FINRAIN: RELIABILITY-AWARE NEWS INGESTION FOR SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM based trading systems often treat financial news ingestion as a sentiment preprocessing task: each article is assigned a directional label and passed to a downstream trading framework. This design leaves two questions unanswered: how much the article should be trusted, and whether that judgment can be made from the model's interior knowledge. We propose **FinRAIN**, a plug-in news ingestion framework that transforms raw news into structured, reliability aware inputs for LLM based trading systems, with access to external knowledge sources. FinRAIN annotates each article for *Market Relevance*, *Source Attribution*, *Information Novelty*, *Headline Exaggeration*, and *Stakeholder Bias* as more interpretable inputs to diverse downstream frameworks. Across five sentiment analysis models, FinRAIN roughly doubles sentiment-and-return correlation and raises directional hit rates from 53.8% to 57.2%; in downstream trading simulations, it improves Sharpe ratios in different frameworks up to 2.5. These results suggest that reliability-aware ingestion is a reusable interface for making news ingestion pipelines more robust, interpretable, and transferable across trading frameworks. The code and sample data are provided in an anonymized repository for review.[1]

1 INTRODUCTION

The standard recipe for LLM-driven equity trading has converged to a three-stage pipeline: collect news and prices, run a generic sentiment stage that summarizes each article into a positive, negative, or neutral label, and send the resulting signal to a downstream trading framework that converts textual evidence into orders (Lopez-Lira & Tang, 2023; Fatouros et al., 2025; Xiao et al., 2024; Yu et al., 2024). The same convention appears in live LLM-trading benchmarks that stream one article at a time per asset (Yu et al., 2025) and in finance-tuned sentiment LLMs that emit a single per-article scalar (Iacovides et al., 2025). As a result, recent work has invested heavily in the predictor and the trading policy, while the ingestion layer is often treated as a solved preprocessing step.

This convention overlooks two structural limitations. First, raw news varies substantially in relevance and reliability, introducing noise into downstream sentiment signals. For example, rumors or exaggerated headlines should not be treated like company statements, or named analyst notes. Second, existing pipelines lack a standardized and reusable mechanism for assessing news reliability before sentiment extraction. Both limitations arise at the ingestion stage, and downstream policy optimization cannot fully recover signal quality once noisy news is inforporated without differentiation.

This paper introduces **FinRAIN** (*Reliability-Aware Ingestion of News*), a structured and model-agnostic dedicated news ingestion framework that evaluates and generates reliability scores of each article from five dimensions: market relevance, source attribution, information novelty, headline exaggeration, and stakeholder bias. The FinRAIN score can then be fed into diverse downstream systems together with raw news to improve sentiment signal generation.

We evaluate FinRAIN from two complementary perspectives. First, we assess the ingestion layer by comparing signed sentiment with next day returns with and without FinRAIN. Second, we insert the same annotations into single-stage and two-stage trading frameworks. The two-stage design is an additional testbed for whether FinRAIN remains effective when domain level context informs ticker

[1] `https://anonymous.4open.science/r/FinRAIN-6BB6`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

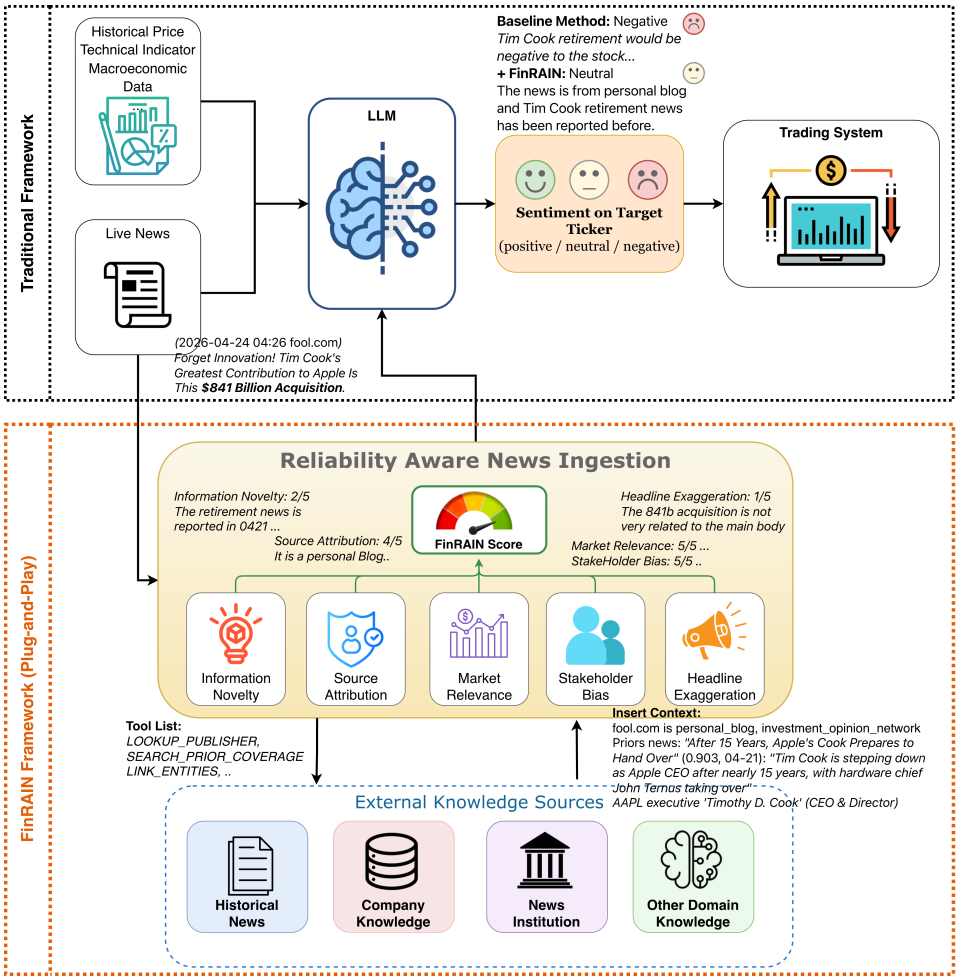


Figure 1: **Overview of FinRAIN.** FinRAIN acts a model agnostic news ingestion framework that transforms raw news into structured, reliability aware inputs for LLM based trading systems, with optional retrieval augmentation from external knowledge sources.

level decisions, rather than a central contribution. Taken together, these experiments examine whether reliability aware ingestion improves both direct article-to-ticker sentiment pipelines and hierarchical trading frameworks.

Our contributions are:

- **A plug-in ingestion framework.** We formulate financial news ingestion as a modular framework that produces reliability measurements while leaving sentiment estimation and trading decisions to any downstream model or strategy.
- **An intuitive reliability schema with unified quantitative measurement.** FinRAIN decomposes article quality into five intuitive reliability dimensions, and provides quantitative evaluation mechanism for each reliability dimension which can be easily aggregated and applied to increase or decrease the influence of individual news articles when deriving sentiment signals.
- **Comprehensive intrinsic and downstream evaluation.** We evaluate FinRAIN effectiveness across multiple LLM backbones and stocks, showing improved next day sentiment alignment. We further integrate FinRAIN into single stage and two stage trading frameworks, where it improved risk adjusted returns across multiple trading policies.

2 Related Work

LLM-derived trading signals. Recent work showed that LLMs can extract trading-relevant signals from financial text. Lopez-Lira & Tang (2023) found that ChatGPT-based sentiment predicted stock returns in a long-short setting. Fatouros et al. (2025) constructed portfolios from GPT-4 interpretations of headline news. Multi-agent systems extended this idea by embedding LLM analysts, traders, and risk managers in end-to-end trading frameworks (Xiao et al., 2024; Yu et al., 2024). Finance-specific foundation models such as BloombergGPT (Wu et al., 2023), FinGPT (Yang et al., 2023a), InvestLM (Yang et al., 2023b), and PIXIU/FinMA (Xie et al., 2023) improved financial-language understanding, while reinforcement-learning frameworks consumed engineered signals downstream (Yang et al., 2020). These systems usually evaluated the full news-to-action pipeline, but the ingestion step was rarely isolated as a distinct object of evaluation.

Financial sentiment baselines. FinBERT (Araci, 2019), DistilRoBERTa-financial-news Romero (2020), and FinMA-7B (Xie et al., 2023) are natural baselines because they mapped financial text to sentiment labels or scores. More recent finance-tuned sentiment LLMs such as FinDPO (Iacovides et al., 2025) likewise emitted a single per-article scalar. Their outputs are useful directional inputs, but they are article-level by design. They do not natively produce market materiality, source attribution, novelty, or other reliability-risk fields.

Financial reliability and factuality. Financial misinformation benchmarks such as Fin-Fact (Rangapur et al., 2023) and FMDLlama (Liu et al., 2025) evaluated whether financial claims were factual or misleading, and recent atomic-claim verification (Guo et al., 2026) and graph-RAG hallucination benchmarks (Kumar et al., 2026) rendered the underlying axes individually testable in financial QA. Broader NLP work studied factual consistency, calibration, self-consistency, and atomic factuality (Guo et al., 2017; Maynez et al., 2020; Elazar et al., 2021; Wang et al., 2023; Min et al., 2023; Chen et al., 2024). FinRAIN is complementary to these efforts. It treats factuality as one part of a broader ingestion problem that also includes source attribution, information novelty, headline exaggeration, and stakeholder bias.

What FinRAIN adds. FinRAIN introduces a reliability aware guard between raw news and downstream LLM based trading systems. Rather than allowing every retrieved article to influence the decision equally, it enables a ingestion framework to identify, weight, and trust news appropriately before sentiment analysis. This distinction is critical because even a directionally capable model can fail when irrelevant, stale, exaggerated, or conflicted articles are treated as equally informative inputs. To our knowledge, prior LLM based trading work has not jointly assessed article level materiality and reliability and quantified their effect under otherwise identical downstream simulations. FinRAIN evaluates this structured assessment through both intrinsic sentiment alignment and controlled trading experiments.

3 FinRAIN Framework

3.1 Structured Ingestion Schema

Let $\mathcal{T} = \{n_1, \ldots, n_N\}$ be a fixed trading universe of N tickers. For each article e and candidate ticker n, FinRAIN answers whether article is about ticker n based on routing prompt and emits five article-level scores shared across candidate tickers:

$$a_e = (D_e^{\mathrm{mr}}, D_e^{\mathrm{sa}}, D_e^{\mathrm{in}}, D_e^{\mathrm{he}}, D_e^{\mathrm{sb}}), \tag{1}$$

corresponding to market relevance, source attribution, information novelty, headline exaggeration, and stakeholder bias. Each $D_e^{\mathrm{reliability}}$ ranges from 1 to 5 while 5 represents most favorable score.

Reliability dimensions. FinRAIN estimates article reliability along five dimensions:

- **Market Relevance** (D_e^{mr}). Measures how directly the article relates to financial markets and whether its content could reasonably affect investor expectations or stock prices. For example, a change in revenue guidance is more market relevant than an announcement about a minor internal company event.

- **Source Attribution** (D_e^{sa}). Measures whether the article supports its claims with identifiable and credible sources. For example, a news published on Yahoo Finance is more reliable than one from an anonymous blog post.
- **Information Novelty** (D_e^{in}). Measures how much new information the article provides relative to recently published coverage of the same topic.
- **Headline Exaggeration** (D_e^{he}). Measures whether the headline overstates, sensationalizes, or misrepresents the tone and substance of the article body.
- **Stakeholder Bias** (D_e^{sb}). Measures whether the article may be influenced by the interests of the individuals or organizations presenting the information. For example, a company executive discussing the firm's outlook may emphasize positive developments while downplaying potential risks.

These dimensions reflect the core checks that analysts apply before using news in a trading decision. FinRAIN converts these qualitative checks into structured quantitative annotations that downstream LLM can consume together with news as additional information while leaving directional prediction and portfolio action to the downstream process.

3.2 External Knowledge Augmentation

Additional external knowledge source can strengthen FinRAIN performance if required information is not contained in the article itself. Among the dimensions introduced in 3.1, information novelty depends on earlier coverage of the same event, source attribution can benefit from publisher and institution metadata, and stakeholder bias can be evaluated more accurately when relationships among companies, executives, and other interested parties are available. Other external sources may also provide additional context for assessing market relevance and determining whether the headline exaggerates the article's claims. FinRAIN therefore adds an optional grounding step: we assemble a small set of reference sources (Section 3.2.1), and expose them to the scoring model through a fixed pipeline or harness (Section 3.2.2). Let z_e denote the external knowledge available for article e. FinRAIN produces its reliability annotations as

$$a_e = F_\theta(e, z_e), \tag{2}$$

where F_θ is the scoring model operating under the fixed FinRAIN rubric. When no external knowledge source is available, $z_e = \varnothing$, and the framework reduces to the article only setting. This design allows FinRAIN to operate across environments with different data access, infrastructure, and computational constraints while preserving the same reliability schema and output interface.

3.2.1 External Knowledge Sources

FinRAIN does not prescribe a particular external database. Instead, each reliability dimension can be connected to the most appropriate evidence source available in a given implementation. In our experiments, we construct a lightweight collection of public reference data that approximates resources commonly maintained by institutional users. All four are constructed offline from stored data sources, so FinRAIN performs no live crawling during scoring. (Appendix J).

- **News archive:** Historical articles are embedded using a encoder (Reimers & Gurevych, 2019) and retrieved through nearest neighbor search. This allows FinRAIN to determine whether an article contains new information or repeats earlier coverage.
- **Company knowledge graph:** The graph connects companies with their industries, peers, competitors, subsidiaries, executives, and macroeconomic themes. It supports entity linking and the identification of potentially interested stakeholders.
- **Publisher metadata:** Publisher domains are mapped to institution type and credibility tier. This information supports the assessment of source attribution and potential stakeholder bias.
- **Price and indicator tape:** Historical returns, volatility, and trend estimates available before the article timestamp provide market context for evaluating claims about price movements and market conditions.

The external sources construction are not a main contribution of FinRAIN. They serve as replaceable knowledge sources that demonstrate how the reliability schema can be grounded when suitable data are available.

3.2.2 Pipeline Retrieval and Harness Tooling

FinRAIN supports two ways of incorporating external knowledge sources into reliability scoring.

Pipeline retrieval. In the pipeline implementation, external sources are retrieved by a fixed upstream procedure before the scoring model is called. For each article, the pipeline can retrieve similar earlier articles, look up the publisher and named institutions, resolve relevant company and stakeholder relationships, and obtain the corresponding historical market records. The retrieved records are then appended to the candidate article as additional scoring context. It is also suitable for systems in which data preparation and model inference are implemented as separate stages. We refer to this setting as *pipeline retrieval* in the experiments.

Harness with provided tools In the harness implementation, the same external knowledge sources are exposed to the scoring model through a collection of read only tools (the full list of tools are listed in Appendix K). Rather than receiving a fully assembled evidence package, the model determines which sources are needed for the article and selects the corresponding tool calls within a fixed interaction budget. For example, it may retrieve earlier coverage when evaluating novelty, inspect a company relationship when evaluating stakeholder bias, or consult the publisher registry when the origin of a claim is unclear. If no tool is called, the model scores the article directly under the same rubric, reducing FinRAIN to the unaugmented setting. We refer to this model directed setting as *agentic retrieval* in the experiments.

The two implementations use the same FinRAIN dimensions, scoring rubric, and output format. They differ only in how external evidence is selected and added in the context of the scoring model. Pipeline retrieval provides consistent prefetched context, whereas harness tooling allows evidence acquisition to adapt to the information needs of each article. In both cases, retrieval remains modular: data sources, retrieval algorithms, and historical coverage windows can be replaced or updated without modifying the FinRAIN schema or any downstream sentiment generation. Temporal validity is also enforced for the augmentation by resolving every query relative to the article timestamp, preventing information published after the article from entering the assessment.

4 Experimental Setup

4.1 Data and Universe

We construct the experimental universe from four major market domains: technology, healthcare, finance, and energy. These domains receive sustained investor attention, generate substantial news coverage, and encompass a broad range of events that can influence stock prices. Within each domain, we select five large, liquid, and widely followed US listed stocks, producing a 20 stocks universe with equal representation across domains (full list provided in Appendix C). Focusing on the large companies with consistent news coverage reduces variation caused by sparse information or irregular trading, enabling more controlled comparisons across tickers and domains.

4.2 News Retrieval and Corpus Construction

We then retrieve news using NewsAPI.org [2], with queries that combine company tickers and domain keywords. Retrieval is restricted to English articles in a single calendar day window that advances daily. Because some sources provide only a headline and summary, we fetch the original article page to extract the full text and use the summary as a fallback when extraction fails. All articles are standardized and de-duplicated by URL and title content before corpus construction. After de-duplication, the corpus contains 25,818 articles from 1,298 distinct publishers covering 132 days from 2025-12-26 to 2026-05-06. The publisher distribution is heavy-tailed, with Yahoo Finance accounting for 40% of articles, a detailed description can be found in Appendix D.

[2] https://newsapi.org/

Table 1: **FinRAIN score distributions and agreement with human annotations.** Distributions are percentages of the full corpus. Agreement is evaluated on 100 randomly sampled articles using quadratic weighted Cohen's κ_w, Spearman's ρ, and mean absolute deviation (MAD). Bold indicates the highest agreement in each metric.

Score / Metric	Information Novelty	Stakeholder Bias	Source Attribution	Market Relevance	Headline Exaggeration
Full-corpus score distribution (%)					
1	14.2	4.7	5.0	49.6	2.9
2	7.3	9.7	20.8	12.7	3.9
3	23.3	17.0	37.1	11.1	2.6
4	33.0	19.8	31.4	9.6	4.9
5	22.1	48.8	5.7	17.0	85.7
Human agreement ($n = 100$)					
$\kappa_w \uparrow$	0.37	0.54	0.54	**0.63**	0.36
Spearman $\rho \uparrow$	0.39	0.58	0.57	**0.81**	0.48
MAD $\downarrow$	1.16	0.85	**0.68**	0.84	**0.68**

4.3 Reliability Scoring

Each dimension can be scored directly by the LLM using its predefined rubric and the candidate article alone. When external knowledge is available, FinRAIN can optionally augment the scoring process with external evidence tailored to the corresponding dimension as mentioned in Section 3.2. The complete scoring protocol and rubric anchors are provided in Appendix E.

Computational cost. FinRAIN scoring uses 2,013 tokens per article without external evidence, 2,576 with pipeline retrieval, and 10,795 with agentic retrieval. Adding FinRAIN annotations increases policy input length but has little effect on output length. Appendix F provides the full token accounting.

4.4 Intrinsic Evaluation Setup

We evaluate whether FinRAIN improves the alignment between news sentiment and subsequent market movements, independently of any trading policy. For each date and ticker bucket, we generate article level sentiment using five sentiment backbones that are released before the earliest date of our day range (2025-12-26): FinBERT (Araci, 2019), DistilRoBERTa (Romero, 2020), FinMA (Xie et al., 2023), Claude Sonnet 4.5 (Anthropic, 2025), and GPT-OSS-120b (OpenAI, 2025). The baseline provide sentiment without reliability assessment, FinRAIN instead weights their sentiment according to the five reliability dimensions. We measure performance using the Spearman correlation between the aggregated signed sentiment signal and the signed one day forward return, together with directional hit rate.

To further assess the contribution of each reliability dimension, we conduct a leave one dimension out ablation under the best performing agentic configuration to measure each dimension's marginal contribution to the complete FinRAIN framework. We repeat the experiment using FinBERT and Claude Sonnet 4.5, providing coverage of both a finance specialized language model and a general purpose LLM.

4.5 Downstream Integration Setup

The downstream diagnostic asks whether the same FinRAIN annotations improve trading decisions when inserted into otherwise unchanged frameworks. We evaluate the risk-adjusted return metrics using sentiment derived trading policies with the same 20-stock universe, $1,000 per leg, 2 bps round-trip cost, and identical price simulator across arms.

Table 2: **Next-session sentiment alignment across sentiment backbones with GPT-OSS-120b-scored FinRAIN.** Each cell reports Spearman correlation ρ / directional hit rate (%) between the aggregated sentiment signal and the signed next-session return. Pipeline and agentic retrieval use external sources. Higher is better for both metrics. Bold indicates the best result for each metric within each backbone.

Method	FinBERT	Distil RoBERTa	FinMA	Claude Sonnet 4.5	GPT-OSS 120b
Baseline	.0561 / 53.9	.0543 / 52.9	.0730 / 53.9	.0779 / 53.8	.0663 / 53.3
FinRAIN configurations					
No external source	.0717 / 54.5	.0728 / 54.8	.0760 / 53.8	.0937 / 55.8	.1091 / 56.3
Pipeline retrieval	.0980 / 54.4	.0873 / 53.5	.1083 / 54.3	.0938 / 54.0	.1111 / 54.5
Agentic retrieval	**.1280 / 55.6**	**.1235 / 55.1**	**.1379 / 55.7**	**.1333 / 57.2**	**.1314 / 56.6**

The *Single-Stage Baseline Framework* follows the spirit of Lopez-Lira & Tang (2023). For each article-ticker pair, the policy LLM answers whether the article moves the ticker upward over the next session using YES, NO, or UNKNOWN. We aggregate the day-level signal as $(\#\text{YES} - \#\text{NO})/N$ and trade when it crosses ± 0.10. The baseline row sees only the article and ticker. The FinRAIN arm receives the same inputs plus reliability annotations.

The *Two-Stage Domain-to-Ticker Framework* is our downstream integration design. Stage 1 forms a domain-level view from the day's domain-tagged news chunks. Stage 2 issues ticker-level action using domain context from Stage 1, ticker-specific articles, and a quantitative panel. This design tests a top-down hierarchy: broad domain news first determines the domain outlook, then ticker-level evidence determines action within that outlook. The baseline setting runs both stages without FinRAIN fields. In the FinRAIN condition, each article is augmented with its FinRAIN annotations.

We evaluate three trading conventions. *Long-Only Intraday* buys at the open and sells at the close. *Long-Short intraday* buys positive signals and shorts negative signals over the same open-to-close window. *Round-Trip Long-Only* buys at the first positive signal and closes according to the framework's next negative signal, with one position per ticker.

5 Experimental Results

Using the data, model configurations, evaluation metrics, and simulation settings defined in Section 4, we now examine the empirical effectiveness of FinRAIN. We first characterize the reliability scores produced across the news corpus in Section 5.1. Then we evaluate whether FinRAIN improves the alignment between news derived sentiment and subsequent price movements and use ablation studies to identify the contributions of individual reliability dimensions in Section 5.2. Finally, we assess whether these improvements translate into stronger performance when FinRAIN is integrated into trading frameworks in Section 5.3.

5.1 Reliability Distribution and Human Agreement

Table 1 shows that the five dimensions capture distinct properties of the corpus. Stakeholder bias and headline exaggeration are concentrated at high scores, whereas market relevance is concentrated at low scores, indicating that many retrieved articles have limited relevance to the target ticker. Information novelty is more dispersed, while source attribution is concentrated near the middle of the scale.

We further compared FinRAIN with independent human annotations on 100 randomly sampled articles. Agreement is strongest for market relevance, with $\kappa_w = 0.63$ and $\rho = 0.81$, and weakest for information novelty and headline exaggeration, with $\kappa_w = 0.37$ and 0.36. MAD remains below one scale point for four dimensions, indicating that most disagreements involve nearby scores rather than opposing judgments. FinRAIN generally assigns higher reliability scores than the human annotators, which helps explain why rank correlation is consistently higher than weighted agreement.

Table 3: **Downstream trading performance across portfolio construction strategies.** PnL is reported in USD using an initial capital of $1,000. Pipeline and agentic retrieval use external sources. Higher is better for all metrics. Bold indicates the best result in each column within each downstream framework, including ties.

	Long-only		Long-short		Round-trip	
Method	**Sharpe ↑**	**PnL ($) ↑**	**Sharpe ↑**	**PnL ($) ↑**	**Sharpe ↑**	**PnL ($) ↑**
Single-stage baseline framework						
Baseline	−1.07	−49.5	−0.54	−20.5	0.39	+11.8
FinRAIN (no external source)	−0.28	−14.1	**0.72**	+29.3	0.85	+27.6
FinRAIN + pipeline retrieval	0.07	+0.5	**0.72**	**+30.0**	0.87	+30.0
FinRAIN + agentic retrieval	**0.12**	**+2.7**	0.26	+8.7	**0.96**	**+32.7**
Two-stage domain-to-ticker framework						
Baseline	0.55	+31.1	−0.73	−54.4	1.28	+66.3
FinRAIN (no external source)	1.58	+82.9	**0.35**	**+14.5**	1.40	+60.7
FinRAIN + pipeline retrieval	**1.60**	**+106.8**	−0.46	−34.5	**2.50**	**+138.2**
FinRAIN + agentic retrieval	0.94	+58.7	−0.63	−46.7	1.28	+64.2

5.2 FinRAIN Intrinsic Effectiveness

We next evaluate whether FinRAIN produces sentiment signals that align more closely with subsequent price movements across multiple sentiment models. Table 2 shows that FinRAIN improves sentiment and return alignment across all five backbones. Baseline correlations range from 0.0543 to 0.0779, whereas FinRAIN without external data sources achieves correlations from 0.0717 to 0.1091. The corresponding directional hit rates increase from 52.9% –53.9% to 53.8%–56.3%. External evidence yields correlations up to 0.1379 and hit rates up to 57.2%. To assess whether these gains extend beyond the selected backbone, we repeat the same evaluation based on FinRAIN scores from Claude Sonnet 4.5. The corresponding results, reported in Appendix G, show a consistent improvement pattern, providing additional evidence that FinRAIN is not tied to a particular LLM.

A decomposition of the improvement shows that FinRAIN's reliability weighting provides a substantial share of the gain. While external context helps, agentic retrieval tends to outperform pipeline retrieval. These results indicate that FinRAIN's primary effectiveness arises from its structured reliability aware weighting, while external knowledge augmentation remains an optional extension whose incremental value may depend on the dataset and evaluation horizon.

Ablation study of reliability dimensions We then conduct ablation studies to isolate the contribution of each reliability dimension and assess the robustness of the observed improvement. Removing any dimension reduces both the Spearman correlation and directional hit rate under both FinBERT and Claude Sonnet 4.5 models as shown in Figure 2, the consistent degradation across both model classes shows that FinRAIN's effectiveness does not depend on a particular sentiment backbone and that the complete multidimensional schema performs better than any reduced variant.

5.3 FinRAIN Impact on Downstream Trading Frameworks

Finally, we examine whether the improvements in intrinsic sentiment alignment translate into stronger downstream trading performance. Table 3 reports annualized Sharpe ratios and absolute PnL normalized to $\$1,000$ balance settings. In the single stage framework, every FinRAIN variant improves over the baseline across all three trading conventions. Standard FinRAIN increases the Sharpe ratio from -1.07 to -0.28 for long only trading, from -0.54 to 0.72 for long short trading, and from 0.39 to 0.85 for round trip trading. The agentic retrieval variant achieves the strongest long only and round trip Sharpe ratios, at 0.12 and 0.96, respectively.

The two stage framework supports the same conclusion that every FinRAIN setting outperforms baseline. Standard FinRAIN performs best under the long-short policy, increasing the corresponding

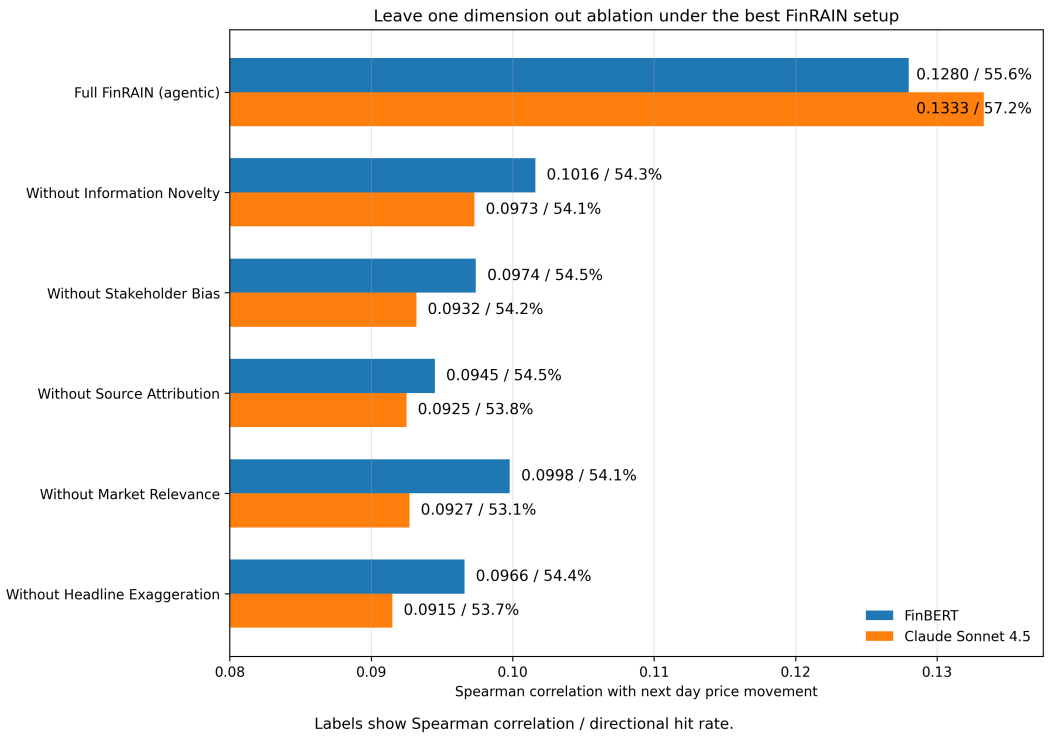


Figure 2: **FinRAIN (with GPT-OSS-120b scoring) leave one dimension out ablation.** Removing any dimension reduces sentiment alignment under both finance specialized LLM (FinBERT) and general purpose LLM (Claude Sonnet 4.5).

Sharpe ratios from -0.73 to 0.35. In contrast, FinRAIN with external source performs best for long-only and round-trip policies, with Sharpe ratio up to 1.60 and 2.50 respectively. Overall, the results support reliability aware ingestion can improve trading decisions.

6 Discussion and Limitations

FinRAIN separates news ingestion from sentiment prediction by assessing whether and how strongly an article should influence downstream decisions. Its five reliability dimensions are independent of the downstream model and policy and support operation with or without external evidence. Across evaluated sentiment backbones, FinRAIN improves alignment with subsequent price movements; ablations indicate complementary contributions across dimensions, and controlled trading simulations show gains in risk-adjusted performance. These findings suggest that improving predictors alone may not address irrelevant, repetitive, exaggerated, or poorly supported news.

These results remain bounded by a 132-day benchmark covering 20 large, liquid US-listed stocks across four domains and a limited set of scoring and sentiment models. Broader generalization requires evaluation across market regimes, asset universes, languages, and model families. Trading simulations use fixed notional, simplified costs, and limited capacity modeling and should not be interpreted as deployment-ready backtests. Reliability dimensions also rely on LLM judgments, motivating stronger retrieval-based or supervised grounding, particularly for source attribution, veracity, and novelty. Future work should address these limitations, adapt reliability assessment to event type and market conditions, and explore integration with multi-agent decision-making and portfolio optimization.

AI USE STATEMENT

Generative AI tools were used to polish author-written text for clarity and language. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All codes to collect source news and construct FinRAIN are uploaded to the Supplementary Materials. And the implementation details are in Appendix A.

REFERENCES

Anthropic. Claude sonnet 4.5. Large language model, 2025. URL `https://claude.ai`.

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1012–1031, 2021.

Georgios Fatouros, Konstantinos Metaxas, John Soldatos, and Dimosthenis Kyriazis. Can large language models beat Wall Street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI. *Neural Computing and Applications*, 37:24893–24918, 2025. doi: 10.1007/s00521-024-10613-4.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. FinGround: Detecting and grounding financial hallucinations via atomic claim verification. *arXiv preprint arXiv:2604.23588*, 2026.

Giorgos Iacovides, Wuyang Zhou, and Danilo Mandic. FinDPO: Financial sentiment analysis for algorithmic trading through preference optimization of LLMs. *arXiv preprint arXiv:2507.18417*, 2025.

Mahesh Kumar, Bhaskarjit Sarmah, and Stefano Pasquali. FinReflectKG–HalluBench: GraphRAG hallucination benchmark for financial question answering systems. *arXiv preprint arXiv:2603.20252*, 2026.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. FMDLlama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25)*, 2025.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. Fin-Fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Manuel Romero. Distilroberta fine-tuned for financial news sentiment analysis. `https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis`, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

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

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF)*, 2020. doi: 10.1145/3383455.3422540.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023a.

Yi Yang, Yixuan Tang, and Kar Yan Tam. InvestLM: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*, 2023b.

Haofei Yu, Fenghai Li, and Jiaxuan You. LiveTradeBench: Seeking real-world alpha with large language models. *arXiv preprint arXiv:2511.03628*, 2025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A HYPERPARAMETERS

LLM scoring. For Sonnet 4.5 and GPT-OSS-120B, decoding is greedy with temperature 0. The per-article body cap is 1,500 characters. Articles are batched by (sector, date) with batch size 8.

Sentiment baselines. FinBERT and DistilRoBERTa-financial use title plus body with 512-token truncation and batch size 64 on an A10G GPU. FinMA-7B uses 1,900-token truncation, batch size 4, fp16, and one A10G GPU. Greedy decoding uses max-new-tokens $= 8$ for the single-label prompt and 200 for the per-ticker prompt in Section 5.2.

Coverage gate. We use $D^{\text{sc}} > 0.3$ for all magnitude and downstream experiments. This gate is treated as a fixed, permissive routing threshold rather than a return-tuned hyperparameter; it is meant to retain weak but plausible ticker links while removing clearly off-scope articles.

Encoder direction threshold. We map encoder softmax outputs to $\{-1, 0, +1\}$ using $\mathrm{sgn}(p_+ - p_-)$ with a threshold of ± 0.10.

Downstream simulator. The downstream diagnostic uses 132 days from 2025-12-26 to 2026-05-06, \$1,000 per leg, and 2 bps round-trip transaction cost. News from $[\text{day}_{t-1}$ 09:30 ET, day_t 09:30 ET) is attributed to trade day t, with weekends rolled to Monday. The single-stage baseline trades BUY if $(\#\text{YES} - \#\text{NO})/N > 0.10$, SELL if the value is below -0.10, and HOLD otherwise.

B MULTI-TICKER PROMPT FOR FINMA-7B

The 20-ticker instruction prompt used in Section 5.2 is reproduced below verbatim:

```
Instruction: For each ticker in the list, what is the sentiment of this news
for that ticker? Choose negative/neutral/positive/not_relevant. Output
one ticker per line as 'TICKER: label'. Tickers: AAPL, MSFT, GOOGL, META,
NVDA, JNJ, PFE, UNH, ABBV, MRK, JPM, BAC, WFC, GS, MS, XOM, CVX, COP, EOG,
SLB. Input: ⟨article⟩. Answer:
```

C TICKER UNIVERSE

The fixed 20-stock U.S. listed stock universe used throughout the paper consists of five large, liquid names from each of four domains. Table 4 lists every ticker by sector.

Table 4: **Full 20-stock universe**, five tickers per sector across four domains. The same universe is used for the corpus retrieval (Section 4.1), the intrinsic diagnostics (Section 5.2), the downstream trading simulations (Section 5.3).

Sector	Tickers
Technology	AAPL, MSFT, GOOGL, META, NVDA
Healthcare	JNJ, PFE, UNH, ABBV, MRK
Finance	JPM, BAC, WFC, GS, MS
Energy	XOM, CVX, COP, EOG, SLB

D CORPUS CONSTRUCTION

The corpus described in Section 4.1 is constructed using a domain based retrieval crawler with **NewsAPI.org** as the primary source. For each domain, we query the `/v2/everything` endpoint using Boolean expressions that combine the corresponding company tickers with curated domain keywords. For example, the technology query combines `AAPL`, `MSFT`, `GOOGL`, `META`, and `NVDA` with terms related to technology, artificial intelligence, software, cloud computing, and semiconductors. Analogous ticker and keyword combinations are used for healthcare, finance, and energy.

Retrieval is restricted to English language articles within a single calendar day query window that advances daily. Because NewsAPI.org may provide only an article headline and summary, the crawler fetches the original article URL and attempts to extract the full body text. When full text extraction is unsuccessful, the available summary is retained as a fallback. All retrieved articles are converted into a common schema and deduplicated using URL and title content matching before final corpus construction. The crawler, domain to ticker mappings, and query keywords will be included in the released code.

After deduplication, the corpus contains $25{,}818$ articles published between December 26, 2025 and May 6, 2026, spanning $1{,}298$ distinct publishers. The publisher distribution is strongly heavy tailed, with Yahoo Finance accounting for approximately 40% of the corpus and the remaining articles distributed across a large number of smaller publishers. Figure 3 presents the distribution of the most frequent publishers.

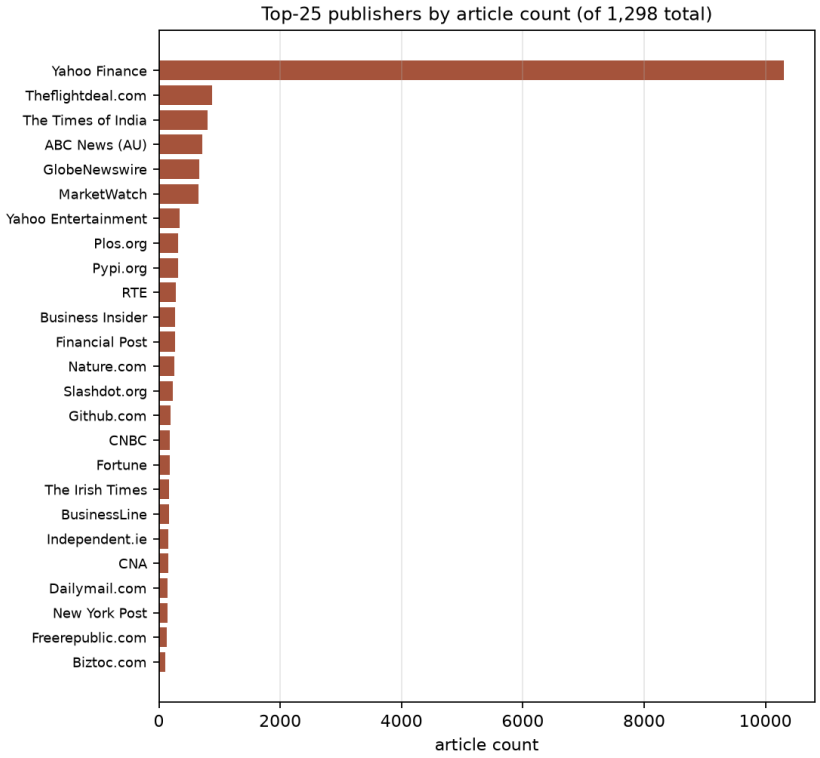


Figure 3: **Distribution of News by Top Publishers**

E Reliability Scoring Protocol

For each candidate article, FinRAIN submits the article and any available retrieval evidence to Claude Sonnet 4.5. A single model call jointly returns the five reliability dimensions, article sentiment, and an indicator of whether additional corroborating evidence is required. Every reliability score is defined by written anchors included directly in the prompt. The model therefore selects the description that best matches the available evidence rather than assigning a number without explicit criteria. No reliability dimension is computed using a fixed arithmetic formula.

- **Information Novelty.** Information novelty is scored from 1 to 5 and measures the amount of new information contributed relative to recent coverage. A score of 5 denotes the first report of a concrete new event. A score of 3 denotes a recognizable continuation of an existing story that contributes at least one meaningful incremental detail. A score of 1 denotes duplicated reporting, a wire rehash, or an article that adds little substantive information. The model is instructed to identify the retrieved prior articles that overlap with the candidate article rather than infer novelty without evidence.
- **Stakeholder Bias.** Stakeholder bias is scored from 0 to 5, with higher scores indicating greater independence from interested parties. A score of 5 represents independent third party reporting. A score of 3 indicates that a party with a direct interest in the event serves as a primary source. A score of 1 indicates sponsored material, press release style reporting, or content substantially controlled by an interested stakeholder.
- **Source Attribution.** Source attribution is scored from 1 to 5 and evaluates both publisher credibility and the transparency of the evidence supporting the article. A score of 5 indicates reporting from a highly credible newsroom that identifies authoritative primary sources. A score of 3 indicates licensed or aggregated reporting that does not clearly identify the original source. A score of 1 indicates poorly attributed content from a low credibility outlet. The prompt also includes the publisher credibility tier obtained from the external institution registry.

- **Market Relevance.** Market relevance is scored from 1 to 5 and measures how clearly and materially the reported event affects a company in the experimental universe. A score of 5 indicates a material and directionally clear event directly associated with a covered ticker. Lower scores represent increasingly indirect, weak, or conditional implications. For example, a score of 2 indicates that the potential market effect depends largely on conditions or assumptions not established within the article.
- **Headline Reliability.** Headline reliability is scored from 1 to 5, with higher scores indicating closer agreement between the headline and article body. A score of 5 denotes a faithful and proportionate summary of the underlying content. A score of 1 denotes a sensational, misleading, or clickbait headline whose framing is not supported by the article body.
- **Additional Outputs.** The same model call produces article sentiment as negative, neutral, or positive, encoded as -1, 0, or +1. It also returns the binary variable "need extra source", which indicates whether the available article and retrieved context provide sufficient evidence or whether additional corroboration is needed. These outputs are retained separately from the five reliability dimensions and are not included as reliability scores.

F TOKEN USAGE WITH VS. WITHOUT FINRAIN

We log per-call token usage from the provider's usage block on every call, for both the reading layer (FinRAIN scoring) and the downstream policy LLM. Table 5 reports the resulting unit costs in tokens. Scoring arms are counted per *article*, since FinRAIN scores six articles per call; policy arms are counted per *decision call*, which is what a trading system issues per ticker-day.

Table 5: **Token usage with and without FinRAIN annotations.** Tokens per unit are averages read from the provider usage block; $\times$ is the multiple over the baseline arm of the same block. Scoring rows cover 2,781 articles (1,703 for the agentic arm, which cannot batch); policy rows cover $8{,}717/8{,}702$ single-stage and $3{,}372/3{,}143$ two-stage calls.

Arm	Unit	In	Out	Total	$\times$
Reading layer (FinRAIN scoring)					
No external source	art.	1,221	792	2,013	1.00
Pipeline retrieval	art.	1,746	830	2,576	1.28
Agentic retrieval	art.	9,727	1,068	10,795	5.36
Routing only, no rubric	art.	242	211	453	0.23
Single-stage policy					
Baseline	call	124	36	160	1.00
+ FinRAIN	call	573	39	612	3.83
Two-stage policy					
Baseline	call	2,770	154	2,924	1.00
+ FinRAIN	call	4,612	160	4,772	1.63

What the extra-source harness costs. Holding model, batch size, and article pool fixed, adding the five evidence blocks costs $+525$ input and $+38$ output tokens per article ($+563$ total, $1.28\times$), or 56.3M tokens per 100,000 articles. The growth is almost entirely on the input side: the rubric output format is fixed, so more evidence does not make the model write more.

What agency costs. Agentic retrieval uses the same rubric and the same five sources, but exposes them as tools and lets the model choose what to fetch. This costs $+7{,}981$ input and $+238$ output tokens per article over pipeline retrieval ($+8{,}219$ total, $4.19\times$), or 822M tokens per 100,000 articles. Two mechanisms drive the gap: the conversation is resubmitted on every turn (2.24 turns per article on average), and batching becomes impossible because the tool results one article needs are specific to that article. The second is a property of the design rather than of this implementation, so we count it as a cost of agency. Given that agentic retrieval also yields the strongest sentiment alignment (Table 2), this is the principal accuracy–cost trade-off in the reading layer.

Table 6: **FinRAIN (with Claude Sonnet 4.5 scoring) improves next session sentiment alignment across multiple sentiment backbones.** Each cell reports Spearman correlation ρ and directional hit rate between the aggregated sentiment signal and the signed next session return, formatted as ρ / hit rate. Bold values indicate the best result in each model column for each metric.

Sentiment Analysis Method	FinBERT	Distil RoBERTa	FinMA	Claude Sonnet 4.5	GPT OSS
Baseline	.0557 / 53.8%	.0534 / 52.9%	.0781 / 54.0%	.0737 / 53.7%	.0630 / 53.2%
FinRAIN, article only	.1287 / 54.6%	.1130 / 55.1%	.1473 / **54.8%**	**.1392** / 55.6%	.1337 / 55.6%
FinRAIN, extra data (pipeline retrieval)	.1222 / 55.6%	.1247 / 55.1%	.1447 / 54.3%	.1257 / 56.4%	.1297 / 56.6%
FinRAIN, extra data (agentic retrieval)	**.1380 / 56.8%**	**.1274 / 56.0%**	**.1533 / 54.8%**	.1311 / **57.3%**	**.1443 / 56.7%**

Validation and coverage. The per-call meter was validated against run-level totals that the policy scripts tallied independently, agreeing to within 0.1% on both single-stage arms (124.1 vs 124.2 and 573.5 vs 573.2 input tokens per call). For the agentic arm, all $3{,}819$ calls charged to the $1{,}703$ scored articles are attributable to a published score, with no failed calls or retries in the metered segment. Metered articles require a body of at least 400 characters and are therefore longer than the live-crawl average, so the counts above are an upper bound on full-corpus figures; we state this rather than rescaling, since the rescaling factor would itself be an estimate. Full-corpus scoring runs ($25{,}464$ articles, 24 workers) took 151.8 minutes with the harness and 137.2 without.

Local baselines. The encoder and quantized-decoder sentiment baselines in Table 2 issue no API calls and consume no API tokens. On a 32-vCPU CPU-only host, FinBERT (110M) processes $\sim$55 articles/s, DistilRoBERTa (82M) $\sim$90 articles/s, and FinMA-7B (4-bit, `llama.cpp`) $\sim$5–7 articles/s. They consume machine time rather than API budget, so they are not cost-free in every sense.

G FinRAIN Intrinsic Effectiveness with Claude Sonnet 4.5 Scoring

To assess whether these gains extend beyond the selected backbone, we repeat the same evaluation using Claude Sonnet 4.5. Table 6 show a consistent improvement pattern, providing additional evidence that FinRAIN is not tied to a particular LLM. Figure 4 shows the leave one dimension out ablation study under using Claude Sonnet 4.5 model for FinRAIN scoring, similar conclusion as GPT-OSS results reported in main paper.

H Case study: GOOGL on 2026-04-14

We use one trading day to illustrate the mechanism behind the aggregate results. On 2026-04-14, GOOGL moves $+2.50\%$ intraday, from an open of $\$324.79$ to a close of $\$332.91$. A correct intraday direction is therefore BUY. The four arms disagree:

Single-stage mechanism: reliability changes weighting. The single-stage baseline routes articles by ticker mention, so both single-stage arms see the same seven GOOGL articles. Without FinRAIN, the framework produces three YES labels and four NO labels, yielding SELL. With FinRAIN, two articles flip from NO to YES, yielding BUY. The flipped articles both have veracity $= 1.00$ and high relevance; the two articles that remain NO have veracity $= 0.50$. Since the article set is unchanged, the mechanism is not scope filtering. FinRAIN changes how the framework weighs article reliability and materiality while making its own directional decision.

Two-stage mechanism: scope changes routing. The two-stage framework adds a routing effect. Without FinRAIN, the ticker-level article binder routes no GOOGL articles into Stage 2, so the framework falls back to sector context and the quantitative panel and chooses HOLD. With FinRAIN, the scope mask routes 25 GOOGL-relevant articles into the ticker-level prompt. The framework then up-weights high-veracity catalysts and drops lower-veracity items, producing BUY. Figure 6 summarizes both mechanisms.

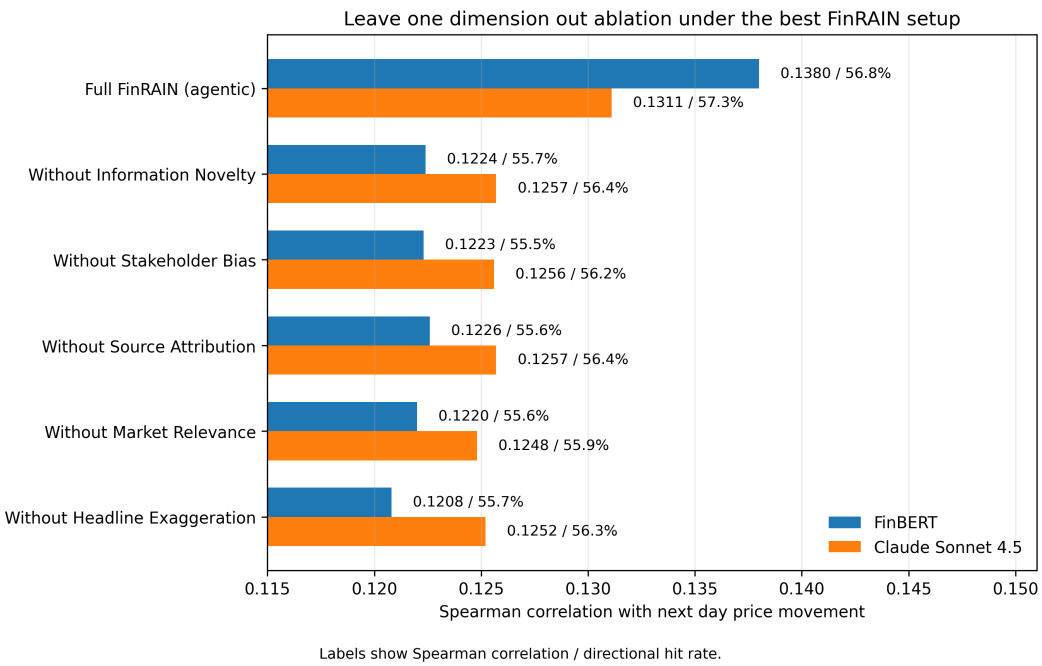


Figure 4: **FinRAIN (with Claude Sonnet 4.5 scoring) leave one dimension out ablation.** Removing any dimension reduces sentiment alignment under both finance specialized LLM (FinBERT) and general purpose LLM (Claude Sonnet 4.5).

This case study separates the two effects that appear in the aggregate results. FinRAIN can improve a framework even when article routing is fixed, by changing how the framework weighs reliability and relevance. It can also improve a framework by changing which articles reach the ticker-level prompt.

I SECTOR BREAKDOWN OF THE DOWNSTREAM TRADING FRAMEWORK

Tables 7, 8, and 9 expand Table 3 to the sector level using the same 132-day window, $1,000 per leg, 2 bps round-trip cost, and simulator. In the two-stage framework, adding FinRAIN improves the corresponding baseline in healthcare, finance, and energy under the long-only intraday convention. Long-short results provide additional sector-level diagnostics. The round-trip variant (long-only with overnight carry) tracks the same direction-of-effect as the intraday long-only configuration: FinRAIN lifts technology and energy when paired with the two-stage policy, while health and finance behave consistently with their long-only profiles.

J CONSTRUCTION OF THE EXTERNAL SOURCES

This appendix details the proxy sources summarised in Section 3.2.1. None is claimed as a contribution; we record their construction so the augmented results are reproducible, and so that a reader with proprietary reference data can see what a substitute must supply.

Historical news archive. Articles are embedded with a sentence encoder (Reimers & Gurevych, 2019) and indexed for approximate nearest-neighbour search. A query is restricted to the candidate article's sector and to timestamps strictly earlier than its own within a fixed lookback window, so retrieved coverage always precedes the article being scored. Maximum cosine similarity to a retrieved prior is reported to the model both as a scalar and as a discretised novelty band, with band edges set at empirical quantiles of the similarity distribution over the corpus.

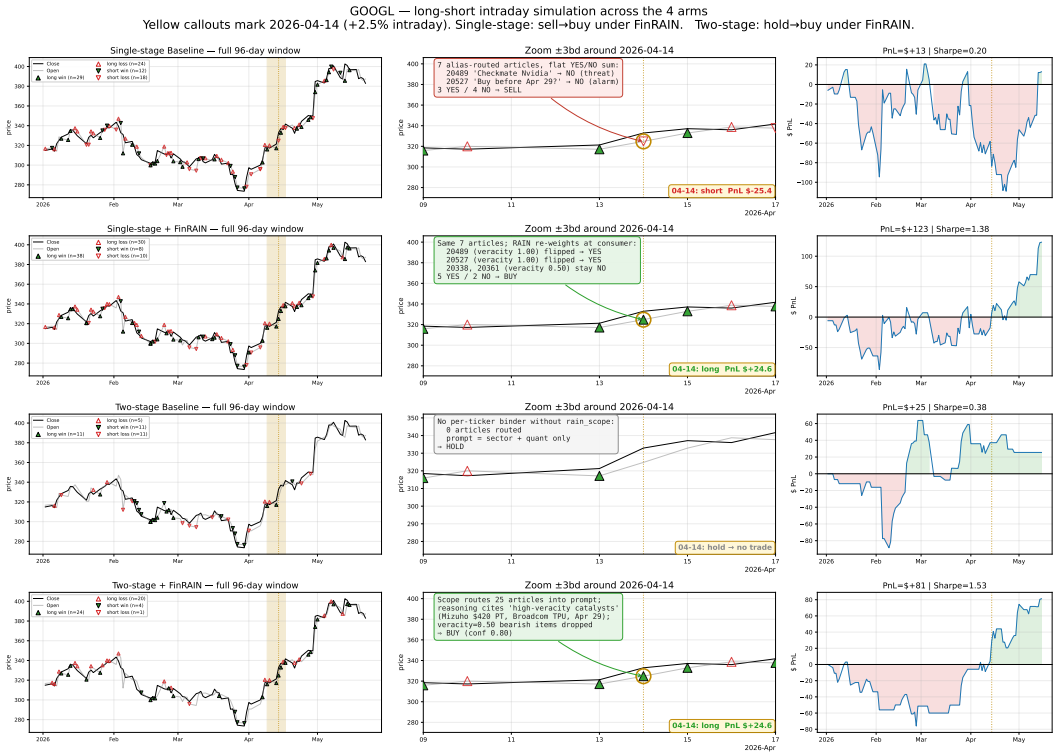


Figure 5: **GOOGL long-short intraday simulation across the four arms.** The figure shows 132 days. The highlighted vertical marker denotes 2026-04-14. The single-stage baseline flips from short to long under FinRAIN, while the two-stage framework flips from hold to long.

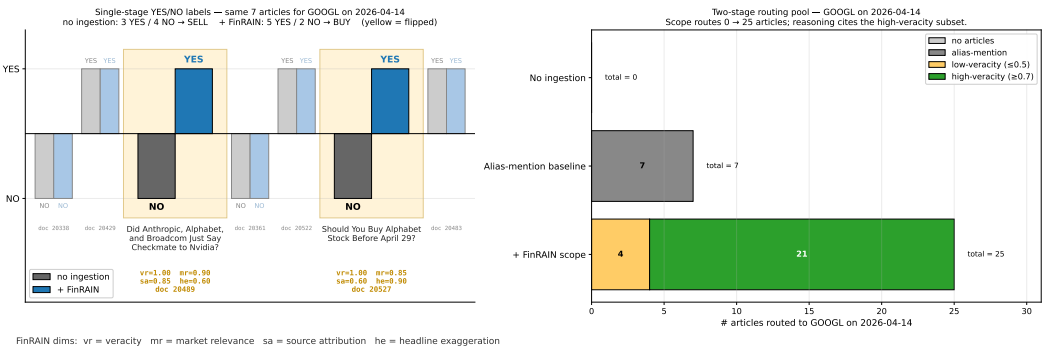


Figure 6: **Mechanisms behind the GOOGL flip on 2026-04-14.** Left: the single-stage baseline sees the same seven articles in both arms, but FinRAIN changes two NO labels to YES by highlighting high-veracity, high-relevance articles. Right: the two-stage framework routes zero ticker-level articles without FinRAIN and 25 in-scope articles with FinRAIN.

Company knowledge graph. Let $G = (V, E)$ be a typed, directed graph anchored on tickers, with nodes drawn from five families and edges from twelve relation types (Table 10). It is assembled in two layers of differing provenance.

The *concept core* is curated and static. Company nodes carry sector, industry, peer-group, index membership, brands, subsidiaries, and named executives, drawn from regulatory filings and a market-data provider. These fields induce the structural relations, and they are what make entity linking auditable: a brand, a subsidiary, or an executive's name resolves to a canonical ticker through a stored edge, so a resulting scope assignment carries a citable reason rather than a model assertion.

Table 7: **Long-only intraday by sector.** Same simulator and window as Table 3.

Sector	Arm	Sh.	PnL	DD%
Tech	1-stage, no ingest.	-0.51	$\$-129$	7.0
	1-stage + FinRAIN	$+0.84$	$\$+128$	5.7
	2-stage, no ingest.	$+2.10$	$\$+268$	3.1
	2-stage + FinRAIN	$+0.90$	$\$+125$	6.5
Health	1-stage, no ingest.	$+2.28$	$\$+241$	2.3
	1-stage + FinRAIN	$+0.99$	$\$+119$	4.1
	2-stage, no ingest.	-2.19	$\$-226$	6.1
	2-stage + FinRAIN	-0.61	$\$-57$	3.4
Finance	1-stage, no ingest.	$+0.42$	$\$+130$	3.3
	1-stage + FinRAIN	$+0.34$	$\$+94$	6.0
	2-stage, no ingest.	-0.58	$\$+6$	5.3
	2-stage + FinRAIN	$+1.12$	$\$+109$	2.4
Energy	1-stage, no ingest.	$+1.99$	$\$+301$	3.1
	1-stage + FinRAIN	$+1.90$	$\$+310$	4.5
	2-stage, no ingest.	$+1.14$	$\$+221$	6.3
	2-stage + FinRAIN	$+1.89$	$\$+442$	7.4

Table 8: **Long-short intraday by sector.** Same simulator and window as Table 3. The largest within-sector swing is technology.

Sector	Arm	Sh.	PnL	DD%
Tech	1-stage, no ingest.	-2.38	$\$-484$	12.0
	1-stage + FinRAIN	$+0.18$	$\$+14$	4.0
	2-stage, no ingest.	$+1.81$	$\$+353$	2.9
	2-stage + FinRAIN	$+1.89$	$\$+314$	4.2
Health	1-stage, no ingest.	$+1.70$	$\$+219$	4.4
	1-stage + FinRAIN	$+1.21$	$\$+143$	3.3
	2-stage, no ingest.	-2.27	$\$-268$	6.0
	2-stage + FinRAIN	-1.09	$\$-129$	5.0
Finance	1-stage, no ingest.	-0.43	$\$-35$	5.7
	1-stage + FinRAIN	$+0.92$	$\$+162$	6.8
	2-stage, no ingest.	-0.40	$\$-17$	8.0
	2-stage + FinRAIN	$+0.20$	$\$+34$	5.3
Energy	1-stage, no ingest.	$+1.10$	$\$+176$	3.0
	1-stage + FinRAIN	$+1.73$	$\$+276$	3.8
	2-stage, no ingest.	$+1.00$	$\$+163$	6.3
	2-stage + FinRAIN	$+1.25$	$\$+304$	7.4

The *event layer* is mined from a ticker-tagged news corpus in which the article–company association is supplied by the provider rather than inferred. Each article is assigned one of nine event classes—earnings, guidance, M&A, regulatory, macro, product, leadership, opinion, other—and articles sharing a (ticker, calendar date, class) triple are merged into a single event node. An event is thus a cluster of reports of one occurrence rather than a headline, which matters because a single earnings print generates many syndicated stories; without clustering, event-driven features would track republication volume. Two properties of this layer are recorded rather than assumed. Classification is *headline-first*, consulting a short opening window of the body only when no rule fires on the headline, and each event stores the fraction of its articles labelled from the headline as an inspectable confidence signal—a description of how a label was obtained, not of whether it is correct. Separately, *commentary is distinguished from occurrence* before any event rule fires: an analyst note reacting to an earnings beat reuses that beat's vocabulary, so recognisable commentary is routed to the *opinion* class, keeping reaction pieces out of event features.

Table 9: **Round-trip long-only by sector** (intraday entry plus overnight carry). Same simulator and window as Table 3.

Sector	Arm	Sh.	PnL	DD%
Tech	1-stage, no ingest.	-1.25	$\$-339$	15.2
	1-stage + FinRAIN	$+0.84$	$\$+181$	9.7
	2-stage, no ingest.	$+1.19$	$\$+281$	11.0
	2-stage + FinRAIN	$+1.56$	$\$+462$	13.4
Health	1-stage, no ingest.	$+1.11$	$\$+224$	5.0
	1-stage + FinRAIN	$+1.75$	$\$+400$	5.2
	2-stage, no ingest.	-0.24	$\$-50$	6.8
	2-stage + FinRAIN	$+0.72$	$\$+169$	5.7
Finance	1-stage, no ingest.	-0.79	$\$-130$	9.6
	1-stage + FinRAIN	-0.19	$\$-18$	6.9
	2-stage, no ingest.	$+0.19$	$\$+130$	7.0
	2-stage + FinRAIN	-0.84	$\$-204$	9.3
Energy	1-stage, no ingest.	$+2.77$	$\$+806$	6.6
	1-stage + FinRAIN	$+2.45$	$\$+774$	9.1
	2-stage, no ingest.	$+2.20$	$\$+1{,}080$	16.3
	2-stage + FinRAIN	$+2.38$	$\$+1{,}237$	16.1

Table 10: **Company knowledge graph schema.** The concept core (company, concept, and person nodes with the structural relations) is curated; the event layer and the weighted EXPOSED_TO edges are mined from ticker-tagged coverage. Graph size is reported in Section 4.

Node kind	Contents
`company`	tracked firms, in universe and their off-universe rivals
`concept`	sectors, industries, peer groups, indices, brands, subsidiaries, macro themes
`person`	named executives
`event`	a dated, typed occurrence, clustered from articles
`event_type`	the nine event classes, queryable as objects
Relation	**From → to**
IN_SECTOR, IN_INDUSTRY, MEMBER_OF, LISTED_IN	company → concept
OWNS_BRAND, OWNS_SUBSIDIARY	company → concept
COMPETES_WITH	company → company
EXECUTIVE_OF	person → company
ABOUT, OF_TYPE, INVOLVES_THEME	event → company, class, theme
EXPOSED_TO	company → theme, weighted

Structural edges cannot express that one macroeconomic driver touches firms in unrelated industries, as a rate decision touches both banks and homebuilders. The graph therefore includes macro *theme* nodes; a theme detected in an article links its event, and aggregating these per company yields a weighted company–theme edge whose weight $W(n, \theta)$ counts the articles tying ticker n to theme θ, so exposure is learned from observed coverage rather than declared by hand. Raw weights cannot rank, since coverage volume dominates them and the most-covered tickers would lead on every theme. We therefore rank by *exposure lift*,

$$\mathrm{lift}(n, \theta) = \frac{W(n, \theta) \big/ \sum_{\theta'} W(n, \theta')}{\sum_{n'} W(n', \theta) \big/ \sum_{n', \theta'} W(n', \theta')}, \tag{3}$$

the share of n's theme-linked coverage spent on θ divided by θ's share of all such coverage, so $\mathrm{lift} = 1$ is average exposure. Because normalisation alone over-promotes companies whose theme coverage is too sparse for the ratio to be meaningful, eligibility also requires a floor on $W(n, \theta)$. The graph stores unnormalised counts, keeping the evidence inspectable, and applies Equation 3 at query time. Theme exposure is associational, and we use it only to inform scope magnitude, never direction.

Publisher registry. Domains are mapped to an institution class, a content category, and a credibility tier ranging from wire services and established financial newsrooms to content farms, together with a flag recording whether the publisher was found in the registry or assigned a fallback tier. The registry informs source attribution, and its stake flags inform stakeholder bias—a fact with no textual trace, since a press-release wire and an independent newsroom can be typographically indistinguishable.

Price and indicator tape. For each ticker we maintain a daily price panel from which prior returns over one, five, and twenty sessions, twenty-session realised volatility, a fourteen-session relative-strength index, position relative to five- and twenty-session moving averages, and an ordinary-least-squares trend estimate with its fit quality are derived. Every query resolves strictly from sessions preceding the article's own calendar date, and a ticker with insufficient prior history returns nothing rather than a partial record. This tape lets a claim about a move be checked against the move that actually occurred.

K TOOLS EXPOSED TO THE SCORING MODEL

Table 11 lists the six tools summarised in Section 3.2.2, with the arguments each accepts. Every tool is read-only and returns a structured record; the model is given at most six turns per article, a budget that in practice does not bind. Tool descriptions state what a tool returns rather than when to use it, so that the choice of evidence remains the model's.

Three tools take no argument that could reach forward in time. LOOKUP_PUBLISHER takes no arguments at all and resolves the publisher of the article currently being scored. SEARCH_PRIOR_COVERAGE and GET_PRICE_TAPE accept only a result count and a ticker respectively; their as-of cut is bound to the candidate article's own timestamp when the tool handler is constructed, which is what makes the no-lookahead property structural rather than a matter of prompt compliance.

Table 11: **Tools exposed to the scoring model.** Each maps onto a source from Section 3.2.1. LOOKUP_PUBLISHER, SEARCH_PRIOR_COVERAGE, and GET_PRICE_TAPE accept no date: their as-of cut is bound to the candidate article's timestamp, so no call can reach evidence that postdates it.

Tool (arguments)	Returns
LOOKUP_PUBLISHER (none)	domain, institution class, category, and credibility tier of this article's publisher, plus whether the publisher was found in the registry or given a fallback tier
LINK_ENTITIES (`text`, optional)	per-ticker mention counts over the universe, resolved through company names, brands, subsidiaries, and executives, with a scope prior that adds one-hop competitor propagation; defaults to this article's own text
GET_COMPETITORS (`ticker`)	the known rivals of a ticker, for implications an article never names
THEME_EXPOSURE (`theme`)	the universe companies most disproportionately exposed to a macro theme, with exposure lift
SEARCH_PRIOR_COVERAGE (`k`, optional)	maximum similarity to any earlier same-sector article in the lookback window, how many priors exceed a high-similarity cut, a novelty band, and the closest priors' titles and timestamps
GET_PRICE_TAPE (`ticker`)	prior close, 1/5/20-session returns, 20-session annualised realised volatility, RSI-14, position versus MA5 and MA20, and 20-session trend with slope and fit quality