points that influence perception; authority, which involves attribution to respected institutions; framing, which involves emphasizing gains or losses; herding, which involves references to other investors' actions; recency, which involves temporal emphasis; and overconfidence, which involves excessive certainty about outcomes. Classify the sentiment of the sentence. Think step by step, then on the last line write your final answer as: FINAL ANSWER: positive, negative, or neutral.
>
> `{sentence}`
>
> Analysis:

Setups 5 and 6 (targeted defense with predicted/gold labels).

> The following financial sentence contains `{bias_type}` bias. The six bias types are anchoring, which involves numerical reference points that influence perception; authority, which involves attribution to respected institutions; framing, which involves emphasizing gains or losses; herding, which involves references to other investors' actions; recency, which involves temporal emphasis; and overconfidence, which involves excessive certainty about outcomes. Ignore the identified bias and focus only on the financial facts. Classify the sentiment:
>
> `{sentence}`
>
> Sentiment:

Setup 5 uses model-detected bias labels. Setup 6 uses ground truth bias labels.

Setups 7 and 8 (rewrite then classify). These setups use the same prompt as Setups 1 and 2, applied to the rewritten sentences from Experiment 2. Setup 7 uses rewrites guided by detected bias labels. Setup 8 uses rewrites guided by ground truth bias labels.

B.5 LLM-as-Judge Prompt

System prompt.

> You are evaluating the quality of financial sentence rewrites that were intended to remove specific cognitive bias triggers while preserving factual content.

The six bias trigger types are: - Anchoring: a numerical reference point that could influence perception of current or reported figures. - Authority: attribution to a respected institution, analyst, or financial data source. - Framing: presenting facts in a way that emphasizes gains over losses or vice versa, without changing the underlying information. - Herding: references to other market participants' actions or positioning. - Recency: temporal emphasis highlighting how recent or fresh the information is, beyond ordinary time references. - Overconfidence: excessive certainty language about future outcomes.

Evaluate the rewrite on three dimensions by comparing against the original sentence. If the model was told to remove a bias type that is not actually present in the biased sentence, mark trigger removal as 1 since there is nothing to remove.

Output ONLY valid JSON with three keys and 0 or 1 values, nothing else.

User prompt. The placeholders `{original_sentence}`, `{biased_sentence}`, `{predicted_biases}`, and `{debiased_sentence}` are replaced with the corresponding values for each rewrite.

Original sentence: `{original_sentence}`

Biased sentence: `{biased_sentence}`

Bias types the model was told to remove: `{predicted_biases}`

Rewritten sentence: `{debiased_sentence}`

Evaluate the rewrite:

1. trigger_removed: 1 if the bias triggers corresponding to the detected bias types are no longer present in the rewrite, 0 if any remain. Compare against the original to determine what was added. If a detected bias type has no corresponding trigger in the biased sentence, mark 1. If a bias trigger is present in the rewrite but was not among the detected bias types, do not penalize.

2. facts_preserved: 1 if all financial figures, entities, time periods, and factual claims from the original sentence are preserved in the rewrite, 0 if any are missing or altered.

3. natural_language: 1 if the rewrite reads like natural financial reporting, 0 if it contains awkward phrasing or unnatural language.

JSON:

C Post-Processing Details

After bias injection, we applied automated post-processing to increase lexical diversity in the generated bias triggers. Zero-shot generation led the injection model to favor certain phrasings repeatedly. Replacements were applied probabilistically (70% replacement rate for authority, herding, and overconfidence; 30% for framing) to maintain a mix of original and diversified phrasings. Recency and anchoring triggers were not post-processed as they already exhibited sufficient lexical diversity. All replacement phrases were verified to preserve the nature of the bias trigger without introducing factual content that could affect sentiment.

Authority. Both the attribution phrase and the source institution were varied. Attribution phrases included "according to", "as reported by", "per data from", "cited by", "in figures compiled by", "as noted by", and "tracked by". Source institutions included Goldman Sachs, JP Morgan, Morgan Stanley, UBS, Barclays, Moody's, S&P, Fitch, Bloomberg, Reuters, and Financial Times.

Herding. Repetitive phrases such as "as market participants adjusted exposure" and "amid broader institutional repositioning" were replaced with alternatives including "as peers repositioned", "amid wider portfolio shifts", "as institutions reallocated", "with funds adjusting positions", "amid parallel moves by investors", "as peers adjusted holdings", "with institutions shifting exposure", "amid sector-wide repositioning", "as portfolios shifted", and "with peers rebalancing".

Overconfidence. Certainty phrases such as "a pattern showing no signs of changing", "expected to hold", "set to remain", and "poised to continue" were replaced with alternatives including "likely to persist", "appearing entrenched", "unlikely to shift", "showing persistence", "appearing stable", "likely to hold", "seeming fixed", "appearing set", "unlikely to change", and "showing stability".

Framing. Spin words were occasionally replaced (30% probability) since the original vocabulary was already reasonably varied. Replacements included "a modest" → "a slim / a limited / a small", "a mere" → "a slim / a bare / a small", and "a solid" → "a stable / a steady / a maintained".

D Model Configurations

Table 7 summarizes the model versions and configurations used across all experiments. All models were run with greedy decoding (zero temperature) and internal reasoning features disabled where applicable. Open-source models were loaded via the HuggingFace Transformers library. Models exceeding 9B parameters were loaded in pre-quantized 4-bit precision from Unsloth repositories using bitsandbytes to fit within available GPU memory. Experiments were conducted on three NVIDIA Quadro RTX 6000 GPUs, with the total com-

pute budget amounting to approximately 1,000 GPU-hours in aggregate across all reported and exploratory experiments. We do not perform hyperparameter search; all models use greedy decoding (temperature 0) with default settings otherwise.

E Evaluation Implementation

Classification metrics (precision, recall, F1, accuracy) were computed using standard scientific Python libraries (NumPy, scikit-learn).

F LLM Judge Validation

Tables 8 and 9 report inter-annotator agreement between the two human annotators, and agreement between each candidate LLM judge and the human annotators on the 198-rewrite validation subset, respectively. We report raw agreement and Gwet's AC1 (Gwet, 2008). Gemma 4 31B was selected as the final judge based on substantially higher agreement with both annotators on trigger removal (0.912 average) compared to Qwen3 32B (0.766 average), which tended to mark nearly all rewrites as successful.

Dimension	Agr. (%)	AC1
Trigger removal	94.4	0.913
Factual preserv.	87.9	0.858
Naturalness	98.0	0.979

Table 8: Inter-annotator agreement between the two human annotators on the 198-rewrite validation subset.

		A1		A2		Avg	
Judge	Dim.	Agr.	AC1	Agr.	AC1	Agr.	AC1
Gemma	Trig.	.914	.868	.909	.861	**.912**	**.865**
	Facts	.955	.950	.864	.837	.910	.894
	Nat.	.965	.963	.985	.985	.975	.974
Qwen	Trig.	.763	.700	.768	.708	.766	.704
	Facts	.980	.979	.889	.872	**.935**	**.926**
	Nat.	.975	.974	.995	.995	**.985**	**.985**

Table 9: Agreement between candidate LLM judges and each human annotator on the 198-rewrite validation subset. Gemma 4 31B was selected based on higher average agreement, particularly on trigger removal.

G Full Detection Results

G.1 Per-Bias-Type F1 (All Models)

Table 10 reports F1 scores per bias type for all evaluated models. Table 11 additionally provides precision, recall, and F1 for the none class specifically, to clarify the source of the low detection scores on clean text.

Model	P	R	F1
GPT-5.4-mini	0.333	0.067	0.111
Claude Haiku 4.5	0.375	0.400	0.387
Gemini 3.1 Flash	0.201	0.447	0.277
Llama 3.1 8B	0.400	0.013	0.026
Llama 3.3 70B	0.250	0.347	0.291
Qwen3 8B	0.119	0.667	0.202
Qwen3.5 9B	0.191	**0.740**	0.303
Qwen3 32B	0.338	0.360	0.348
Gemma 4 31B	**0.483**	0.487	**0.485**

Table 11: Precision, recall, and F1 for the none class (150 base sentences with no bias triggers).

G.2 Single-Bias vs Pairwise-Bias Detection

Table 12 compares macro F1 on single-bias sentences (one trigger) against pairwise-bias sentences (two co-occurring triggers) for each model.

Model	Single	Pairwise
Closed-source		
GPT-5.4-mini	0.592	0.633
Claude Haiku 4.5	0.591	0.635
Gemini 3.1 Flash	0.512	0.488
Open-source		
Llama 3.1 8B	0.427	0.468
Llama 3.3 70B	0.422	0.475
Qwen3 8B	0.371	0.419
Qwen3.5 9B	0.450	0.533
Qwen3 32B	0.490	0.528
Gemma 4 31B	**0.653**	**0.687**

Table 12: Macro F1 on single-bias versus pairwise-bias subsets of TriggerBank for each model.

G.3 Pairwise Combination Results

Table 13 reports F1 per pairwise bias combination for each model. Combinations involving framing are evaluated only on positive and negative sentences, as explained in (§3.1).

H Dataset Examples

Table 14 shows representative examples of single-bias injections from TriggerBank, one per bias type. Table 15 shows representative examples of pairwise bias injections. In each case, the injected bias trigger is <u>underlined</u> for clarity.

I Flagged Annotation Examples

Table 16 shows representative examples of sentences where at least one annotator identified an is-

Model	Version / Checkpoint	Precision
Closed-source		
GPT-5.4-mini	gpt-5.4-mini-2026-03-17	API
Claude Haiku 4.5	claude-haiku-4-5-20251001	API
Gemini 3.1 Flash	gemini-3.1-flash-preview (no snapshot)	API
Open-source		
Llama 3.1 8B	unsloth/Llama-3.1-8B-Instruct	Full
Llama 3.3 70B	unsloth/Llama-3.3-70B-Instruct-bnb-4bit	4-bit (bnb)
Qwen3 8B	unsloth/Qwen3-8B	Full
Qwen3.5 9B	unsloth/Qwen3.5-9B	Full
Qwen3 32B	unsloth/Qwen3-32B-bnb-4bit	4-bit (bnb)
Gemma 4 31B	unsloth/gemma-4-31B-it-unsloth-bnb-4bit	4-bit (bnb)

Table 7: Model versions and precision settings used across all experiments.

Bias Type	GPT-5.4m	Claude H.	Gemini F.	Llama 8B	Llama 70B	Qwen3 8B	Qwen3.5 9B	Qwen3 32B	Gemma 4
Anchoring	0.493	0.578	0.523	0.442	0.494	**0.582**	0.542	0.504	0.565
Authority	0.941	0.946	0.833	0.932	0.908	0.851	0.804	**0.960**	0.959
Framing	0.540	0.585	0.647	0.369	0.456	0.291	0.326	0.333	**0.660**
Herding	0.898	0.863	0.578	0.773	0.700	0.456	0.762	0.825	**0.918**
Recency	0.837	0.701	0.545	0.480	0.598	0.452	0.562	0.625	**0.874**
Overconfidence	0.598	0.673	0.286	0.150	0.050	0.213	0.606	0.342	**0.750**
None	0.111	0.387	0.277	0.026	0.291	0.202	0.303	0.348	**0.485**

Table 10: F1 score per bias type for all models on TriggerBank.

sue during injection validation. Flagged issues predominantly concerned naturalness and other problems, while bias identifiability was rarely flagged and factual preservation had perfect agreement across all 210 sentences.

J Rewriting Error Examples

Table 17 shows representative examples of rewriting errors across bias types, illustrating common failure modes.

K Full Rewriting Results

Table 18 reports rewriting quality per bias type, as scored by the Gemma 4 31B LLM judge. Results are shown separately for single-bias and pairwise-bias sentences.

L Full Defense Results

Tables 19 and 20 report sentiment classification accuracy broken down by bias type for single-bias and pairwise-bias subsets of TriggerBank, respectively.

Combination	GPT-5.4m	Claude H.	Gemini F.	Llama 8B	Llama 70B	Qwen3 8B	Qwen3.5 9B	Qwen3 32B	Gemma 4
Anch.+Auth.	**0.286**	0.279	0.253	0.285	**0.286**	0.264	**0.286**	0.285	**0.286**
Anch.+Fram.	0.257	0.250	0.205	**0.284**	0.268	0.233	0.252	0.213	0.241
Anch.+Herd.	**0.272**	0.260	0.210	0.241	0.244	0.189	0.255	0.252	**0.272**
Anch.+Overc.	0.225	0.229	0.184	0.161	0.147	0.172	0.222	0.188	**0.242**
Anch.+Rec.	0.267	0.247	0.200	0.224	0.229	0.272	0.263	**0.283**	0.276
Auth.+Fram.	0.241	0.229	0.219	**0.256**	0.186	0.195	0.252	0.188	0.250
Auth.+Herd.	0.272	0.268	0.160	0.249	0.236	0.194	0.264	0.252	**0.274**
Auth.+Overc.	0.236	0.237	0.146	0.156	0.145	0.139	0.239	0.185	**0.251**
Auth.+Rec.	0.263	0.245	0.127	0.206	0.182	0.200	0.245	0.268	**0.270**
Fram.+Herd.	0.219	0.232	0.107	0.198	0.179	0.145	0.194	0.161	**0.240**
Fram.+Overc.	0.209	0.227	**0.241**	0.158	0.144	0.134	0.212	0.137	0.215
Fram.+Rec.	0.225	0.211	**0.229**	0.191	0.201	0.147	0.165	0.171	0.221
Herd.+Overc.	0.190	0.204	0.148	0.123	0.091	0.073	0.168	0.136	**0.219**
Herd.+Rec.	0.260	0.254	0.181	0.203	0.215	0.200	0.235	0.264	**0.274**
Overc.+Rec.	0.208	0.210	0.138	0.124	0.126	0.125	0.200	0.190	**0.245**

Table 13: F1 per pairwise bias combination for all models on TriggerBank.

Bias Type	Original	Biased
Anchoring	Vacon's cash flow from operations grew to EUR 37.1 mn from EUR 21.9 mn a year ago.	Vacon's cash flow from operations grew to EUR 37.1 mn from EUR 21.9 mn a year ago, above the EUR 18.4 mn recorded in 2008.
Authority	The total project duration is three years and it is valued at some EUR11.5 m.	The total project duration is three years and it is valued at some EUR11.5 m, according to Reuters.
Framing	Ruukki Group calculates that it has lost EUR 4mn in the failed project.	Ruukki Group calculates that it has contained its loss at EUR 4mn in the failed project.
Herding	Diluted earnings per share (EPS) rose to EUR 0.52 versus EUR 0.09.	Diluted earnings per share (EPS) rose to EUR 0.52 versus EUR 0.09, with institutions shifting exposure.
Recency	The forecast for 2012 is 3.3%.	The forecast for 2012 is, most recently, 3.3%.
Overconfidence	Konecranes Oyj KCR1V FH fell 5.5 percent to 20.51 euros, the biggest fall since June.	Konecranes Oyj KCR1V FH fell 5.5 percent to 20.51 euros, the biggest fall since June, a pattern showing no signs of changing.

Table 14: Representative single-bias injection examples from TriggerBank. Injected bias triggers are underlined.

Combination	Original	Biased
Framing + Herding	In Q1 of 2009, the company's operating loss totalled EUR 0.3 mn, compared to a profit of EUR 3.6 mn in Q1 of 2008.	In Q1 of 2009, the company's operating loss totalled a contained EUR 0.3 mn, as market participants adjusted exposure, compared to a profit of EUR 3.6 mn in Q1 of 2008.
Authority + Recency	Its board of directors will propose a dividend of EUR0.12 per share for 2010, up from the EUR0.08 per share paid in 2009.	Its board of directors will propose, according to JP Morgan, a dividend of EUR0.12 per share for 2010, most recently up from the EUR0.08 per share paid in 2009.
Anchoring + Overconfidence	The Baltimore Police and Fire Pension, which has about $1.5 billion, lost about $3.5 million in Madoff Ponzi scheme.	The Baltimore Police and Fire Pension, which has about $1.5 billion and oversaw roughly $1.2 billion in 2007, lost about $3.5 million in Madoff Ponzi scheme, a pattern of exposure appearing entrenched.

Table 15: Representative pairwise bias injection examples from TriggerBank. Each injected trigger is underlined separately.

Biased Sentence	Bias Type	Dimension	A1	A2
Earnings per share EPS amounted to a recent EUR0.01.	recency	identifiability	1	0
Cerberus Capital Management LP-backed printing paper maker NewPage Corp. has posted mixed second-quarter results, placing a limited cloud over its planned initial public offering.	framing	identifiability	0	1
Viking Line's cargo revenue increased by 5.4% to EUR 21.46 mn, as market participants adjusted exposure, and cargo volume increased by 2.4% to 70,116 cargo units.	herding	naturalness	0	1
Last year the company recently raised its turnover to approximately 7 million litas EUR 2 mln, from 6.1 million litas in 2004.	framing, recency	naturalness	0	0
The total investment in the Vantaa plant extension will amount to around 10 million euro, and Okmetic's share of the investments will recently be worth around 2.7 million.	recency	naturalness	0	1
It generated an operating loss of EUR 96.3 mn, according to Bloomberg, against an operating profit of EUR 28.7 mn in 2008, down from a profit of EUR 43.9 mn.	anchoring, authority	naturalness	0	1
Sampo Bank's market share of lending was 13.6%, as market participants adjusted exposure, down from 14.4% in the first quarter of 2008 and below the 15.1% recorded in 2007.	anchoring, herding	other	1	0
Operating profit margin was a solid 8.3%, compared to 11.8% a year earlier.	framing	identifiability	1	0

Table 16: Representative sentences flagged during injection validation. Underlined text highlights the injected trigger. A1 and A2 denote annotator judgments (1 = satisfied, 0 = not satisfied) on the flagged dimension.

Bias Type	Error Type	Biased Sentence	Model Rewrite
Anchoring	Trigger retained	Earnings per share amounted to EUR0.01, after standing at EUR0.04 in 2008.	Earnings per share amounted to EUR0.01, compared with EUR0.04 in 2008.
Herding	Trigger retained	Operating loss totalled EUR 0.9 mn, as market participants recently adjusted exposure, down from a profit of EUR 2.7 mn.	Operating loss totaled EUR 0.9 mn, as exposure was adjusted, down from a profit of EUR 2.7 mn.
Framing	Fact damaged + Trigger retained	The Baltimore Police and Fire Pension, which has about $1.5 billion, saw a limited loss of about $3.5 million in Madoff Ponzi scheme.	The Baltimore Police and Fire Pension saw a limited loss of about $3.5 million in the Madoff Ponzi scheme. ($1.5 billion omitted)
Overconfidence	Fact damaged	Both operating profit and turnover increased, respectively from EUR0.9m and EUR8.3m, as compared to 2005, a pattern showing no signs of changing.	Both operating profit and turnover increased to EUR0.9m and EUR8.3m compared to 2005. ("from" changed to "to")
Recency	Unnatural	Vacon's cash flow from operations recently grew to EUR 37.1 mn from EUR 21.9 mn a year ago, amid sector-wide repositioning.	Vacon's cash flow from operations increased to EUR 37.1 mn from EUR 21.9 mn, reflecting sector-wide repositioning. (herding trigger retained)
Authority	Unnatural	Production capacity will increase from 36,000 to 85,000 tonnes per year, according to Reuters, and the raw material is showing persistence recycled paper and board.	Production capacity will increase from 36,000 to 85,000 tonnes per year, and the raw material is showing persistence in recycled paper and board. (ungrammatical phrasing retained)

Table 17: Representative rewriting errors across bias types. Bias triggers are underlined. Parenthetical notes describe the error.

Bias Type	Subset	GPT-5.4m	Claude H.	Gemini F.	Llama 8B	Llama 70B	Qwen3 8B	Qwen3.5 9B	Qwen3 32B	Gemma 4
Trigger Removal (%)										
Anchoring	Single	44.7	36.7	99.3	8.7	10.0	24.0	18.0	**48.7**	45.3
	Pairwise	32.2	35.5	99.1	23.0	22.1	24.4	23.8	47.9	**52.1**
Authority	Single	98.7	**99.3**	100.0	93.3	97.3	97.3	96.7	98.7	**99.3**
	Pairwise	73.6	76.6	99.4	74.0	76.6	67.6	57.2	79.2	**86.3**
Framing	Single	95.0	94.1	99.0	93.1	88.1	**98.0**	94.1	94.1	96.0
	Pairwise	82.0	84.6	99.4	80.2	83.4	74.5	74.7	83.4	**86.7**
Herding	Single	86.7	91.3	98.7	84.7	89.3	86.0	75.3	**94.7**	**94.7**
	Pairwise	69.2	74.5	98.6	67.8	76.3	64.9	50.5	81.2	**84.2**
Recency	Single	92.0	89.3	99.3	86.0	82.7	**96.0**	93.3	94.0	94.7
	Pairwise	69.8	76.9	99.3	72.3	75.2	72.9	69.5	81.3	**83.3**
Overconfidence	Single	84.0	96.7	98.0	89.3	94.7	94.0	**98.0**	95.3	90.7
	Pairwise	71.0	81.3	98.9	69.8	78.6	71.2	68.2	**84.6**	83.0
Factual Preservation (%)										
Anchoring	Single	90.7	86.7	2.0	88.7	88.7	88.7	92.0	88.7	**95.3**
	Pairwise	93.6	93.2	1.6	91.0	94.2	91.0	93.9	92.3	**95.9**
Authority	Single	96.0	94.7	9.3	96.0	94.7	96.7	**100.0**	94.7	97.3
	Pairwise	97.1	95.6	3.7	95.9	95.7	95.0	**98.3**	93.9	98.0
Framing	Single	94.1	91.1	4.0	91.1	89.1	93.1	96.0	84.2	**97.0**
	Pairwise	95.2	92.9	1.6	94.7	93.5	92.5	**96.8**	90.7	**96.8**
Herding	Single	95.3	90.0	7.3	95.3	94.7	94.7	**98.0**	90.0	95.3
	Pairwise	96.0	92.7	3.9	95.4	96.4	94.4	**97.4**	93.0	96.0
Recency	Single	96.7	91.3	6.7	93.3	94.7	93.3	97.3	90.0	**98.0**
	Pairwise	95.9	93.6	3.3	93.3	93.4	92.4	96.4	93.2	**97.0**
Overconfidence	Single	93.3	94.7	4.7	95.3	93.3	93.3	**98.7**	92.7	97.3
	Pairwise	95.1	94.2	2.7	93.7	94.2	94.3	**96.9**	93.0	96.6

Table 18: Rewriting quality per bias type on TriggerBank, as scored by the Gemma 4 31B LLM judge. Naturalness scores are omitted as they exceed 96% across all bias types and models excluding Gemini 3.1 Flash. Best results per row shown in bold; Gemini 3.1 Flash excluded from bolding due to low factual preservation (Table 5).

Bias Type	Setup	GPT-5.4m	Claude H.	Gemini F.	Llama 8B	Llama 70B	Qwen3 8B	Qwen3.5 9B	Qwen3 32B	Gemma 4 31B
Anchoring	No def.	0.807	0.813	0.793	**0.84**	0.76	0.833	**0.847**	0.793	0.807
	Generic	0.793	**0.853**	0.787	0.813	0.713	**0.853**	**0.853**	0.78	0.773
	CoT	**0.813**	0.547	0.327	0.687	0.613	0.807	0.773	0.78	0.78
	Targeted (det.)	0.807	0.787	0.733	0.8	0.687	**0.873**	0.827	0.787	0.76
	Targeted (gold)	0.78	0.787	0.747	0.813	0.687	**0.88**	0.827	0.78	0.76
	Rewrite (det.)	0.86	0.787	0.520	0.82	0.753	0.847	0.807	0.873	**0.893**
	Rewrite (gold)	0.853	0.773	0.520	0.82	0.76	0.86	0.853	**0.92**	0.907
Authority	No def.	**0.96**	0.927	0.947	0.927	0.867	0.947	0.973	0.973	0.953
	Generic	0.953	0.947	0.933	0.933	0.82	0.953	0.967	**0.987**	0.953
	CoT	**0.953**	0.647	0.333	0.78	0.647	0.893	0.907	0.947	0.94
	Targeted (det.)	0.953	0.96	0.913	0.967	0.793	0.887	0.947	**0.98**	0.947
	Targeted (gold)	0.947	0.96	0.907	0.96	0.793	0.86	0.953	**0.98**	0.947
	Rewrite (det.)	0.96	0.927	0.547	0.913	0.807	0.927	**0.98**	0.953	0.94
	Rewrite (gold)	0.96	0.953	0.633	0.92	0.847	0.953	**0.98**	0.96	0.94
Framing	No def.	**0.901**	0.881	0.851	0.584	0.861	0.822	0.822	0.861	0.891
	Generic	0.941	0.822	**0.960**	0.683	0.931	0.723	0.822	0.812	0.95
	CoT	0.861	0.733	0.000	0.703	0.772	0.673	**0.881**	0.723	0.871
	Targeted (det.)	0.96	0.931	**0.980**	0.762	0.941	0.653	0.802	0.802	0.97
	Targeted (gold)	0.96	0.95	**0.970**	0.713	0.95	0.574	0.871	0.782	**0.97**
	Rewrite (det.)	0.95	0.931	0.376	0.871	0.931	0.931	0.95	0.96	**0.99**
	Rewrite (gold)	0.921	0.921	0.446	0.871	0.941	0.941	0.95	0.941	**0.98**
Herding	No def.	**0.987**	0.947	0.960	0.907	0.893	0.927	**0.987**	**0.987**	0.967
	Generic	0.953	0.953	0.953	0.913	0.887	0.92	**0.973**	**0.973**	0.953
	CoT	**0.94**	0.633	0.333	0.813	0.68	0.893	0.893	0.913	0.92
	Targeted (det.)	0.94	**0.96**	0.900	0.94	0.847	0.86	**0.96**	**0.96**	0.953
	Targeted (gold)	0.927	**0.967**	0.913	0.96	0.827	0.847	0.96	**0.967**	0.947
	Rewrite (det.)	0.953	0.92	0.473	0.907	0.833	0.94	**0.973**	0.94	0.92
	Rewrite (gold)	0.96	0.947	0.620	0.907	0.827	0.94	**0.98**	0.96	0.94
Recency	No def.	**0.967**	0.927	0.927	0.927	0.833	0.947	**0.967**	0.953	0.947
	Generic	0.953	0.953	0.933	0.933	0.813	0.94	0.967	**0.987**	0.947
	CoT	0.953	0.62	0.333	0.787	0.713	0.953	0.873	0.953	**0.96**
	Targeted (det.)	0.96	0.947	0.920	0.953	0.78	0.913	0.953	**0.973**	0.94
	Targeted (gold)	0.96	0.947	0.907	0.953	0.767	0.907	0.933	**0.967**	0.94
	Rewrite (det.)	0.96	0.907	0.507	0.92	0.853	0.94	**0.967**	0.927	0.94
	Rewrite (gold)	0.967	0.947	0.673	0.92	0.853	0.953	**0.98**	0.96	0.947
Overconf.	No def.	**0.94**	0.92	0.900	0.913	0.853	**0.94**	0.913	**0.94**	0.92
	Generic	0.927	0.92	0.920	0.953	0.873	0.947	0.927	**0.967**	0.927
	CoT	**0.94**	0.613	0.333	0.8	0.68	0.893	0.747	0.887	0.873
	Targeted (det.)	0.94	0.913	0.900	**0.947**	0.833	0.907	0.927	**0.947**	0.913
	Targeted (gold)	0.927	0.92	0.893	**0.96**	0.82	0.88	0.9	0.947	0.907
	Rewrite (det.)	0.94	0.92	0.520	0.92	0.833	0.933	**0.967**	0.94	0.933
	Rewrite (gold)	**0.96**	0.933	0.667	0.94	0.9	0.947	**0.967**	0.947	0.94

Table 19: Sentiment classification accuracy per bias type across all defense setups on the single-bias subset of TriggerBank.

Setup	GPT-5.4m	Claude H.	Gemini F.	Llama 8B	Llama 70B	Qwen3 8B	Qwen3.5 9B	Qwen3 32B	Gemma 4
No def.	0.888	0.878	0.890	0.806	0.831	0.868	**0.891**	0.889	0.89
Generic	0.896	0.889	0.896	0.824	0.827	0.846	0.892	0.854	**0.9**
CoT	**0.877**	0.625	0.252	0.716	0.662	0.781	0.819	0.798	0.86
Targeted (det.)	**0.894**	0.884	0.873	0.843	0.796	0.805	0.878	0.851	0.886
Targeted (gold)	**0.892**	0.886	0.877	0.85	0.79	0.793	0.876	0.844	0.883
Rewrite (det.)	0.91	0.889	0.499	0.854	0.826	0.908	0.923	0.921	**0.928**
Rewrite (gold)	0.909	0.891	0.572	0.856	0.839	0.906	**0.934**	0.927	0.932

Table 20: Sentiment classification accuracy across all defense setups on the pairwise-bias subset of TriggerBank.