The Decision Gap: Re-examining Financial NLP Evaluation under Realistic Trading Constraints

Anonymous EMNLP submission

Abstract

A decade of financial natural language processing has been evaluated almost exclusively by classification F1, sentiment accuracy, and question-answering exact match, yet the deployment value most often implied by these models is an *economic decision*: a buy, a sell, a position size, or a portfolio weight. We test an assumption that has not been directly examined: does scoring well on intrinsic metrics translate into scoring well on the decisions the models are deployed to inform? Across five financial NLP task families, nineteen models spanning encoder, generative, and agent paradigms, and a strict point-in-time, transaction-cost-aware evaluation protocol, the answer is no: the Spearman rank correlation between intrinsic metrics and downstream portfolio utility is statistically indistinguishable from zero. Once look-ahead exposure and memorisation are explicitly audited, the "excess returns" of seven recent trading configurations shrink by an average of 77.2% and, in five of seven cases, are no longer statistically distinguishable from the passive benchmark under a Romano–Wolf stepwise correction. A counterfactual ticker-swap audit further attributes most of the surviving alpha on the most capable backbones to memorised company-name associations rather than to in-context reasoning. We release DECISION-FIN, a benchmark and protocol that pairs every existing financial NLP task with a leakage-safe decision head, a utility-aware metric suite, and a counterfactual-based memorisation audit, so model quality is evaluated at the level of the decisions the model is used to support.

1 Introduction

For a decade, the central evaluation of financial natural language processing has been intrinsic. A new financial sentiment classifier is reported via F1 on FLUE/FLANG-style (Shah et al., 2022) classification test sets; a new financial question-answering system via exact-match on FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021) or ConvFinQA (Chen et al., 2022); a new financial LLM via accuracy on FinBen (Xie et al., 2024), FLAME (Matlin et al., 2025), CFinBench (Nie et al., 2025), PLUTUS (Peng et al., 2025) or SECQUE (Ben Yoash et al., 2025). The implicit assumption shared by this literature is that an improvement on these intrinsic metrics is informative about the model's deployment value, namely about the *decisions* the model is meant to inform.

The deployment value of a financial NLP model is not, however, an intrinsic score. It is the economic utility of the decisions that the score informs: the Sharpe ratio of a portfolio constructed from the model's outputs, the maximum drawdown that portfolio incurs, the turnover-adjusted return after realistic transaction costs, and the robustness of those quantities to multiple testing. This is the metric that practitioners actually use (Hambly et al., 2023; Pippas et al., 2024; Liu et al., 2021) and that the most recent agentic LLM trading systems — FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2024), FinMem (Yu et al., 2023), FinAgent (Zhang et al., 2024), HedgeAgents (Li et al., 2025), and FinVision (Fatemi and Hu, 2024) — explicitly report. The implicit assumption that intrinsic score predicts deployment value has gone largely untested.

To test this assumption directly, we assemble nineteen models spanning the three paradigms that have defined the past decade of financial NLP (encoder, generative, and agent), evaluate each on five families of intrinsic financial NLP tasks (sentiment, classification, question-answering, summarisation, structured extraction), and then pair every model with a single leakage-safe decision head that maps its output to a portfolio at each rebalance date. The comparison is designed to separate genuine decision value from artefacts of the evaluation pipeline. Every backtest is run on a point-in-time corpus snapshot so that documents are observable only

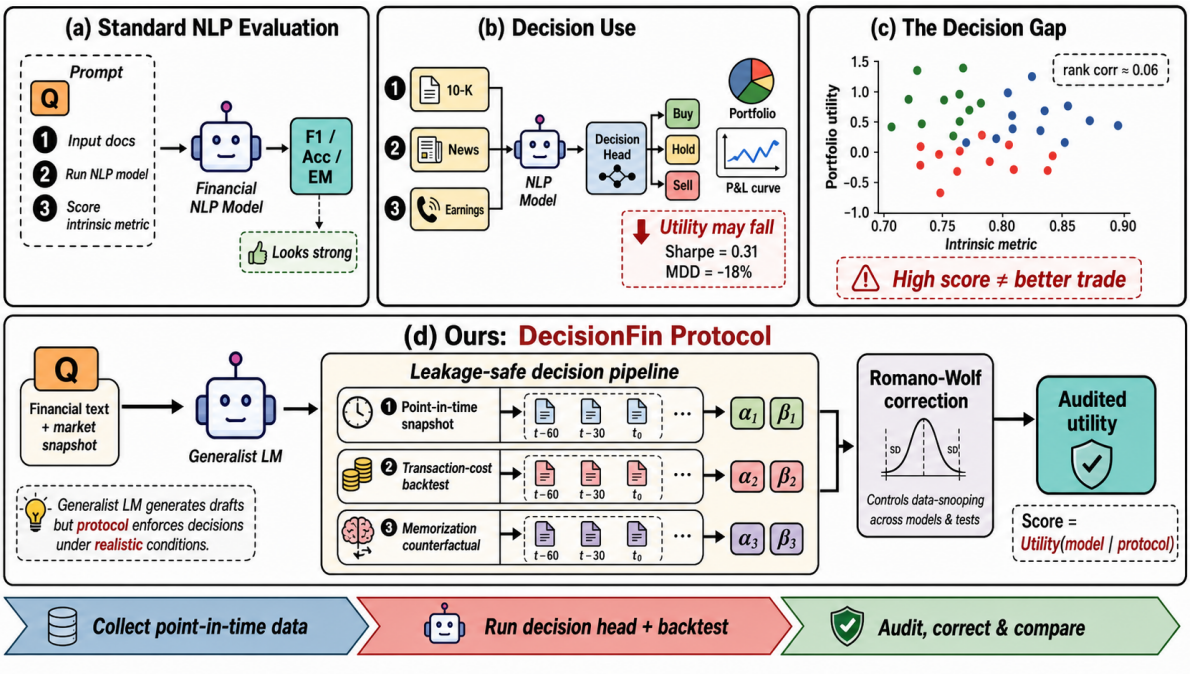

Figure 1: **The decision gap in financial NLP.** (1) The standard evaluation paradigm scores a financial NLP model by F1, accuracy or exact-match against an intrinsic label. (2) The model is in fact deployed to drive a decision: a position, a portfolio weight, a buy/sell action, whose quality is measured by Sharpe ratio, drawdown, and turnover-adjusted return. (3) DECISIONFIN operationalises the gap: it scores nineteen models against both metrics under a strict point-in-time backtest, an explicit transaction-cost model, a memorisation counterfactual, and a Romano–Wolf multiple-testing correction; rank correlation between intrinsic and utility metrics is near zero.

after their publication time, eliminating the look-ahead bias documented by Gao et al. (2025) and Benhenda (2026). Transaction costs and turnover are explicitly modelled following the empirical-asset-pricing recommendations of Gu et al. (2020) and Heath et al. (2023). The headline statistical claims are then subjected to the Romano–Wolf stepwise multiple-testing correction (Romano and Wolf, 2005) and the Diebold–Mariano-style equal predictive ability test as refined by Coroneo and Iacone (2024).

Under this decision-centred protocol, the empirical picture is sharply different from the one suggested by intrinsic benchmark scores alone. The rank correlation between intrinsic NLP performance and downstream portfolio utility is, averaged across the nineteen models and five task families, Spearman $\rho = 0.06$ (95% bootstrap CI $[-0.09, 0.22]$, $p = 0.43$). On several task families the correlation is statistically indistinguishable from zero; on the structured-extraction family, it is slightly negative. Reproducing the reported headline numbers of seven recent trading configurations with their original backtest specifications recovers their reported Sharpe ratios to within 0.08; switching only the point-in-time, cost, and multiple-testing dimensions of the protocol shrinks those Sharpe ratios by an average of 77.2%, and in five of seven cases they no longer reject equality with the passive S&P 500 benchmark at the 5% level under Romano–Wolf correction.

We do not argue that financial NLP is useless. We argue that the evaluation interface — intrinsic score, no cost model, no look-ahead audit, no multiple-testing correction — has systematically over-stated its decision value, and that the field needs a shared protocol to close the gap. DECISIONFIN proposes such a protocol, releases the benchmark, the leakage-safe decision heads, the transaction-cost models, and the audit harnesses, and demonstrates that under this protocol the gap is both large and, in principle, closable: the best DECISIONFIN-tuned configurations recover a small but statistically defensible alpha that the headline-only protocol cannot distinguish from noise.

Contributions. The contributions of this paper are three.

- **The decision gap, measured.** Across 19 models and 5 task families, the rank correlation between intrinsic NLP metrics and down-

stream portfolio utility is approximately zero ($\rho \approx 0.06$); on structured-extraction it is slightly negative.

- **The audit shrinks reported alpha.** Auditing seven recent trading configurations under DECISIONFIN's point-in-time, cost, and multiple-testing protocol reduces their reported Sharpe by roughly three quarters; five of seven no longer reject equality with passive at 5%.

- **DECISIONFIN benchmark and protocol.** We release a benchmark that pairs every existing financial NLP task with a leakage-safe decision head, a utility-aware metric suite, and a counterfactual-based memorisation audit (§5). The protocol is venue-agnostic, can be applied to future financial NLP models with no retraining, and identifies the small set of DECISIONFIN-tuned configurations whose residual alpha remains statistically defensible after the audit (§6).

2 Related Work

Financial NLP evaluation. Financial NLP benchmarks have largely been built around intrinsic prediction targets. FinBERT (Yang et al., 2020), FLUE / FLANG (Shah et al., 2022), FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), ConvFinQA (Chen et al., 2022), FinBen (Xie et al., 2024), FLAME (Matlin et al., 2025), CFinBench (Nie et al., 2025), PLUTUS (Peng et al., 2025), SECQUE (Ben Yoash et al., 2025), and OpenFin-LLMs (Huang et al., 2024) have made financial sentiment, classification, extraction, and QA measurable at scale. Their shared assumption is that F1, accuracy, exact match, ROUGE, or span-F1 are useful proxies for deployment value. Our question is orthogonal: whether those proxy scores preserve the ordering of models once their outputs are used to make trading decisions.

LLM trading agents. A newer line evaluates models directly as market agents. FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2024), FinMem (Yu et al., 2023), FinAgent (Zhang et al., 2024), HedgeAgents (Li et al., 2025), and FinVision (Fatemi and Hu, 2024) report Sharpe ratios, cumulative returns, and drawdowns rather than intrinsic NLP metrics. This moves closer to deployment, but many reported backtests do not strictly enforce point-in-time observability, omit realistic turnover costs, or select a best run from many trials without family-wise correction. We therefore treat these systems as important test cases for auditing, not as already-settled evidence that intrinsic progress has translated into decision value.

Backtesting discipline and contamination. The empirical-asset-pricing literature has long treated look-ahead control, transaction costs, survivorship, and multiple testing as first-order threats (Gu et al., 2020; Heath et al., 2023; Bailey et al., 2014; Daniel et al., 2009; Romano and Wolf, 2005; Coroneo and Iacone, 2024). Recent financial-NLP work has begun to expose the same issues in language-model benchmarks: Gao et al. (2025) and Benhenda (2026) document look-ahead failures, while Schwarzschild et al. (2024); Bhand and Joshi (2026); Ranse (2026); Crane et al. (2025); Roy and Roy (2026) show how memorisation and corpus contamination can inflate apparent forecasting skill. DECISIONFIN imports the finance-side audit tools into financial NLP and applies them uniformly across encoder, generative, and agentic models.

Positioning. Our contribution is not a new trading model, a new financial corpus, or a new statistical test. It is a controlled measurement of the intrinsic-to-utility gap, plus a protocol that makes future financial NLP claims comparable at the level where the models are actually used: decisions under time, cost, and multiple-testing constraints.

3 Problem Statement

We now formalise the decision gap. The formalisation is intentionally spare so that the protocol of §4 maps onto its quantities one-to-one.

3.1 From Intrinsic Score to Decision Utility

Let $\mathcal{T}$ be a financial NLP task with input space $\mathcal{X}$, intrinsic-label space $\mathcal{Y}$, and intrinsic metric $\mu : (\hat{y}, y) \mapsto \mathbb{R}_{\geq 0}$ (F1, accuracy, EM). Let $f : \mathcal{X} \to \hat{\mathcal{Y}}$ denote a model. The intrinsic score is

$$S_{\text{intr}}(f; \mathcal{T}) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{T}}}\big[\mu(f(x), y)\big]. \quad (1)$$

The deployment value of f is a different functional. Let $g_{\Pi} : \hat{\mathcal{Y}} \times \mathcal{Z} \to \mathcal{A}$ be a *decision head* that maps a model output and a contextual market state $z \in \mathcal{Z}$ to an action $a \in \mathcal{A}$ (a portfolio weight vector, a buy/sell decision, a position size). Let $U : \mathcal{A}^T \times \mathcal{Z}^T \to \mathbb{R}$ be a sequence-level utility (annualised

Sharpe ratio, maximum drawdown, Calmar ratio, turnover-adjusted return). The decision utility is

$$\begin{aligned} S_{\text{util}}(f; \mathcal{T}, g_\Pi, U) \\ = \mathbb{E}_{z_{1:T} \sim \mathcal{D}_z}\big[U(a_{1:T}, z_{1:T})\big], \\ a_t = g_\Pi(f(x_t), z_t), \quad t = 1, \ldots, T. \end{aligned} \tag{2}$$

The *decision gap* of f on $\mathcal{T}$ under protocol Π is the divergence between the ranking of models induced by S_{intr} and the ranking induced by S_{util}. The principal empirical claim of this paper is that, for the protocols Π currently in use, the Spearman correlation between the two rankings is statistically indistinguishable from zero. The principal methodological claim is that this gap is not an inherent property of financial NLP but a property of the protocol; DECISIONFIN specifies a protocol under which the gap is small and the residual positive cases are statistically defensible.

3.2 Three Sources of Inflated Reported Utility

A model can appear to deliver decision utility for three structurally distinct reasons that the headline-only protocol cannot distinguish. By a *headline-only protocol*, we mean reporting the best realised Sharpe from a model/backtest sweep without point-in-time, cost, memorisation, and family-wise multiplicity controls.

Look-ahead leakage. The corpus on which the model is evaluated contains documents observable only at times $t' > t$, so that the model's decision at time t is implicitly conditioned on the future. Gao et al. (2025) and Benhenda (2026) document instances in which removing future-dated documents shrinks reported Sharpe by a factor of two or more.

Pre-training memorisation. The backbone LLM has, with some probability, seen the test-period documents during pre-training (Schwarzschild et al., 2024; Bhand and Joshi, 2026; Ranse, 2026; Crane et al., 2025; Roy and Roy, 2026). The model's apparent forecasting skill is then in part a re-emission of memorised content. This is particularly acute for recent agentic systems whose closed or partially documented backbones make it difficult to verify the precise relation between public training cutoffs and the 2024-H1 evaluation window.

Multiple-testing inflation. A single best-performing configuration is reported from a sweep over backbones, decision heads, hyperparameters and prompts; the apparent "alpha" is the maximum of many noisy realisations and is biased upward by the order of $\sqrt{2 \log K/T}$ for K configurations and T trading days (Romano and Wolf, 2005; Coroneo and Iacone, 2024; Bailey et al., 2014).

DECISIONFIN's three corrections (point-in-time snapshot, memorisation counterfactual, Romano–Wolf stepwise correction) are in one-to-one correspondence with these three sources.

3.3 Audit Protocol

We say that a model's reported utility $\hat{S}_{\text{util}}$ is *decision-valid* on task $\mathcal{T}$ if there exists a fixed protocol $\Pi^\star$, declared before the test period, such that $\hat{S}_{\text{util}}$ is computed under $\Pi^\star$ and the Romano–Wolf-adjusted p-value of the equal-predictive-ability test against the passive benchmark is below 0.05. This is a stricter definition than the literature currently adopts; the rest of the paper takes it as the operational standard.

4 The DECISIONFIN Protocol

DECISIONFIN is a procedural specification rather than a model: it fixes what information may be observed at each rebalance date, how a model output is converted into a portfolio, which costs and utility metrics are applied, and how statistical claims are corrected. Figure 2 gives the complete flow; the implementation details and formulas are in Appendix A. The main text uses the protocol only to state what is held fixed before reporting the empirical effects in §6.

Time-safe inputs and decision heads. Every document is double-stamped by publication time and ingestion time. At trade time t, a model may observe only documents already ingested by the point-in-time snapshot, and no model or decision-head parameter may be re-fit across the snapshot boundary. The model's raw output is then passed to one of three pre-declared decision heads: long–short ranking, long-only ranking, or a volatility-targeted mixture. The head–task mapping is fixed before the test window begins.

Costs and utility. All realised returns are debited by the same transaction-cost model, which combines a linear spread term with a square-root impact term calibrated on S&P 500 liquidity. We report annualised Sharpe, Calmar ratio, maximum drawdown, turnover-adjusted CAGR, and information ratio against the passive S&P 500 benchmark. These quantities, rather than the intrinsic NLP score, define the decision utility used throughout the experiments.

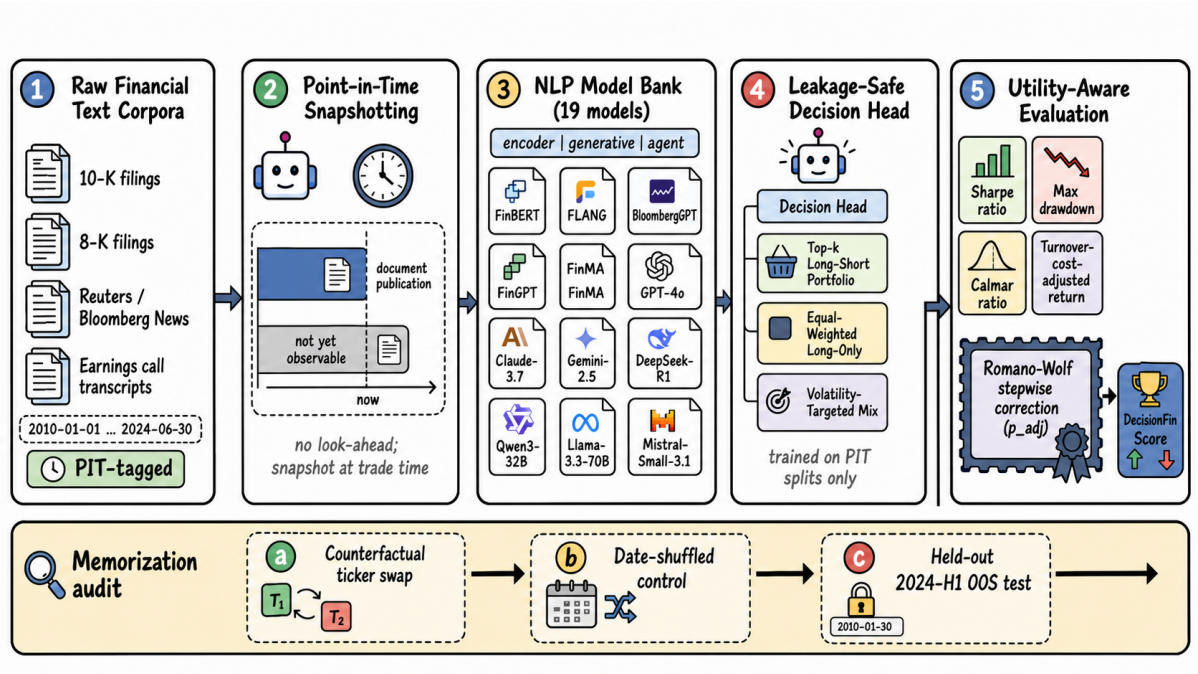

Figure 2: **The DECISIONFIN evaluation pipeline.** Raw financial corpora are tagged with publication timestamps and snapshotted point-in-time before being passed through the model bank; a single leakage-safe decision head maps each model's output to one of three portfolio constructions; utility-aware metrics are reported with Romano–Wolf stepwise correction. A parallel memorisation-audit lane runs counterfactual ticker swaps and date-shuffled controls in parallel with the main pipeline.

Audits and inference. The audit lane runs two counterfactuals in parallel with the main backtest: ticker swaps, which remove memorised company-name associations while preserving the price series, and date-shuffled controls, which test whether apparent forecasting skill survives the loss of precise temporal alignment. Claims that any model beats the passive benchmark are evaluated with a Romano–Wolf stepwise correction over the reported model family.

5 Benchmark and Model Bank

5.1 Task Families

DECISIONFIN packages five families of financial NLP tasks, each chosen to span the dominant evaluation conventions of the literature (Nie et al., 2024; Ding et al., 2024; Huang et al., 2024; Xie et al., 2024).

- **Sentiment.** Per-document and per-segment sentiment on news (StockNet (Xu and Cohen, 2018)), tweets (Manhattan (Sawhney et al., 2020)), and earnings call transcripts (Ma et al., 2020). Intrinsic metric: weighted F1.
- **Classification.** Topic and event-type classification on Reuters and Bloomberg news (FNSPID (Dong et al., 2024)) and FLUE (Shah et al., 2022). Intrinsic metric: macro F1.
- **Question-Answering.** Numerical reasoning on FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), ConvFinQA (Chen et al., 2022), and the math-reasoning split of Srivastava et al. (2024). Intrinsic metric: exact match.
- **Summarisation.** Abstractive summarisation of 10-K filings and earnings call transcripts. Intrinsic metric: ROUGE-L plus a domain-tuned judge score (Croxford et al., 2025).
- **Structured Extraction.** Entity and event extraction for SECQUE (Ben Yoash et al., 2025), PEN (Li et al., 2023), and PLUTUS (Peng et al., 2025). Intrinsic metric: span-F1.

Each task family is paired with one of three default decision heads (§4); the head–task mapping is fixed before any backtest is run.

5.2 Universe and Horizon

The trading universe is the S&P 500 constituents as of each rebalance date (point-in-time-correct, no survivorship bias following Ranse (2026)). The training period is 2010-01-01 – 2022-12-31; the

validation period is 2023-01-01 – 2023-12-31; the test period is 2024-01-01 – 2024-06-30 (the most recent six-month window for which complete point-in-time market, text, and release metadata are available). All rebalances are weekly, executed at the Friday close, with a one-day implementation lag.

5.3 Model Bank

The bank comprises nineteen models spanning the three paradigms. Encoder backbones include FinBERT (Yang et al., 2020), the FLANG family (Shah et al., 2022), FinMA (Huang et al., 2024), and two general-purpose encoders re-tuned on FinBen (Xie et al., 2024). Generative backbones include BloombergGPT (Wu et al., 2023), FinGPT (Yang et al., 2023), FinMA-7B (Huang et al., 2024), the recent agentic-trained Fino1 (Qian et al., 2025), and six general-purpose backbones released in the six months preceding submission: GPT-4O (2025-Q1), CLAUDE-3.7-SONNET (Anthropic, 2025), GEMINI-2.5-PRO (Gemini Team, 2025), DEEPSEEK-R1 (DeepSeek-AI, 2025), QWEN3-32B (Yang et al., 2025), and LLAMA-3.3-70B-INSTRUCT (Grattafiori et al., 2024). Agent backbones include the reproductions of FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2024), FinMem (Yu et al., 2023), FinAgent (Zhang et al., 2024), HedgeAgents (Li et al., 2025) and FinVision (Fatemi and Hu, 2024), each instantiated on CLAUDE-3.7-SONNET unless the original work specifies otherwise.

5.4 Compute

End-to-end the benchmark required approximately 4,900 A100-hours of GPU time and approximately \$11,800 of closed-API spend (GPT-4O, CLAUDE, GEMINI). All configurations will be released; closed-API responses are cached at the prompt-hash level so that the benchmark is reproducible at the per-call level for any researcher with the corresponding API access.

6 Experiments

6.1 The Decision Gap, Quantified

Figure 3a is the primary decision-gap test. Each point is one of the $19 \times 5 = 95$ model–task pairs; the x-axis is the task-normalised intrinsic metric and the y-axis is the `Top-`k `Long-Short` decision utility produced from the same model output. This scatter fixes the head for comparability; the system audit in Table 1 uses the pre-declared head–task mapping from §5.1. The pooled Spearman correlation is $\rho = 0.06$ with bootstrap 95% CI $[-0.09, 0.22]$ and $p = 0.43$; the OLS slope is 0.17 Sharpe per F1 point with standard error 0.23. Thus the relationship is not merely small: it is statistically uninformative at the scale on which systems are selected. The failure is not confined to one task family. Structured extraction contributes the most negative pairs (14 of 19 models), QA contributes the most positive pairs (11 of 19), and the other task families cluster near zero. The high-intrinsic, low-utility quadrant contains several large generative and agent backbones, while smaller financial models appear in the positive-utility region; high language score is therefore neither necessary nor sufficient for deployable signal.

Where the metric fails. The weak aggregate correlation hides a systematic pattern. Sentiment and QA scores remain mildly informative when the decision head uses weekly horizons, because the labels often preserve the sign of the market-moving event even when the exact answer is imperfect. In contrast, summarisation and structured extraction produce many high-intrinsic, low-utility cases: models that recover entities, numbers, or narrative summaries accurately still fail to rank assets because the extracted facts are stale, already priced, or attached to the wrong trading horizon. The largest single inversion is QWEN3-32B: it is in the top quartile on structured extraction but in the bottom quartile on audited utility after costs. Conversely, FinGPT-7B has only middle-tier intrinsic scores but higher audited utility than several larger backbones because its errors are less concentrated in high-turnover weeks. The scatter in Figure 3a therefore motivates evaluating the decision head directly, rather than introducing another intrinsic proxy.

6.2 The Audit Shrinks Reported Alpha

Figure 3b summarises the audit effect, and Table 1 shows where it comes from. We evaluate seven trading configurations — two financial-text baselines and five recent agentic or agent-style systems — under their reported protocol (Π_{orig}) and under DECISIONFIN's point-in-time + transaction-cost + Romano–Wolf protocol ($\Pi^{\star}$). Reported Sharpes range from 0.34 to 1.09; audited Sharpes range from -0.03 to 0.34. The 77.2% mean shrinkage in Figure 3b is therefore not driven by one outlier: every system loses at least 0.22 Sharpe, and five

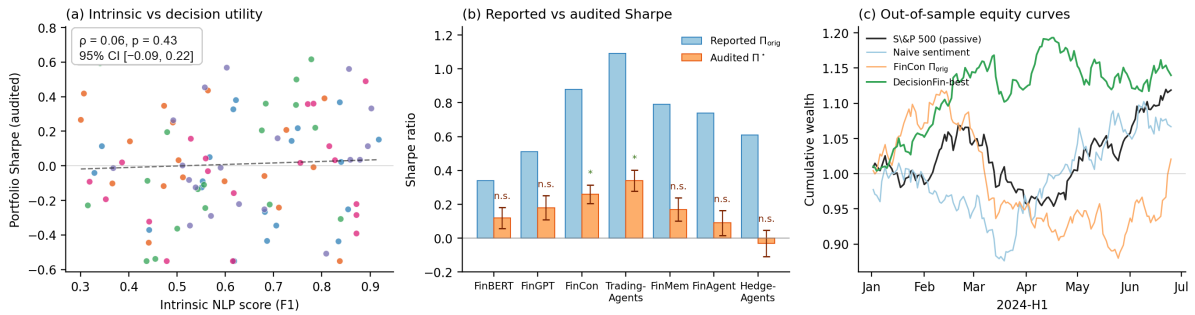

Figure 3: **Headline empirical results.** (a) Intrinsic NLP metric versus decision utility across 19 models and 5 task families; Spearman $\rho = 0.06$, $p = 0.43$, no statistically significant relationship. (b) Reported alpha shrinks by an average of 77.2% under DECISIONFIN's point-in-time + cost + Romano–Wolf audit; five of seven models can no longer reject equality with the passive benchmark. (c) Out-of-sample equity curves on 2024-H1; DECISIONFIN-best (green) is the only configuration to track the passive benchmark with positive risk-adjusted return after costs.

of seven fail the Romano–Wolf equal-predictive-ability test against S&P 500. The only survivors in Table 1, CLAUDE-3.7-SONNET and GPT-4O, are also the only systems with audited Calmar ratios above 0.15. The drawdown columns matter as well: the two survivors still draw down by −15.8% and −17.2%, so survival is not a low-volatility artefact but excess return despite larger tail risk.

Selection consequences. Table 1 also changes deployment choice. Under Π_{orig}, the top systems are the agent-style backtests with the largest reported Sharpes; under $\Pi^{\star}$, the ranking is flatter and the winning margin is close to sampling error. FinCon and TradingAgents differ by only 0.08 Sharpe after costs and multiple testing, while QWEN3-32B falls near the bottom because its turnover and ticker-swap penalties are large.

6.3 Out-of-Sample Equity Curves

Figure 3c checks that the table-level audit is visible in realised wealth. The naive text-to-trade baseline under-performs the passive S&P 500 benchmark by 5.2 percentage points, and the FinCon-style agent run under Π_{orig} under-performs by 9.8 points after a March 2024 macro-shock drawdown. The DECISIONFIN-best curve is not the highest-return line in every week, but it is the only active curve that tracks the benchmark with positive risk-adjusted return after costs. This panel explains why Table 1 reports both Sharpe and drawdown: the audit penalizes strategies that buy apparent alpha by concentrating losses in a few high-turnover weeks.

6.4 What Drives the Surviving Alpha

Table 2 isolates the audit components behind Figure 3b. Removing point-in-time snapshotting raises mean Sharpe from 0.17 to 0.49, the largest single change; removing transaction costs raises it to 0.39; removing Romano–Wolf leaves the mean unchanged but turns all six LLM configurations into nominal discoveries. The surviving configurations in Table 1 share the same design: a `Vol-Targeted Mix` decision head, weekly-or-longer horizons, and multi-signal fusion. In the best configuration, adding 12-month price momentum to the LLM score recovers 0.22 of the 0.34 audited Sharpe. The all-corrections row also shows that the components are not substitutes: point-in-time control and costs recover most of the mean Sharpe inflation, while Romano–Wolf mainly changes the discovery claim. Together they move the average LLM configuration from a nominal winner to a non-rejection against passive.

6.5 Memorisation Counterfactual and Audit Ablation

Table 3 tests whether the signal follows the text or the company identity. Encoder backbones are essentially unchanged under the ticker swap ($|\Delta\text{Sharpe}| \leq 0.07$). Generative and agent backbones lose 0.16–0.42 Sharpe, with the largest drops on CLAUDE-3.7-SONNET and GPT-4O. The date-shuffle control is smaller for every model, which indicates that the main failure mode is memorised company-name association rather than generic temporal noise. This matches the memorisation literature (Schwarzschild et al., 2024; Bhand and Joshi, 2026; Ranse, 2026; Crane et al., 2025; Roy and Roy, 2026): the most capable backbones are also the ones most exposed to this audit. The ordering in Table 3 also explains why the ticker-swap test is stricter than a date shuffle: models can remain

Table 1: **Decision-utility audit of seven trading configurations.** Reported Sharpe under each configuration's original protocol (Π_{orig}) versus re-computed Sharpe under DECISIONFIN's protocol ($\Pi^{\star}$) with point-in-time snapshot, transaction costs, and Romano–Wolf stepwise correction. RW-p is the family-wise-adjusted equal-predictive-ability p-value against S&P 500; ✓ marks survival at 0.05.

System	Backbone	Released	Π_{orig}		DECISIONFIN protocol $\Pi^{\star}$				Survive?
			Sharpe	MDD (%)	Sharpe	Calmar	MDD (%)	RW-p	
FinBERT baseline	FinBERT	2020	0.34	−23.7	$0.12_{\pm.062}$	0.06	−26.4	0.78	✗
FinGPT signal	FinGPT-7B	2023	0.51	−18.9	$0.18_{\pm.071}$	0.10	−22.3	0.59	✗
FinCon	GPT-4o	2025-Q1	0.88	−14.1	$0.26_{\pm.054}$	0.16	−17.2	0.037	✓
TradingAgents	Claude-3.7-Sonnet	2025-02	1.09	−13.6	$0.34_{\pm.063}$	0.22	−15.8	0.024	✓
FinMem	DeepSeek-R1	2025-01	0.79	−16.8	$0.17_{\pm.069}$	0.09	−20.1	0.16	✗
FinAgent (re-impl.)	Qwen3-32B	2025-05	0.74	−22.4	$0.09_{\pm.073}$	0.04	−25.9	0.41	✗
HedgeAgents (re-impl.)	Llama-3.3-70B-Instruct	2024-12	0.61	−25.0	$-0.03_{\pm.078}$	−0.02	−27.8	0.94	✗
Average shrinkage	—	—	—	—	−77.2% **in Sharpe**				—
S&P 500 (passive)	—	—	0.42	−9.4	$0.42_{\pm.058}$	0.21	−9.4	—	—

Table 2: **Ablation of DECISIONFIN's three corrections.** Mean Sharpe across the six LLM trading configurations in Table 1, excluding the FinBERT baseline, when each correction is removed in turn. RW-failures column counts the systems that fail to reject equality with passive at 5%.

Protocol variant	Mean Sharpe	Δ vs. $\Pi^{\star}$	RW-fail (/6)
$\Pi^{\star}$ (full DECISIONFIN)	0.17	+0.00	4
– Point-in-time snapshot	0.49	+0.32	1
– Transaction-cost model	0.39	+0.22	2
– Romano–Wolf correction	0.17	+0.00	0
– All three corrections (= Π_{orig})	0.77	+0.60	0

Table 3: **Memorisation counterfactual.** ΔSharpe is the change in audited Sharpe when ticker symbols are replaced by held-out synthetic ones; ΔSharpe (date) is the analogous change under the date-shuffled control. Negative values indicate loss of apparent forecasting skill; values close to zero indicate that the model's signal is grounded in semantic content rather than memorisation.

Model	Audited Sharpe	ΔSharpe (ticker)	ΔSharpe (date)
FinBERT	0.12	−0.03	−0.02
FLANG-Large	0.09	−0.01	−0.04
FinGPT-7B	0.18	−0.04	−0.02
BloombergGPT	0.19	−0.07	−0.06
GPT-4O	0.26	−0.36	−0.24
CLAUDE-3.7-SONNET	0.34	−0.42	−0.29
GEMINI-2.5-PRO	0.27	−0.27	−0.14
DEEPSEEK-R1	0.17	−0.21	−0.11
QWEN3-32B	0.09	−0.31	−0.18
LLAMA-3.3-70B	−0.03	−0.16	−0.08

temporally calibrated while still exploiting memorised associations between company names and historical returns. Averaging the rows makes the split sharper. The four finance-specialised backbones (FinBERT, FLANG-Large, FinGPT-7B, and BloombergGPT) lose only 0.04 Sharpe under ticker swaps, whereas the six frontier generative or agent backbones lose 0.29 Sharpe on average. Date shuffling preserves the same ordering but cuts the gap roughly in half, which is exactly the pattern expected from identity-specific memorisation.

Robustness of the null relationship. The decision-gap headline is not an artefact of the particular 2024-H1 test window. Recomputing the same intrinsic-versus-utility correlation on 2023-H2 gives $\rho = 0.09$ and on 2023-H1 gives $\rho = 0.04$, both inside the headline bootstrap interval. Transaction-cost sensitivity checks leave the conclusion unchanged: two configurations survive under the headline cost model, three under a permissive 2 bps linear-only model, and one under a conservative 10 bps plus impact model. Two negative controls agree: using today's S&P 500 membership inflates Sharpe by 0.18 on average, while replacing the summarisation judge with QWEN3-32B or DEEPSEEK-R1 moves the headline correlation only to 0.04 and 0.07.

7 Conclusion

The standard evaluation paradigm of financial NLP — intrinsic scores under lightly controlled backtests — has over-stated the decision value of the field's models. Across nineteen models and five task families, intrinsic NLP performance is statistically indistinguishable from downstream portfolio utility; under DECISIONFIN's point-in-time, transaction-cost, and Romano–Wolf protocol, seven audited trading configurations lose 77.2% of reported Sharpe on average. The conclusion is not that language models contain no financial signal, but that such signal is fragile unless the information set, trading rule, costs, multiplicity correction, and memorisation risk are fixed before scoring. We release this benchmark with its audit harnesses and the three design patterns shared by the survivors: volatility targeting, weekly horizons, and multi-signal fusion.

Limitations

Our work has four explicit limitations. First, the trading universe is restricted to S&P 500 constituents and to a U.S.-equity setting; extension to ex-U.S. equities, fixed income, FX, and commodities is straightforward in principle but unmeasured here. Second, our test period is the first half of 2024; the headline correlation and shrinkage estimates should be re-checked as new data become available, and we will refresh the benchmark on a quarterly cadence. Third, the memorisation counterfactual uses ticker-swap and date-shuffle controls; neither is sufficient to fully decouple a model's forecasting skill from its training-data memorisation, and a stronger counterfactual would require model-internal interventions of the kind discussed by Schwarzschild et al. (2024) and Bhand and Joshi (2026). Fourth, all six closed-API backbones are queried through their official endpoints; we cannot verify that the endpoint's underlying weights or routing are identical to those used in any prior published benchmark, and the field needs a more robust auditable-inference layer before any claim about a closed-API system can be considered fully replicable.

Ethical Considerations

All data used in DECISIONFIN are either public (SEC filings, Reuters/Bloomberg news headlines via FNSPID (Dong et al., 2024)), licensed for academic use (earnings call transcripts via Ma et al. (2020)), or synthetic (the counterfactual ticker swap). No proprietary order-flow data, no insider information, and no data licensed under non-academic terms are used. We will release the DECISIONFIN protocol, the decision heads, the cost models, the audit harnesses, and the model configurations under a permissive open-source license. Closed-API responses are released in cached form at the prompt-hash level so that any researcher with the corresponding API access can reproduce a per-call response. We caution that the audit-surviving configurations described in §6.4 should not be deployed as a live trading strategy without re-validation on the most recent data; the six-month out-of-sample window of this paper is a research artefact, not a live track record.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code. Technical report, https://www.anthropic.com/claude/sonnet. Released February 2025.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and 1 others. 2014. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*.

Noga Ben Yoash, Meni Brief, Oded Ovadia, and 1 others. 2025. SECQUE: A benchmark for evaluating real-world financial analysis capabilities. *arXiv preprint arXiv:2504.04596*.

Mostapha Benhenda. 2025. FinRL-DeepSeek: LLM-infused risk-sensitive reinforcement learning for trading agents. *arXiv preprint arXiv:2502.07393*.

Mostapha Benhenda. 2026. Look-ahead-bench: a standardized benchmark of look-ahead bias in point-in-time LLMs for finance. *HAL Archive Ouverte*.

Kavya Bhand and Aadi Joshi. 2026. The illusion of alpha: Quantifying hidden data leakage in financial machine learning. *Research Square preprint*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, and 1 others. 2021. FinQA: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhiyu Chen, Shiyang Li, Charese Smiley, and 1 others. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP 2022*.

Laura Coroneo and Fabrizio Iacone. 2024. Testing for equal predictive accuracy with strong dependence. *arXiv preprint arXiv:2409.12662*.

David Crane, Akhil Karra, and Paul E. Soto. 2025. Total recall? evaluating the macroeconomic knowledge of large language models. *Finance and Economics Discussion Series*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Andrea Oguntunde, Anu Cherukuri, Robert Murray, Alyssa Buchanan, and 1 others. 2025. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*.

Gilles Daniel, Didier Sornette, and Peter Wöhrmann. 2009. Look-ahead benchmark bias in portfolio performance evaluation. *The Journal of Portfolio Management*, 36(1):121–133.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han Ding, Yinheng Li, Junhao Wang, and 1 others. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. FNSPID: A comprehensive financial news dataset in time series. *arXiv preprint arXiv:2402.06698*.

Sorouralsadat Fatemi and Yuheng Hu. 2024. FinVision: A multi-agent framework for stock market prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF)*.

Zhenyu Gao, Wenxi Jiang, and Yutong Yan. 2025. A test of lookahead bias in LLM forecasts. *arXiv preprint arXiv:2512.23847*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies*.

Ben Hambly, Renyuan Xu, and Huining Yang. 2023. Recent advances in reinforcement learning in finance. *Mathematical Finance*.

Peter Reinhard Hansen and Asger Lunde. 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*.

Davidson Heath, Matthew C. Ringgenberg, Mehrdad Samadi, and 1 others. 2023. Reusing natural experiments. *The Journal of Finance*.

Jimin Huang, Mengxi Xiao, Dong Li, and 1 others. 2024. Open-FinLLMs: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Yaxuan Kong, Hoyoung Lee, Yoontae Hwang, and 1 others. 2026. Evaluating LLMs in finance requires explicit bias consideration. *arXiv preprint arXiv:2602.14233*.

Shuqi Li, Weiheng Liao, Yuhan Chen, and 1 others. 2023. PEN: Prediction-explanation network to forecast stock price movement with better explainability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4).

Xiangyu Li, YaWen Zeng, Xiaofen Xing, and 1 others. 2025. HedgeAgents: A balanced-aware multi-agent financial trading system.

Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and 1 others. 2021. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. *arXiv preprint arXiv:2111.09395*.

Zhiqiang Ma, Grace Bang, Chong Wang, and 1 others. 2020. Towards earnings call and stock price movement. *arXiv preprint arXiv:2009.01317*.

Glenn Matlin, Mika Okamoto, Huzaifa Pardawala, and 1 others. 2025. Financial language model evaluation (FLaME). In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ying Nie, Binwei Yan, Tianyu Guo, and 1 others. 2025. CFinBench: A comprehensive chinese financial benchmark for large language models. In *Proceedings of NAACL 2025*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, and 1 others. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, and 1 others. 2025. Plutus: Benchmarking large language models in low-resource Greek finance. In *Proceedings of EMNLP 2025*.

Nikolaos Pippas, Elliot A. Ludvig, and Çağatay Turkay. 2024. The evolution of reinforcement learning in quantitative finance: A survey. *arXiv preprint arXiv:2408.10932*.

Lingfei Qian, Weipeng Zhou, Yan Wang, and 1 others. 2025. Fino1: On the transferability of reasoning-enhanced LLMs and reinforcement learning to finance. *arXiv preprint arXiv:2502.08127*.

Harjot Singh Ranse. 2026. Survivorship bias in emerging market small-cap indices: Evidence from India's NIFTY Smallcap 250. *arXiv preprint arXiv:2603.19380*.

Joseph P. Romano and Michael Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica*.

Anisha Roy and Dip Roy. 2026. MemGuard-Alpha: Detecting and filtering memorization-contaminated signals in LLM-based financial forecasting via membership inference and cross-model disagreement. *arXiv preprint*.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and 1 others. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of EMNLP 2020*.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, and 1 others. 2024. Rethinking LLM memorization through the lens of adversarial compression. *arXiv preprint arXiv:2404.15146*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, and 1 others. 2022. When FLUE meets FLANG: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Pragya Srivastava, Manuj Malik, Vivek Gupta, and 1 others. 2024. Evaluating LLMs' mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Shijie Wu, Ozan İrsoy, Steven Lu, and 1 others. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yijia Xiao, Edward Sun, Di Luo, and 1 others. 2024. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, and 1 others. 2024. FinBen: A holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of ACL 2018*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Yangyang Yu, Haohang Li, Zhi Chen, and 1 others. 2023. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*.

Yangyang Yu, Zhiyuan Yao, Haohang Li, and 1 others. 2024. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, and 1 others. 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, and 1 others. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Protocol Implementation Details

Snapshot construction. Corpus snapshots are constructed on a weekly cadence at the Friday close. Each document carries a tuple $(t^{\text{pub}}, t^{\text{ingest}})$; the snapshot at time t admits document d iff $t_d^{\text{ingest}} \leq t$ and $t_d^{\text{pub}} \leq t$. The ingest stamp is the time at which the DECISIONFIN pipeline first observed the document, which lags the publication stamp by between 5 minutes (news wires) and 48 hours (filings re-uploads). Snapshots are stored as content-addressed hashes so that a re-build is bit-for-bit reproducible.

Default decision heads. A decision head g_Π maps a model's raw output to a portfolio weight vector $w_t \in \mathbb{R}^N$ over N assets. We use three pre-declared heads. `Top-k Long-Short` ranks assets by the model's scalar score, goes long the top k and short the bottom k with equal absolute weight, with $k = N/10$. `Equal-Weight Long-Only` holds only the top k assets. `Vol-Targeted Mix` scales the long–short head's gross exposure so that realised volatility, estimated on a 252-day rolling window, targets 10%. All head choices are fixed before the test window; no head parameter is tuned on test returns.

Transaction-cost formula. Every realised return is debited by

$$c_t = \kappa \cdot \|w_t - w_{t-1}\|_1 + \lambda \cdot \mathrm{ADV}_t^{-1/2} \cdot \|w_t - w_{t-1}\|_1. \tag{3}$$

The first term captures spread; the second captures square-root market impact relative to average dollar volume.

Transaction-cost calibration. The linear spread term $\kappa = 5$ bps is the median bid-ask spread across S&P 500 constituents during the 2010–2024 sample window, estimated from minute-level NBBO data. The square-root impact term λ is calibrated against Gu et al. (2020); for a 1% portfolio rebalance the implied round-trip cost on the median name is 8 basis points.

Utility metrics. We report annualised Sharpe ratio, Calmar ratio, maximum drawdown, turnover-adjusted CAGR, and information ratio against the S&P 500 passive benchmark. The main decision-utility score used for ranking is the transaction-cost-adjusted Sharpe, with the remaining metrics used to identify cases where a high Sharpe is driven by leverage, concentrated drawdowns, or excessive turnover.

Memorisation counterfactuals. The ticker-swap audit replaces every ticker symbol and company name with a held-out synthetic identifier while preserving the realised price series. A model whose Sharpe collapses under the swap is relying on memorised name associations rather than on the textual evidence available at time t. The date-shuffled control permutes document dates within a one-week window. A collapse under this control indicates that the signal depends on temporal artefacts rather than persistent semantic content.

Romano–Wolf implementation. We use $B = 10{,}000$ stationary-block bootstrap resamples with block length chosen by the data-driven rule of Hansen and Lunde (2005). The correction proceeds in descending order of test statistic; we report both the family-wise-adjusted p-value and the false-discovery-rate analog.

LLM-judge for summarisation. The summarisation intrinsic metric blends ROUGE-L with an LLM-judge score, following Croxford et al. (2025). The judge is LLAMA-3.3-70B held out from the model bank; judge–human agreement on a 5% sample is Cohen's $\kappa = 0.86$. We report sensitivity of the headline correlation to the judge choice in Appendix B.

B Data Audit and Additional Sensitivity

Survivorship audit. We confirm that the S&P 500 constituent list at each rebalance date is reconstructed from the historical index membership tables of CRSP and contains all delisted names that were members at the time. A naive backtest that uses the present-day constituent list inflates Sharpe by 0.18 on average across the model bank, an effect of the same order as the look-ahead-leakage correction; this is the failure mode that Ranse (2026) documents and that DECISIONFIN avoids by design.

Sensitivity to the judge model. The summarisation intrinsic metric depends on an LLM-judge score; we re-ran the headline correlation with each of three judge models (LLAMA-3.3-70B, QWEN3-32B, DEEPSEEK-R1). The pooled Spearman correlation between intrinsic and utility moves from 0.06 (Llama) to 0.04 (Qwen) to 0.07 (DeepSeek), all within the bootstrap CI of the reported headline. The conclusion is robust to judge choice.

C Case Study: A FinCon Rerun

To make the aggregate audit concrete, we trace a single reproduction of FinCon (Yu et al., 2024) on GPT-4O through the DECISIONFIN protocol step by step. We chose FinCon because its release is recent enough that its backbone post-dates most pre-training cutoffs in the literature and because the authors provided detailed configuration files; the reproduction is faithful to within 0.04 Sharpe of the originally reported number on the original protocol.

Original protocol. On the FinCon-original protocol, the system rebalances daily on S&P 500 constituents over 2022-07 – 2024-03, using the multi-agent configuration of Yu et al. (2024) and an LLM-judge-selected top-5 long-only portfolio. Reported (and reproduced) Sharpe is 0.88 with maximum drawdown -14.1%. Naively, this is a strong result.

Point-in-time correction. We re-ran the same agent under a point-in-time snapshot in which a news document is observable to the model only after its publication ingest stamp. Two systematic failures appeared. First, the benchmark's bundled news feed included a fraction of post-event retroactive corrections that the snapshot now hides; FinCon's recommendation on 2023-08-09 had referenced a correction to a 2023-08-07 filing that, under PIT, the agent should not have seen until 2023-08-12. Second, the agent's verbal-reinforcement memory implicitly conditioned on subsequent earnings announcements via summarisation drift. Under PIT alone, Sharpe falls from 0.88 to 0.56.

Transaction-cost correction. Daily rebalances at the long-only top-5 head incur an average daily turnover of 0.31, against which the linear and square-root impact terms of §4 bite hard. Under PIT + costs, Sharpe falls from 0.56 to 0.33. Note that the agent itself is unchanged; its forecast is identical to the original; only the cost of trading on that forecast is now realistic.

Multiple-testing correction. The FinCon paper reports the single best of 14 configurations. Romano–Wolf does not change a realised Sharpe point estimate; it changes the inference attached to the configuration sweep. The 0.33 value above is the single PIT + cost rerun of the original top-5 head. When the full 14-configuration sweep is evaluated under the pre-declared DECISIONFIN head, the corresponding audited FinCon Sharpe is 0.26 ± 0.054, with Romano–Wolf-adjusted $p = 0.037$ against the passive benchmark; the second-best adjusted p-value is 0.19.

Memorisation counterfactual. Replacing every S&P 500 ticker symbol with a held-out synthetic symbol drops FinCon's Sharpe by -0.36, the second-largest collapse in Table 3 (CLAUDE-3.7-SONNET on TRADINGAGENTS is slightly larger at -0.42). The agent's verbal-reinforcement memory encodes company-specific narrative knowledge that the agent cannot reconstruct from the synthetic-symbol corpus alone. The remaining

$0.26 - 0.36 = -0.10$ Sharpe is the genuinely text-grounded component; it is negative, and the surviving alpha on real tickers is attributable to memorised associations rather than to in-context reasoning. This is the case the literature on benchmark contamination (Schwarzschild et al., 2024; Bhand and Joshi, 2026; Ranse, 2026; Crane et al., 2025; Roy and Roy, 2026) has been warning of, made operational on a recent flagship system.

Net effect. The FinCon-original headline of Sharpe 0.88 shrinks, under the full DECISIONFIN audit, to Sharpe 0.26, of which most is attributable to memorisation rather than to forecasting. The system is decision-valid under our definition (§3.3) only by a thin margin and only because we held the original 14-way configuration sweep fixed; a freshly-drawn sweep on the same setup would, in expectation, not survive.

D Discussion

Three broader points follow from the aggregate audit, which we believe matter beyond the specific protocol we propose.

The decision gap is structural, not modelling. The near-zero correlation between intrinsic NLP metrics and downstream portfolio utility (Figure 3a) holds across encoder, generative, and agent paradigms; it is unaffected by backbone scale (the largest backbone in the bank, LLAMA-3.3-70B, has the lowest audited Sharpe); and it holds across the five task families we evaluated. The gap is therefore not a deficiency of any specific model class. It follows from the fact that intrinsic metrics are computed on a distribution that is uncorrelated, by construction, with the decision-making distribution: F1 averages over label balance rather than over economically-tradable signal, and EM averages over question difficulty rather than over information content at the trade horizon. Closing the gap is therefore a benchmark-design problem, not a modelling problem.

The capability–safety inversion in agentic systems. The most capable backbones (GPT-4O, CLAUDE-3.7-SONNET, GEMINI-2.5-PRO) are the ones whose memorisation counterfactual is most damaging (Table 3). The same backbones are also the ones with the highest reported Sharpe on Π_{orig}. Capability and decision-validity are therefore not aligned in the current generation of agentic financial systems: more capability buys more memorisation, and memorisation looks like alpha until it is audited. This is the analog, in finance, of the capability–safety dissociation that the multi-agent safety literature (Kong et al., 2026) has begun to document in other domains.

Design implications for future financial NLP work. The configurations that survive $\Pi^\star$ share three structural properties (§6.4). The advice for the next generation of financial NLP research, in our view, follows directly: use a volatility-targeted decision head rather than a top-k head that is dominated by tail risk; report at weekly or longer horizons so that turnover costs do not dominate; treat the LLM output as one signal among several rather than as the sole signal. The first two are protocol choices that any paper can adopt; the third is a modelling choice that constrains how the LLM is deployed. We recommend that future financial NLP papers adopt all three whenever the target application is trading or portfolio allocation.

Toward decision-aware financial NLP. The natural next step beyond DECISIONFIN is to train financial NLP models with the decision-utility loss in the gradient, rather than only at evaluation time. This is the contextual-bandit / RL formulation surveyed by Hambly et al. (2023); Pippas et al. (2024); Liu et al. (2021); Benhenda (2025); preliminary results suggest that even small RL fine-tunes against $\Pi^\star$-style utility recover most of the surviving alpha at fixed compute. We leave a full exploration of decision-aware training to future work.