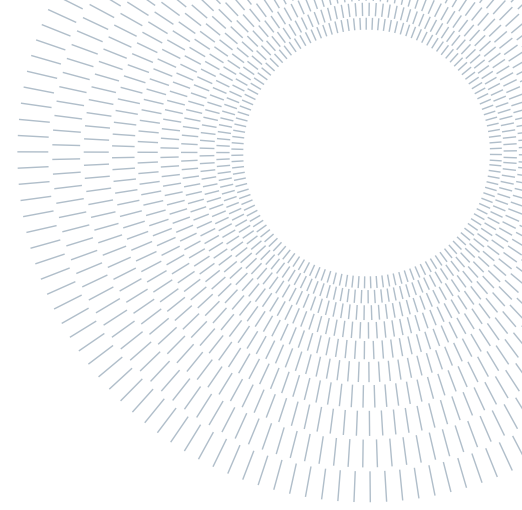




POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE



Does Multi-Source Data Beat Price Alone? A Controlled Ablation for Next-Day Stock Direction Prediction

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Mouadh Ltifi, 247697

Advisor:
Prof. Fabrizio Pittorino

Co-advisors:
Francesco Puoti
Prof. Molnár Bálint

Academic year:
2025-2026

Abstract: The multi-source financial-prediction literature reports gains over price-only baselines, paper after paper: news, social media, macroeconomic indicators, and inter-stock graphs, fused through increasingly elaborate architectures, are said to beat price alone. Yet the empirical-asset-pricing literature finds that most published return predictors fail to replicate and decay sharply once published, which sets a low prior for any new predictability claim on liquid US equities. This article takes that prior seriously and tests the multi-source claim for daily next-day direction prediction of 55 large-cap US equities (2015–2023), under a single controlled ablation of 2,468 experiments that holds the universe, folds, seeds, and Bonferroni-corrected paired statistics constant across every comparison. *No* combination of news, social, macro, or graph signals significantly beats the price-only baseline, and the null holds across every axis we vary: three fusion strategies, two prediction targets, four text encoders, and a tenfold parameter range across two structurally distinct architectures. Inside the null, adding backward-looking news *degrades* direction prediction in four of the five folds. We then reproduce one prominent published positive (MSGCA) and show, from the authors' own released code, that its reported gain does not survive a leakage-free protocol: the reported model is selected on the test set, and under leakage-free validation-based selection on the same data and model the advantage disappears. A second system (CAMEF) shows a different pathway: its headline reproduces from the authors' released weights but collapses under a chronological train/test split, which the released code leaves to a non-default flag. On this problem the estimator and the rigour of its evaluation matter more than which data sources are added.

Key-words: financial market prediction, multi-source data integration, ablation study, model-selection bias, reproducibility

1. Introduction

The multi-source financial-prediction literature is full of positive results. Methods that combine price with news, social media, macroeconomic indicators, or inter-stock graph structure, fused through tensor products, gated cross-attention, or graph neural networks, regularly report gains over price-only baselines [17, 18, 20, 29]. The

empirical-asset-pricing literature, looking at large cross-sections of US equities, paints a different picture: most published return predictors fail to replicate, and the survivors decay sharply once published [8, 9, 16].

These two records are usually read apart, and they are not flatly contradictory, because they test different things. The asset-pricing evidence concerns broad cross-sections of return predictors and sets a low prior: a new predictability claim on liquid US equities should be doubted until it survives disciplined evaluation. The multi-source results are individual systems on particular benchmarks, each reporting a gain over its own baseline, and each can hold in its own setting. The multi-source claim has rarely been tested against that prior in one controlled ablation, with a single universe, a tuned price-only baseline, and corrected statistics. This article fills that gap. The answer also matters in practice, since it decides whether scarce engineering effort should go into acquiring and fusing more data or into the estimator and its evaluation. Given the low prior, a careful negative result on this question is more useful than another optimistic positive.

This article runs a controlled empirical test of that question on a single calibrated apparatus. The data, the temporal folds, the fixed seeds, and the statistical machinery are held constant across every comparison, so that the only thing varying between a baseline and a candidate is the data source under test. The target is deliberately the hardest case for any predictive-signal claim: daily next-day direction prediction of liquid large-cap US equities, where information is absorbed into price quickly and any residual edge is small and contested (Section 3.1 develops why this makes a null an upper bound on the field’s claims).

Three research questions structure the study:

- **RQ1.** Does integrating multiple heterogeneous data sources improve prediction performance over a single-source (price-only) baseline?
- **RQ2.** What is the relative marginal contribution of each data source, and which sources provide the most value?
- **RQ3.** Under what conditions, if any, does multi-source integration help rather than hurt?

The answer to RQ1 is negative, and it is robust. No combination of news, social, macroeconomic, or graph signals significantly beats the price-only baseline after multiple-testing correction, and the null survives every robustness axis we test (enumerated in C1 below). The answers to RQ2 and RQ3 run with the same grain: no source carries positive marginal value — the only non-null marginal effect is backward-looking news, and it *degrades* prediction — and the only conditions under which a source changes the result are adverse or estimator-bound, never a genuine multi-source gain. This article makes two contributions:

- **C1 — A robust null.** On daily direction prediction of 55 US large-caps, no heterogeneous data source significantly improves over price alone, across two prediction targets, four text encoders, three fusion strategies, and two architectural families. Inside the null, the one consistent non-null signal is a degradation: adding backward-looking news lowers direction performance in four of the five folds.
- **C2 — A reproducibility audit of two reported positives.** Two prominent systems that report multi-source gains release their code, and we check whether the gains survive a leakage-free evaluation. For MSGCA [29] the reported headline is selected on the test set; under leakage-free validation-based selection on the same data and model, the advantage disappears. For CAMEF [28] the reported headline reproduces from the authors’ released weights, but it holds only under a non-temporal train/test split (a chronological split, set by a non-default flag, makes the same recipe diverge). Presented as reproduction results rather than verdicts on the systems, the two give concrete, reproducible reasons the field’s reported gains and the null reported here can coexist.

The remainder is organised conventionally. Section 2 is a background section that, for a reader new to the area, defines the prediction task and how models are compared, introduces the building blocks of a multi-source model, situates the work against the three literatures it sits between, and positions the specific prior systems against this study’s test in a comparison table. Section 3 then describes the data and method together, source by source; Section 4 reports the experiments; Section 5 interprets them; and Section 6 concludes.

2. Background

This section collects the concepts the rest of the article relies on, for a reader approaching multi-source financial prediction for the first time. It defines the prediction task and the way two models are compared (§2.1), then the vocabulary of multi-source modelling and the building blocks from which the later architectures are assembled (§2.2); and it situates the work against the three literatures it sits between, with the prior systems gathered in a comparison table (§2.3). A reader already at home in the area can skip to Section 3.

2.1. The prediction task, and how models are compared

The task. The problem studied here is *next-day directional prediction*: given everything known about a stock by the close of trading on day T , predict whether it closes higher or lower on day $T+1$. This is a binary

classification problem — up or down — deliberately simpler than predicting the size of the move, and it is the form in which most of the multi-source literature states its claims. Days on which the price barely moves carry no decisive direction and are set aside by a small “dead-zone” (§3.1), so the model is asked only about moves large enough to matter.

Why not accuracy, and what MCC measures. The obvious score for a binary predictor, the fraction of days it calls correctly, is misleading on market data, because the two outcomes are not equally common and one can be predicted trivially. If a market rises on 60% of days, a model that simply answers “up” every day is 60% accurate while having learned nothing. The *Matthews correlation coefficient* (MCC) [15] removes this loophole. Computed from all four cells of the confusion matrix (right and wrong calls of each direction), it is bounded in $[-1, +1]$: $+1$ is perfect prediction, -1 perfectly inverted, and, crucially, 0 **is chance**, the score the always-up model earns. Throughout, MCC is the measure of skill, and any MCC near zero should be read as “no better than guessing.” Because daily direction is genuinely hard, even a real edge yields only a small positive MCC, a few thousandths of a point, and the comparison method below is built to detect an effect that small, reliably.

Comparing two models fairly: paired testing. A model’s MCC swings far more from one market period to another than it does from adding a data source, so comparing two models by their average scores would drown the effect of interest in market noise. The study therefore compares models in *matched pairs*: the baseline and the candidate are run under identical conditions (the same time period, the same fixed seed), and it is the *difference* within each matched run that is measured. Because both models met the same market in each pair, the shared good or bad luck cancels, and what remains is the contribution of the data source alone. This is a standard paired comparison, and the quantity it produces is a per-pair difference in MCC.

How large, and how sure: effect size and correction. Two questions are asked of every such difference. *How large is it?* answered by the effect size, *Cohen’s d* [2]: the mean difference expressed in units of its own run-to-run variability, with the usual rough anchors of 0.2 (small), 0.5 (medium), and 0.8 (large). *How sure are we it is real?* answered by a p -value, the probability of seeing a difference this large if the data source truly did nothing. A subtlety matters here: testing many configurations gives many chances to stumble on a lucky one, so the p -values are *corrected for multiple comparisons* by the Bonferroni rule, which multiplies each by the number of tests in the family, guarding against a false positive produced simply by running enough tests. A result is treated as real only when it is both large enough (by d) and survives this correction.

2.2. Multi-source models: fusion and building blocks

What “multi-source” means, and how sources are combined. A price-only model sees one stream of numbers: the stock’s own past prices and quantities derived from them. A *multi-source* model adds others (the words of financial news, the volume and tone of social-media posts, macroeconomic indicators, or a graph of which stocks move together) in the hope that they carry information price alone does not. Each stream is first turned into a fixed-length numeric vector by its own *encoder*; the vectors are then *fused* into a single representation the predictor can use. Fusion is usually classed as data-level (combine raw inputs), feature-level (combine the encoded vectors), or model-level (combine separate models’ outputs) [20]; for time series the same taxonomy has been extended along fusion, alignment, and transference axes [10]. The three fusion strategies compared in this study are all feature-level and differ only in how they mix the encoded vectors:

- *concatenation* stacks the vectors and lets the predictor sort out the interactions;
- *self-attention* lets the streams weigh one another before pooling;
- *gated cross-attention* treats price as primary and lets a learned gate decide how much of each other stream to admit.

The building blocks. The models in this study are assembled from a small set of standard components, named here so the method section can use the terms freely:

- *technical indicators* — textbook functions of past price and volume (momentum, moving-average relationships, volatility measures) that summarise the price stream;
- *text encoder* — turns a sentence into a vector. This study uses FinBERT [1], a version of the BERT language model fine-tuned on financial text, kept *frozen*: a fixed translator from text to sentiment numbers, not trained further;
- *graph attention network (GAT)* [21] — blends each stock’s vector with those of its neighbours in a relation graph, so a stock’s representation reflects the firms it moves with;

- *price-sequence models* — a *feedforward* network sees a single day, a *long short-term memory* (LSTM) network reads an ordered window of recent days, and a *Temporal Fusion Transformer* (TFT) [13] is a larger forecaster that also learns which inputs and which past days deserve weight. They differ mainly in how much history they consume, and they let the study ask whether the null survives an order-of-magnitude increase in model capacity.

2.3. The three literatures this study sits between

This study sits between three literatures: the multi-source field that reports fusion gains over price, the empirical-asset-pricing record that sets a low prior for any such claim on liquid US equities, and a complexity-sceptical strand in time-series forecasting; the prior systems it is positioned against are gathered in Table 1.

Table 1: Representative multi-source prediction work: data used, headline claim, and relation to this study’s controlled ablation. The relation column is the point of the table: each system holds something fixed (a market, horizon, label rule, or baseline) that this study does not, so the reported gains are not mutually comparable and motivate the common-denominator test here. “best-on-test” denotes model/epoch selection on the test set as implemented in the released code (Section 4.2).

Study	Modalities / data	Reported claim	Relation to this study
MSGCA [29]	price, news, inter-stock graph; gated cross-attention	MCC $\approx$ 0.11 on BigData22; gains over baselines	We reproduce it and show the headline rests on best-on-test selection (Sec. 4.2).
CMTF [17]	price, news, financials, macro (tensor fusion)	direction-classification gains (precision/recall/F1), FTSE 100	Multi-source claim on a small universe; we test it at scale with seeds and folds.
MSMF [18]	price, news, sentiment (multi-scale + completion)	accuracy/F1 gains	Tests whether elaborate fusion beats plain concatenation; our fusion ablation finds no difference.
FinMulti-Time [24]	price, news, tables, K-line charts (S&P 500 + HS 300)	multimodal fusion gives moderate gains (Transformers)	Recent large-scale four-modal benchmark; we re-test the moderate-gain claim in our controlled ablation.
STONK [11]	price, news (5 text encoders)	MCC spread $\approx$ 0.09–0.33 by encoder	Motivates our four-encoder screen; no encoder beats price-only.
THGNN [23]	price, dynamic correlation graph	accuracy gains, S&P 500 / CSI 300	Our graph modality (GAT on GICS plus rolling correlation) is neutral.
CAMEF [28]	price, macro event text (causal)	event-window gains, FOMC	Our daily macro config does not beat price; its headline reproduces but holds only under a non-temporal split (§4.2).
StockNet [25]	price, tweets	tweets improve movement prediction	Our social modality (StockTwits) helps only an LSTM, not overall.
Ferraro & Sperli [5]	price, social media (GameStop)	finance-only data beats finance+social	Equity social-media ablation matching our social null.
Koval et al. [12]	forward-looking vs backward-looking text	forward text helps; backward does not	Explains our finding that backward-looking news degrades prediction.
Lopez-Lira and Tang [14]	GPT / news sentiment	Sharpe 6.54→1.22 (decaying)	Same-day news edge arbitrated away; corroborates the null at the daily horizon.
Tan et al. [19]	LLM time-series forecasters	simpler ablations win 67/78	Complexity does not substitute for signal; supports our capacity result.
Zhang et al. [27]	16 datasets, 7 domains (multimodal TS)	$p = 0.789$ (MSE) multimodal vs unimodal	Cross-domain null; this study is the within-domain depth complement.
Hou et al. [9]	452 anomalies (asset pricing)	65–82% fail to replicate	Sets the prior under which our null is the expected outcome.

The multi-source promise. The field holds that combining price with additional modalities — financial news, social-media sentiment, macroeconomic series, and inter-stock graphs — beats price alone, through progressively more elaborate fusion: from plain concatenation to tensor products, gated cross-attention, and graph

neural networks. But each system in Table 1 reports its gain against a different market, horizon, label rule, and often unspecified baseline, so the claims are not mutually comparable and none is a clean “gain over a tuned price-only model”. This study supplies the missing common denominator — one universe, one baseline, one protocol, with each modality added and removed in isolation. Two of the benchmarks double as setups for the tests here: STONK’s MCC spread of 0.09 to 0.33 across five encoders motivates the four-encoder screen, and StockNet’s tweets-plus-price result is what the social contrast re-tests.

Why a null is the expected outcome. The base rate for return-predictability on liquid US equities is low and independently established. Hou et al. [9] re-examine 452 anomalies under a common, microcap-robust protocol and find 65% fail a nominal $t > 1.96$ (a t -statistic is an estimated effect divided by its standard error; $|t| > 1.96$ is the conventional 5% significance threshold), rising to 82% under a multiple-testing-aware bar; McLean and Pontiff [16] track 97 predictors and find returns decay about 26% out of sample and 58% after publication; Harvey et al. [8] show that with hundreds of factors tested the conventional threshold guarantees false discoveries, and argue for $t > 3.0$. Independent in method, the three agree: an isolated positive on this universe is weak evidence until it survives disciplined evaluation, so the expected outcome of a careful test is a null.

Complexity is not a substitute for signal. A parallel forecasting strand finds that when information is the bottleneck, added capacity does not fill it. Tan et al. [19] replace the language model in three LLM forecasters with random or empty blocks, and the simpler ablations win 67 of 78 comparisons across 13 datasets; Zeng et al. [26] match or beat five transformer forecasters with a one-layer linear model on nine benchmarks; Zhang et al. [27] find no multimodal benefit across sixteen datasets in seven domains ($p = 0.789$ on MSE, 0.247 on MAE), a breadth-first null this study complements with within-domain depth. The text channel adds its own version: Koval et al. [12] find only forward-looking text, not backward-looking news, carries forecasting value, and Lopez-Lira and Tang [14] show news-sentiment Sharpe ratios decaying from 6.54 to 1.22 as the edge is arbitrated away, with finance-tuned encoders even posting negative drift. Together these set the expectation that more data and more capacity, absent new information, leave a price-only baseline unmoved; and where a released protocol can be checked, its rigour varies, as Section 4.2 shows for one prominent headline.

3. Method and data

The study is a falsification-oriented controlled evaluation [22]: a fixed prediction task, a single price-only baseline, and a grid of modality additions, each judged against a pre-specified statistical bar. This section describes the universe and target, then each data source together with the evidence that motivates testing it, then the architectures and the statistical protocol.

3.1. Universe, target, and folds

The universe is 55 large-cap US equities — the five largest by market capitalisation in each of the eleven sectors of the Global Industry Classification Standard (GICS),¹ ranked as of the end of the sample period and held fixed across all folds (a balanced panel with full 2015–2023 coverage; no mid-sample constituent changes are modelled). Two features of this universe are deliberate. The sector balance prevents any one industry from dominating the pooled effect estimate, so a result is not an artefact of, say, the technology sector’s behaviour in a particular window. The large-cap restriction is deliberately adversarial to the hypothesis: large-cap US equities have the deepest liquidity, the heaviest analyst coverage, and the fastest absorption of public information into price, which makes them the hardest case for any claim that an alternative data source contains exploitable residual signal. If multi-source integration helps anywhere, it should help least here. A null on this universe is therefore an upper bound on the field’s positive claims, and a positive here would have generalised to less efficient markets far more readily than the reverse.

Ranking by end-of-sample size invites one selection caveat, which deserves a direct answer: it makes the universe a survivor panel — these are the firms that grew to dominate their sectors. Whatever a survivor’s upward drift does to raw predictability, it does to the price-only baseline and to every multi-source configuration alike, because they all run on the same universe. The quantities this study tests are paired deltas of a configuration against that baseline on matched folds and seeds, so a universe-wide trend is common to both sides of each contrast and cancels from the marginal effect. Survivorship is therefore unlikely to manufacture, or to mask, a marginal source effect: it would have to act differently on one data source than on price, and nothing in the design points that way.

¹Developed jointly by MSCI and S&P Dow Jones Indices.

With the universe fixed, the prediction target is the sign of the next-day log return $r_{t,d+1} = \log(p_{t,d+1}/p_{t,d})$, with a symmetric 0.5% dead-zone that drops near-flat days so that each retained label reflects a decisive move rather than microstructure noise; the dead-zone removes roughly a third of the days and the retained set is close to balanced (about 38,800 down against 44,300 up). The primary metric is the Matthews correlation coefficient (MCC; defined and motivated in §2.1), on which zero is chance and a real edge on daily direction shows up as a small positive value. The secondary target is next-day realised volatility (the absolute return), evaluated by R^2 ; it tests whether the null is specific to the direction problem or extends to a continuous target with different statistical properties (§4.5).

That target is evaluated across time on five expanding-window temporal folds (F0–F4), each training on all data up to a cut-off and testing on the following window, so no future information enters training. The cut-off advances across folds, so each fold’s training set strictly contains the previous one’s; the smallest fold (F0) trains on roughly 23,000 stock-days and the largest (F4) on roughly 54,000 (Table 2). The five windows were chosen to span qualitatively distinct market regimes (the pre-COVID period and crash, the 2020–2021 recovery, the 2021–2022 transition, the 2022 bear market, and the 2023 rally), so the same model is evaluated across conditions rather than on a single hold-out. This matters for a null result: a configuration that helped only in one regime would surface as a positive in the corresponding fold, and the design gives every configuration five such chances. All statistical comparisons are paired on the (fold, seed) cell, which absorbs the large fold-to-fold variance in MCC. Every configuration is run under the same three fixed seeds, identical across configurations, so each run is exactly reproducible and a baseline–candidate pair differs only in the source under test.

Table 2: The five expanding-window folds. Training windows nest (each fold contains the previous) and the test windows are disjoint and chronological. Sizes are stock-days for the price-only baseline after the dead-zone filter; the validation set is the chronologically last 20% (by calendar days) of each fold’s training window, carved per fold and disjoint from the test window, with early stopping on validation MCC and the test set read once.

Fold	Test window	Test-window regime	Train	Val	Test
F0	2019-07 – 2020-06	pre-COVID and the COVID crash	23,116	6,372	10,100
F1	2020-07 – 2021-06	2020–2021 recovery	30,474	9,112	9,851
F2	2021-07 – 2022-06	2021–2022 bull-to-bear transition	38,468	10,972	9,979
F3	2022-07 – 2023-06	2022 rate-hike bear market	46,657	12,721	9,994
F4	2023-07 – 2023-12	2023 generative-AI rally	54,070	15,339	4,514

Figure 1 draws the same folds to scale, each test window placed on the market path it must predict.

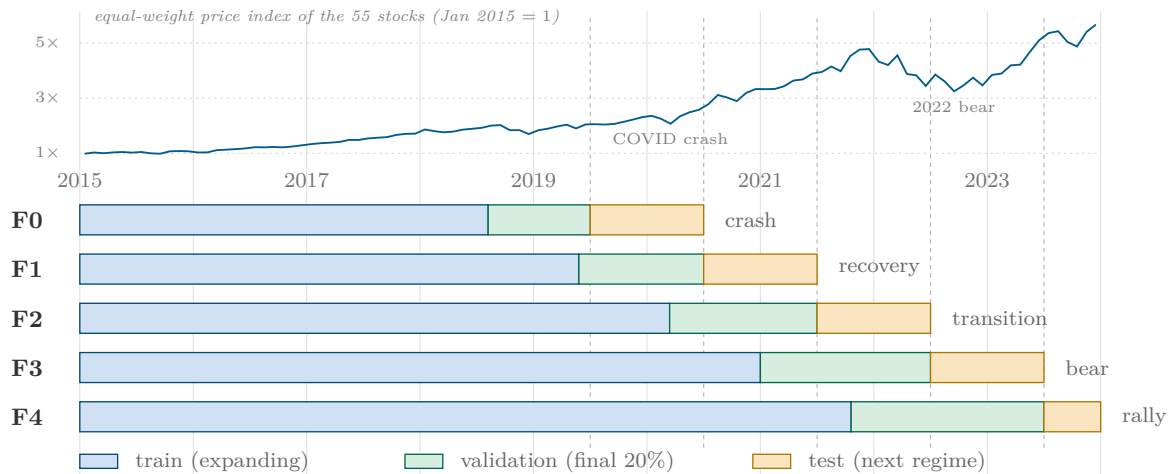


Figure 1: The five expanding-window folds on the 2015–2023 calendar, drawn to scale. Top: the equal-weighted price index of the 55-stock universe (January 2015 = 1), the market path the folds must predict. Bottom: each training span (blue) extends to its cut-off and strictly contains the previous fold’s, with its final 20% (green) held out for validation and early stopping; each test span (orange) is the disjoint window that immediately follows — one year each, six months for F4 — and the dashed verticals mark the cut-off dates. Window sizes in stock-days, and each window’s market regime, are in Table 2.

3.2. Data sources and why each is tested

Each modality is included because the literature reports it as predictive; the ablation tests whether that holds under a common baseline and protocol. All non-price modalities are aligned to be observable at prediction time (no look-ahead).

Price and technical indicators (baseline). Daily open-high-low-close-volume bars from a public provider, augmented with ten standard technical indicators: log returns at one, five, and twenty days; the relative strength index; the moving-average convergence-divergence; an exponential-moving-average ratio; Bollinger %B (where the latest price sits within its rolling volatility bands); the average true range; realised volatility; and a volume ratio. Each indicator is standardised by a 252-trading-day (one-year) rolling z -score, so the feature seen on any day uses only information available by that day; no population statistic leaks backward in time. This price-and-technical block is the single-source baseline (configuration P) against which every other configuration is measured, and it is deliberately a *strong* baseline: technical indicators already encode momentum, mean-reversion, and volatility-regime information, so a multi-source configuration must beat price that is already enriched with the standard quantitative-finance feature set, not raw price.

Financial news sentiment. Published company news from the FNSPID corpus [3] (244,764 articles tagged to the universe), encoded along two parallel paths so the news pathway is not hostage to a single encoder. One path runs each article through a frozen finance-tuned FinBERT [1], takes the three-class sentiment probabilities, and aggregates them per stock-day into an 11-statistic feature block (mean and dispersion of sentiment, extremes, article count, three-day rolling averages, and a has-news flag). The other path encodes each article with a frozen Qwen3-0.6B embedding model and reduces the mean-pooled 1,024-dimensional vector to 128 dimensions by a per-fold PCA. Both apply a one-business-day lag, so $T-1$ news predicts T direction. News is the most-claimed multi-source signal in the field [11, 17, 29]; it is also where the replication concern is sharpest, because published articles are backward-looking by construction [12, 14], a timing gap that §4.4 shows is what drives the news result.

Social-media sentiment. StockTwits² messages (14.6M, all 55 stocks covered), aggregated into 18 daily features (message and author counts, user-self-labelled bullish/bearish ratios, net sentiment, and their three-day momentum), lagged one business day for the same reason as news. Coverage ends in December 2022; for the 2023 test window the social features are zero-filled with an explicit no-data flag, so the model sees an absence-of-data signal rather than imputed sentiment. Social sentiment is reported to improve movement prediction [25], motivating its inclusion.

Macroeconomic indicators. Six FRED³ series (the federal funds rate, the 10-year Treasury yield, CPI, unemployment, GDP, and the VIX, chosen to span the channels macro-aware prediction work conditions on: the policy rate, the term structure, inflation, the labour market, growth, and implied volatility), plus a binary flag for Federal Open Market Committee (FOMC) meeting windows. Timing here needs care, because macroeconomic series become public on two different schedules. The 10-year yield and the VIX are daily market quotes, observable at the close of day T and used as such to predict the overnight $T \rightarrow T+1$ return. The other four are *releases*: the value for a reference month or quarter only becomes public weeks later, so feeding it to the model on its reference date would leak the future. Each release series therefore enters the feature set shifted to its publication schedule — CPI by 30 business days, unemployment by 27, the monthly federal-funds average by 23, and GDP by 88, each at least the historical delay from reference date to release — and the rolling normalisation is computed on the shifted series, so the model only ever sees macro values that were public by the close of day T . Because these features are shared by all 55 stocks on a given day, macro can only express market-level timing, not stock-specific signal. Macro indicators are reported to help, especially at event windows [28]; the ablation tests whether daily macro context helps stock-specific direction.

Inter-stock relationship graph. A graph over the 55 stocks combining static GICS-sector co-membership (110 edges) with a dynamic graph of 20-day rolling return correlations thresholded at $|\rho| > 0.3$, encoded with a two-layer graph attention network [21] into 64-dimensional node features (the 0.3 threshold was chosen from a sweep over $\{0.1, 0.2, 0.3, 0.4\}$). The classifier family trains the GAT end-to-end and integrates it per sample; the forecaster family precomputes node embeddings once per fold with the GAT frozen at its random initialisation (a choice made for compute and leakage-safety, discussed in §5.1), the cut-off enforced,

²<https://stocktwits.com/>; the bullish/bearish tags are self-labelled by the posting users.

³Federal Reserve Economic Data, maintained by the Federal Reserve Bank of St. Louis: <https://fred.stlouisfed.org/>.

and a dedicated leakage test. Dynamic relational structure is reported to help [23]; the ablation tests it under the common protocol.

Table 3 summarises the sources, and Table 4 defines the nine ablation configurations.

Table 3: The five data sources and how each enters the model. The lag column records the alignment to prediction time (the per-source detail and the evidence motivating each are in §3.2); nothing the model sees on day T was unavailable by then.

Source	Provider / coverage	Lag
Price + technical	public OHLCV, 10 indicators, 2015–2023	—
News sentiment	FNSPID, 244,764 articles; FinBERT-11 / Qwen3-PCA128	$T-1$
Social sentiment	StockTwits, 14.6M msgs; 18 features	$T-1$
Macroeconomic	FRED, 6 series + FOMC flag	pub. date
Inter-stock graph	GICS + rolling correlation; GAT-64	$T-1$

Table 4: The nine ablation configurations, grouped by the role each plays in the design. Every configuration includes the price+technical block P; the letters N, M, S, G denote the news, macro, social, and graph sources added to it. The graph source is tested in isolation and over macro; it is not combined with the text and social sources. (The identifiers under which these appear in the result files are listed, in order, in Appendix A.4.)

Configuration	Added to price
<i>Baseline</i>	
P	— (price + technical indicators only)
<i>Single source over price — the RQ2 add-one-in tests</i>	
P+N	news
P+M	macro
P+S	social
P+G	graph
<i>Each source anchored over macro</i>	
P+N+M	news + macro
P+M+S	macro + social
P+M+G	macro + graph
<i>Joint ceiling</i>	
P+N+M+S	news + macro + social

The nine configurations are a purposive subset of the sixteen possible source combinations (Table 4), chosen so each research question is answered by a specific contrast rather than by reading a full 2^4 grid: the single-source group gives the add-one-in test for RQ2; the macro-anchored group reads each information source a second time over a market-context baseline, which is what makes the news-over-macro degradation contrast (P+N+M versus P+M) and its social and graph counterparts possible; and P+N+M+S is the joint ceiling that supports the leave-one-out reading. The graph source, whose encoder is a network rather than a feature block, is tested in isolation and over macro but not combined with the text and social sources. Keeping the grid to nine rather than sixteen lets every cell run at full seed and fold resolution rather than spreading the same compute thinly across all combinations.

Figure 2 traces the end-to-end pipeline: each raw source passes through its own encoder, and the five 64-dimensional outputs are concatenated into the per-sample input tensor that both architectural families consume. Every fold-dependent transform (the per-modality scalars, the per-fold news PCA, the graph precompute) is fit on the training slice only, so the diagram is also the leakage boundary.

3.3. Two architectural families

A natural objection to any null is that the architecture was too weak to find the signal. The study pre-empts this by testing two structurally distinct families under the same grid, folds, seeds, and statistics. The two are not the whole design space (a graph-native backbone, for instance, is out of scope), but they bracket a tenfold capacity range and two different ways of consuming the price sequence, which is the axis the objection turns on.

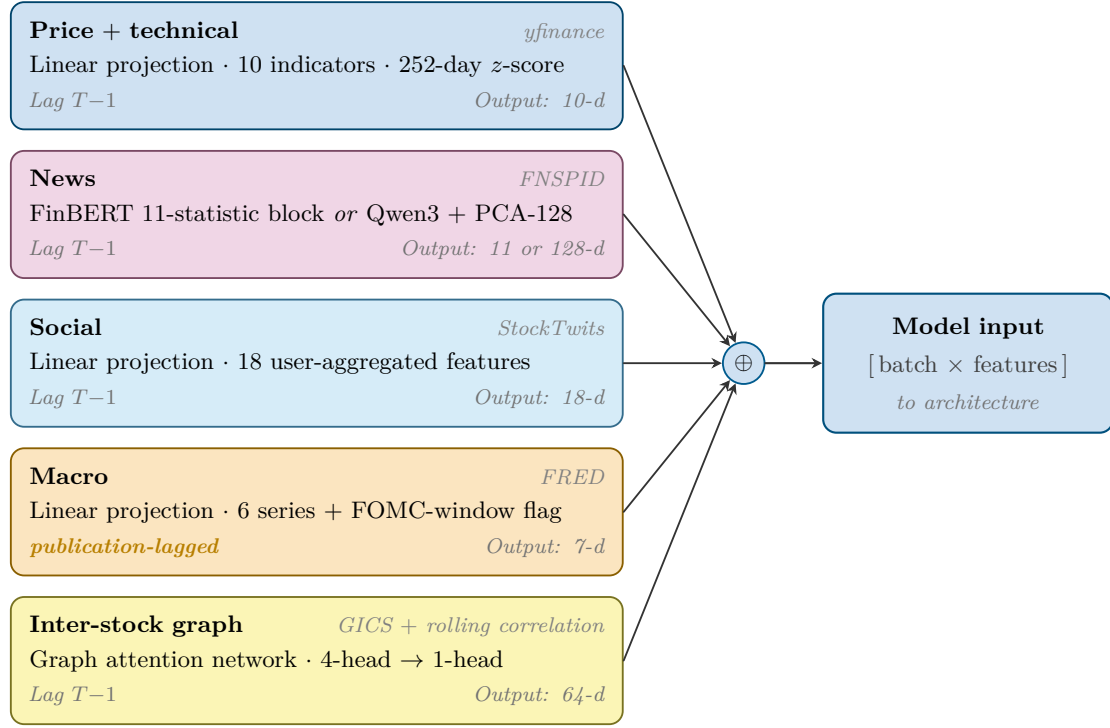


Figure 2: End-to-end data flow. Five raw sources feed per-source encoders; the 64-dimensional outputs are concatenated ($\oplus$) into the per-sample input tensor consumed by both architectural families. Source lags are listed in Table 3 (§3.2).

The *classifier-shaped* family is the dominant published shape: light per-modality encoders, each projecting its source to a 64-dimensional vector through a linear-ReLU-LayerNorm block, a fusion module, and a thin two-layer head. Its two variants differ only in how price enters: a one-day feedforward vector or a 20-day LSTM sequence. Across the nine configurations and three fusion strategies the parameter count ranges from 9,251 (price-only feedforward) to 88,035 (the largest multi-modal configuration), and the single heaviest block is the fusion projection rather than any encoder. This matches the structure of MSGCA [29] and the designs in [11, 17]: most capacity sits in front of a shallow predictor, with no dedicated forecasting body. Three published fusion strategies are compared within this family — simple concatenation, gated cross-attention, and multi-head self-attention. Figure 3 shows the wiring.

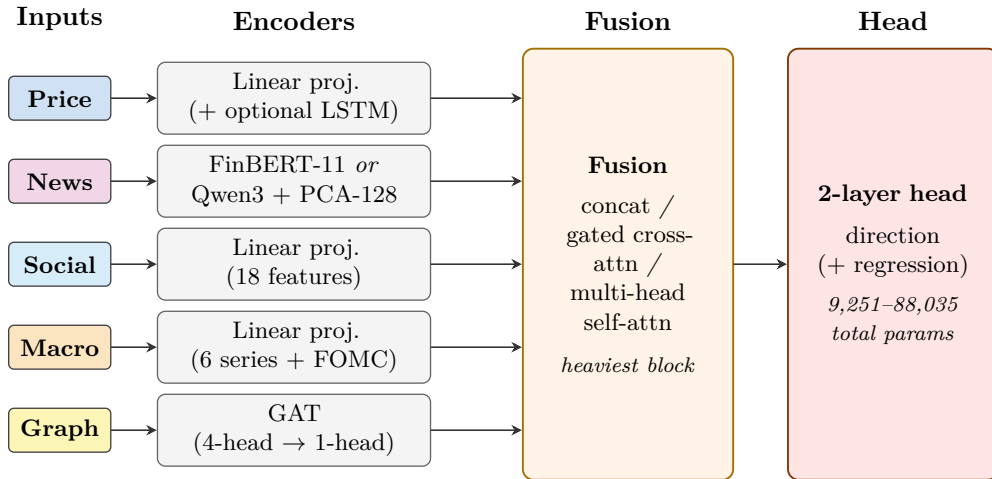


Figure 3: Classifier-shaped family. Five light per-modality encoders project to 64 dimensions; a fusion module (one of three strategies) integrates them; a two-layer head emits the prediction. There is no dedicated forecasting body — the single heaviest block is the fusion projection, and the total is 9,251–88,035 parameters depending on configuration. This is the dominant published shape.

The three fusion strategies differ only in how the five 64-dimensional encoder outputs $\{h_1, \dots, h_5\}$ are combined before the head, and they span the range of designs the literature uses. *Concatenation* simply stacks them, $h = [h_1; \dots; h_5]$, and lets the head learn any interaction; it is a strong baseline because it imposes no structure.

Multi-head self-attention treats the five encoder outputs as a length-five sequence and lets each attend to the others before pooling. *Gated cross-attention*, the mechanism MSGCA [29] reports as its advantage, designates price as a primary stream that queries each secondary modality h_m and admits it through a learned gate,

$$h = h_{\text{price}} + \sum_m g_m \odot \text{Attn}(h_{\text{price}}, h_m), \quad g_m = \sigma(W_g [h_{\text{price}}; h_m]), \quad (1)$$

so the gate $g_m \in (0, 1)$ can suppress an uninformative modality entirely. The premise behind Equation (1) is that a learned gate extracts signal that naive concatenation dilutes; whether it does so is tested directly in Section 4.3, where the three strategies prove statistically indistinguishable.

The *forecaster-shaped* family inverts that shape. It places most of its capacity in a Temporal Fusion Transformer body [13] (variable-selection networks, static covariate encoders, an LSTM encoder–decoder, and interpretable multi-head attention), fed by the same light per-modality projections. The output is five conditional quantiles of the next-day log return, the same variable whose sign defines the direction target; both direction (the sign of the median quantile) and a volatility estimate (the quantile spread) are derived from it without separate heads. Parameter counts run from about 195,000 (price-only) to about 278,000 (the full P+N+M+S configuration), with the variable-selection network as the single heaviest block and 95–99% of parameters in the forecasting body rather than the projections. Figure 4 shows the body’s components and where the capacity sits: the variable-selection network dominates at roughly 110,000 parameters, while the quantile head that emits the prediction is a single linear layer of 325. The cross-family comparison then asks the question the objection raises: does a tenfold increase in forecasting capacity, placed where a forecaster would place it, recover a stock-specific multi-source signal that the lighter family misses? Figure 5 shows the two families’ parameter budgets side by side.

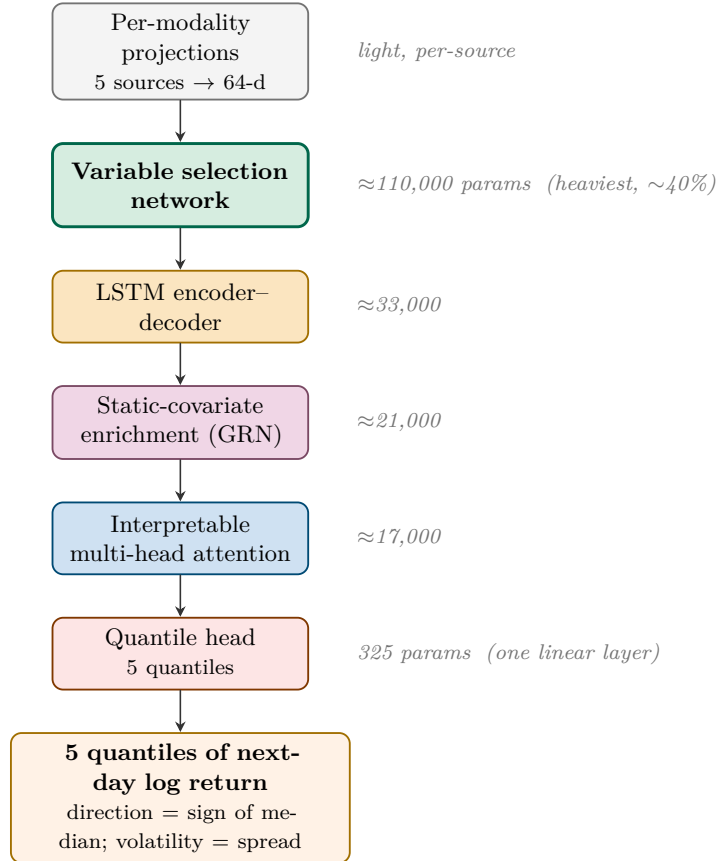


Figure 4: Forecaster-shaped family (Temporal Fusion Transformer body). The five projected sources pass through a variable-selection network, an LSTM encoder–decoder, static-covariate enrichment, and interpretable attention, ending in a single head that emits five conditional quantiles of the next-day log return. Direction (sign of the median quantile) and a volatility estimate (quantile spread) are both *derived* from that output — there are no separate task heads. Per-component parameter counts (the P+N+M+S configuration) show the capacity sits in the body: the variable-selection network alone is $\approx 110,000$ of the $\approx 278,000$ total, while the quantile head is a single 325-parameter linear layer.

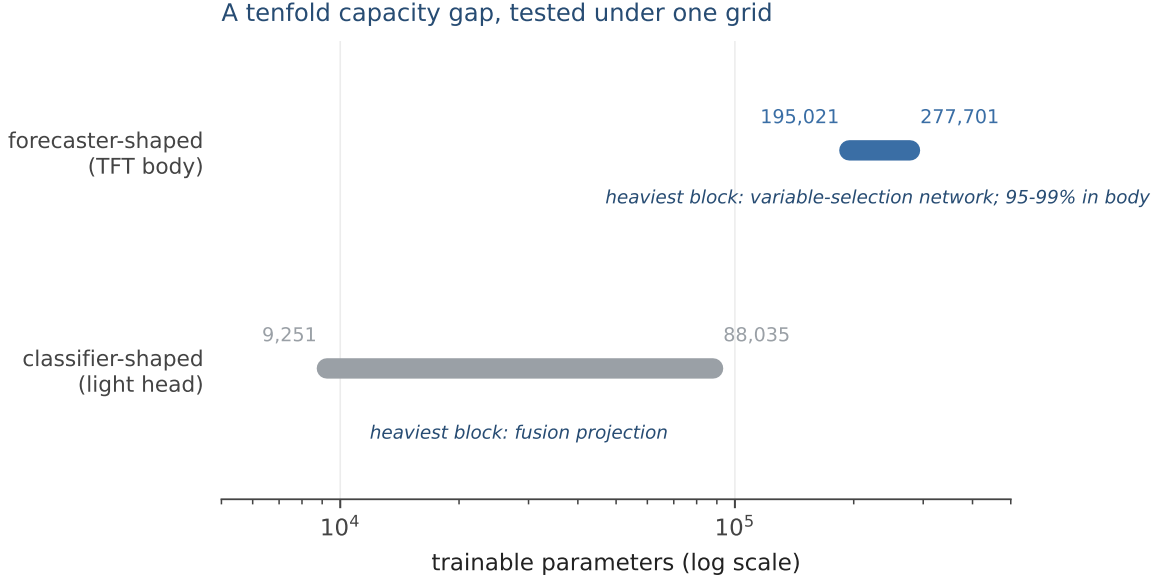


Figure 5: Parameter budgets of the two architectural families (log scale): the classifier-shaped family’s light head against the forecaster-shaped TFT, about an order of magnitude apart, with the heaviest block of each marked (counts in the text). The cross-family null over this capacity range is reported in Section 4.5.

3.4. Statistical protocol

Each configuration is compared to the price-only baseline with the paired procedure introduced in §2.1, matched on the (fold, seed) cell so that the large fold-to-fold variance — which would otherwise swamp a per-configuration delta of a few thousandths of an MCC point — cancels within each pair. For every contrast we report the three quantities of §2.1 jointly: the paired effect size Cohen’s $d = \bar{\Delta}/s_{\Delta}$ (mean paired difference over its standard deviation); the Bonferroni-corrected p -value [4] $p_{\text{bonf}} = \min(1, k \cdot p_{\text{raw}})$, with family size $k = 8$ for the per-configuration modality grid; and the sample size n . A finding counts as real only when all three agree, so that neither a significant-but-negligible effect at large n nor a large-but-noisy effect at small n is mistaken for a result. The pairing unit varies with the contrast, and we report n with every test so it is explicit: the primary per-configuration tests average the three fusion strategies within each (fold, seed, architecture) cell and pair on those $n = 30$ cells, while a contrast that holds the modality structure fixed and treats fusion as a replicate, such as the news leave-one-out of §4.4, pairs on all $n = 90$ (fold, seed, fusion, architecture) cells. Every contrast that involves the macro source is computed within a single matched batch, candidate and baseline run under the identical publication-lag protocol of §3.2, so no comparison mixes macro timing conventions.

The promotion bar was fixed *before* the forecaster-shaped family was run — pre-specified in the project repository, though not lodged on an external registry: a candidate is promoted only if it shows a positive paired delta, $p_{\text{bonf}} < 0.05$ at $k = 8$, $|d| \geq 0.3$, consistency across folds, and survival of an extension-seed retest. No configuration crossed it. The grid totals 2,468 valid experiments (2,268 classifier-shaped, 200 forecaster-shaped; Appendix A.4 reconciles the count, including the macro re-runs) across the nine configurations, five folds, three seeds, and, in the classifier family, three fusion strategies.

A null is only meaningful if the design could have caught a real effect. A power analysis says what it could catch: at the primary resolution of $n = 30$ paired observations, with the Bonferroni-corrected threshold ($\alpha = 0.05/8$), the design has an 80% chance of detecting an effect of Cohen’s $d \geq 0.70$ (and $d \geq 0.56$ at the larger $n = 45$ resolution). Anything smaller could be present on a single contrast and slip through — which is exactly why the same null is then re-checked across four independent robustness axes (§4.5): a small true effect might hide from one underpowered test, but not from all of them while never once clearing correction.

Training protocol. The hyperparameters were fixed once, during a calibration phase that preceded the main grid, and then held constant across every configuration so that no configuration is advantaged by per-model tuning — itself a form of test-set leakage when tuning is informed by the evaluation. All models use the Adam optimiser, batch size 64, hidden dimension 64, and gradient clipping at norm 1.0, with early stopping on the validation MCC at patience 30; the learning rate is 10^{-4} in the classifier-shaped family and 3×10^{-4} for the Temporal Fusion Transformer (its tuned-once value), held fixed across all configurations within each family. The validation set on which early stopping is judged is the chronologically last 20% (by calendar days) of each fold’s training window, carved per fold and disjoint from the test window (sizes in Table 2); the test set is never

read during training or selection. This is the discipline whose absence the MSGCA re-run of §4.2 quantifies, so the study applies it to itself by construction. The price look-back is the one input contract that differs by architecture, matched to how each consumes the sequence: one day for the feedforward variant, twenty for the LSTM, sixty for the Temporal Fusion Transformer. Training is short and stable: models converge in 54–72 epochs on average depending on configuration, with early-stopping triggered in 93–100% of runs (Appendix A.5), and the mean validation-to-test MCC gap is small (0.018–0.029), indicating the early-stopping criterion is not itself overfitting the validation split. The full direction grid (750 runs) took roughly 217 hours of compute on a single Apple M1 Max laptop (Appendix A.10).

Reproducibility. The apparatus is built to be re-checkable end to end, because a negative result is only as credible as its implementation. The two architectural families are implemented in two independent codebases over a shared data pipeline; both pass a fixed-seed integration test that produces CPU bit-identical result records across consecutive runs, so a reported number is a property of the configuration and not of run-to-run noise. Every fold fits its scalars, its per-fold news PCA, and its graph snapshots on the training slice only, each guarded by an explicit leakage assertion — the discipline the MSGCA diagnostic (§4.2) shows missing in a published system. All reported confidence intervals are bootstrap intervals over 10,000 resamples under a fixed RNG seed. The model code, configurations, and the scripts that regenerate every table and figure in this article are available from the authors on request.

4. Experiments and results

This section follows the research questions. It opens with the headline result, whether any source beats the price-only baseline (§4.1, answering RQ1 and RQ2), which then prompts its own question: if no source helps here, why does the literature so consistently report that one does? Three subsections answer it: the fusion strategies are equivalent (§4.3), the two effects detectable inside the null are bounded (§4.4), and two prominent published positives, audited from their released code, are shown to rest on where their result comes from rather than on the data they add (§4.2). The section then establishes that the null is not fragile across a second prediction target, a change of encoder, of fusion, and a tenfold capacity increase (§4.5), and draws the threads together (§4.6). Throughout, each contrast is the paired comparison of §2.1, reported with effect size, corrected significance, and sample size together.

4.1. No data source beats the price-only baseline

This section answers RQ1 and RQ2. Everything is measured against the baseline, which is itself only weakly predictive: price plus the standard technical indicators scores a pooled mean MCC of 0.0095, marginally above the chance level of zero ($t = 2.061$, $p = 0.048$; 63% of runs positive). A small edge is exactly what efficient-market reasoning predicts for liquid large-caps, and it is the bar a useful source must clear.

None does. Table 5 gives each of the eight configurations’ paired MCC delta against the baseline with effect size and corrected significance; Figure 6 plots the same deltas with 95% confidence intervals. Both show a flat line at zero. The best is +0.0005 MCC (social, P+S), the worst -0.0097 (news+macro, P+N+M); every delta sits within ± 0.010 of zero and none clears Bonferroni. The contrast closest to significance ($p_{\text{bonf}} = 0.14$) is that worst case, which makes prediction *worse*. Six of the eight deltas are negative and the other two rest on zero: noise tilted faintly against the added sources, not for them.

This holds in both model families. The forecaster-shaped Temporal Fusion Transformer, with roughly ten times the capacity, reproduces the null exactly: no configuration crossed the promotion bar pre-specified in §3.4, its extension-seed retest included. And the same verdict holds for the secondary volatility target, treated in full under robustness (§4.5): every configuration, the baseline included, earns a negative R^2 , and none beats the baseline after correction.

On the relative contribution of each source (RQ2), the picture is flat. Adding any single source moves MCC by at most $|d| = 0.25$, and the two largest single-source effects, macro ($d = -0.25$) and news ($d = -0.19$), are both *negative*; social and graph are indistinguishable from zero. Removing news or social from the full P+N+M+S model changes nothing either. No source carries positive marginal value; the only non-null marginal effect is backward-looking news, and it degrades, consistently across folds though not at strict five-fold significance (§4.4). Within the resolution of the design, ranking the sources by usefulness is ranking noise.

Table 5: Per-configuration paired statistics on next-day direction (MCC) against the price-only baseline P, shown as a forest plot in Figure 6. Classifier-shaped family; $n = 30$ paired observations per row; Bonferroni $k = 8$. Single-source configurations first, then the macro-anchored pairs, then the joint P+N+M+S model.

Config	Δ MCC	d	p_{bonf}
P+N	-0.0034	-0.191	1.000
P+M	-0.0063	-0.251	1.000
P+S	+0.0005	+0.030	1.000
P+G	+0.0000	+0.001	1.000
P+N+M	-0.0097	-0.458	0.143
P+M+S	-0.0052	-0.248	1.000
P+M+G	-0.0057	-0.260	1.000
P+N+M+S	-0.0048	-0.241	1.000

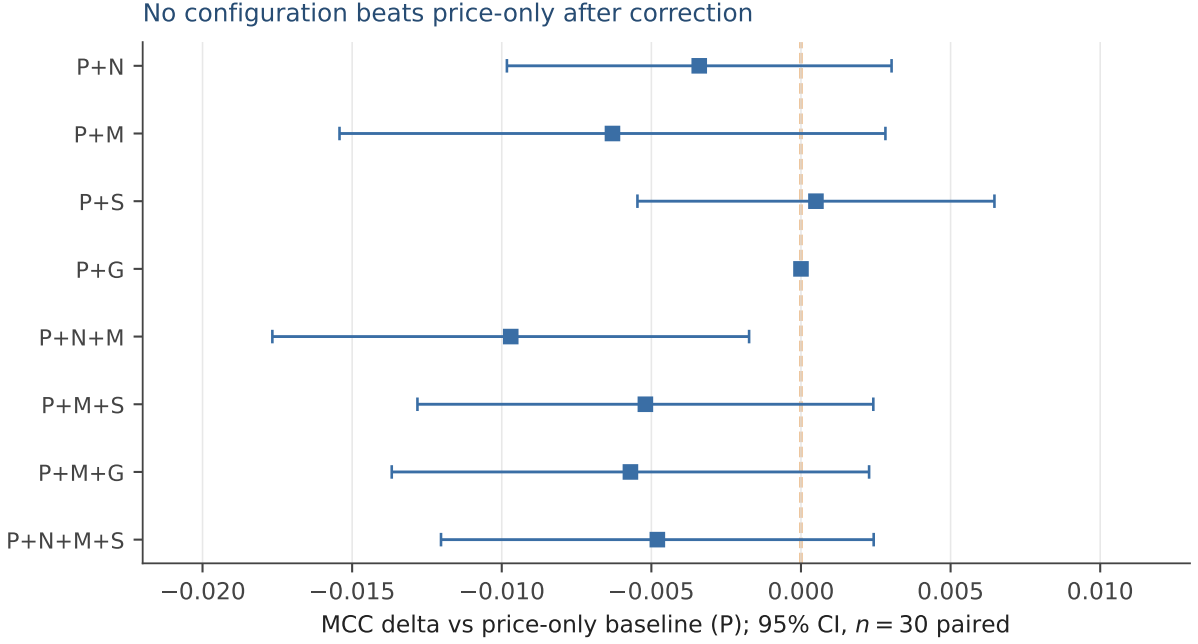


Figure 6: The numbers of Table 5 read as a forest plot: each point is a configuration’s paired MCC delta against the price-only baseline, the bars its 95% confidence interval. The dashed line at zero is the baseline (chance-level skill); an interval that straddles it denotes a configuration indistinguishable from price alone. Every interval straddles zero, and the cluster sits slightly to its left: no configuration clears Bonferroni correction, and the additions tilt faintly negative.

4.2. Two reported positives, audited at the level of their code

This subsection is the second contribution. The result so far is a null: on this universe, under this protocol, no source beats price. Much of the published literature reports the opposite. The two need not be in conflict — they test different universes, horizons, and baselines (§2.3). But because some of these systems release their code, the gap can be examined directly: one can ask not whether their result is plausible but whether, under a leakage-free evaluation, the reported positive still holds. We do this for two prominent systems, setting out for each what the released code does, what we ran, and what we changed, and reporting the outcome as a reproduction result rather than a verdict on the system in its own setting. The aim is to make concrete two distinct ways a reported gain can fail under a leakage-free protocol, which is why it cannot be read at face value.

MSGCA: the reported gain tracks test-set model selection. MSGCA [29] fuses price, news, and an inter-stock graph by gated cross-attention and reports a headline MCC of 0.1112 ± 0.0037 (five-seed mean) on the BigData22 benchmark. Its released code⁴ shows how the reported number is chosen. The training loop (`main.py`: the split at lines 50–51, the per-epoch test evaluation at line 134, the running maximum at

⁴<https://github.com/changzong/MSGCA>, branch `master`, commit 253925d (2024-12-01), retrieved 2026-02-09. Statements refer to this version; it contains the current training script and an earlier one under `old/`, both of which implement the rule described here.

lines 137–138, and the printed result at line 141) draws a random 80/20 train/test split with *no validation set*, scores the test set after every epoch, and keeps the maximum test MCC over all epochs as the reported figure. Model selection is thus performed on the test set. The split is moreover random rather than chronological — so training and test data span the same calendar period instead of holding out a future window, a look-ahead concern in its own right — though our diagnostic holds that split fixed and changes only the selection rule, so the inflation it measures is attributable to selection alone and is, if anything, a lower bound on the gap to a fully temporal protocol. The paper’s text instead describes a chronological train/validation/test split with no selection step, so this rule is visible only in the code; our statements concern the code that runs, not the protocol the paper describes.

What we ran, and what we changed. We ran the released code unmodified on its own BigData22 data to reproduce the headline, then re-ran it changing *only* the selection rule — carving a chronological validation set out of the training data, selecting the epoch by validation MCC, and reading the test set once at that epoch.⁵ Holding data, model, and training length fixed, the like-for-like measurement is within-run: choosing the reported epoch by the test set rather than by held-out validation raises the reported MCC by +0.04 to +0.07, and by similar amounts at the 200-, 300-, and 400-epoch budgets, so it is not an artefact of training length (Table 6). The absolute levels agree: best-on-test reproduces mean MCC ≈ 0.097 at the released 300–400-epoch budget, close to the paper’s 0.1112, while leakage-free selection on the same model and data gives chance (≈ 0 ; Figure 7). The reported positive does not survive leakage-free selection.

Two qualifications keep this precise. The bar-to-bar gap between the best-on-test reproduction (≈ 0.097) and the leakage-free result (≈ 0) is about 0.10 MCC, larger than the within-run inflation, but part of it is not selection: the leakage-free arm must carve its validation set out of training and so trains on about 20% less data. We therefore lead with the within-run figure, which holds data and length fixed; the cross-arm gap is the looser number. And this is one reproduction, not a survey: it shows that this particular headline tracks test-set model selection, gives a mechanism the related-work table (Table 1) suggests recurs wherever a protocol leaves model selection unspecified, and says nothing about whether other systems’ gains arise the same way. This study fixes the selection rule to leakage-free validation for every run (§3.4).

Table 6: MSGCA on BigData22 under the two model-selection rules, at the three training budgets, with the same data and model — the numbers behind Figure 7. *Best-on-test* is the released code’s rule (the maximum test MCC over all epochs); *validation-best* selects the epoch on held-out validation and reads the test set once. Best-on-test reproduces the paper’s reported level (0.1112) at the 300–400-epoch budget, while validation-best on the same model collapses to chance. The last column is the *within-run* inflation — the like-for-like gap between the two rules measured on identical validation-arm runs — which is the clean measure of the selection effect. It is smaller than the bar-to-bar difference between the first two columns, because that difference also reflects the validation arm’s roughly 20%-smaller training set.

Training budget	Best-on-test	Validation-best	Within-run inflation
200 epochs	0.030	−0.002	+0.040
300 epochs	0.097	−0.007	+0.056
400 epochs	0.097	−0.015	+0.069

The mechanism is easiest to see epoch by epoch (Figure 9). The test-set MCC wanders with training. Best-on-test reports its running maximum, the single luckiest epoch on the test set, whereas validation-based selection commits to whichever epoch looked best on held-out validation data and reports the test score there: the first is the peak of a noisy curve, the second an ordinary point on it. Best-on-test reads the test set twice, once to choose the epoch among many noisy candidates and once to report that epoch’s score, so the reported maximum is biased upward by the very test-set noise the selection picks up. Validation-based selection keeps the two roles separate (Figure 8 sets the two rules out side by side), which is why it alone gives a number that would survive on genuinely unseen data. Selecting on validation and reporting on test is correct because it keeps the test set a single-use estimate of generalisation; selecting on test turns the test set into part of the training loop.

CAMEF: a genuine headline that holds only under a non-temporal split. The second system is CAMEF [28], a KDD 2025 model forecasting five US assets around macroeconomic announcements; it too

⁵The re-run script `msgca_diagnostic_rerun.py`, the eight (validation-split, initialisation) seeds it iterates over, and the per-split, per-epoch test- and validation-MCC histories are retained in the project archive (available from the authors on request), so Table 6 and Figures 7–9 regenerate from it. The reproduction reuses the authors’ published embedding cache for their fixed seed-42 outer 80/20 split; extending the audit to a k -fold outer split would require regenerating those text embeddings through the original embedding API.

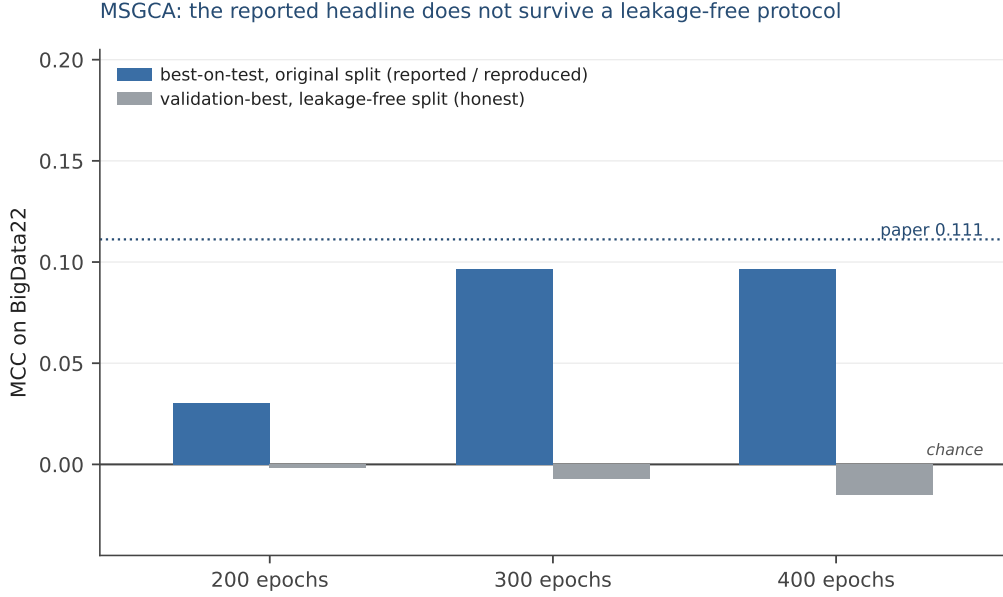


Figure 7: MSGCA on BigData22: the reported headline does not survive a leakage-free protocol (the numbers of Table 6). The released code on its original split (best-on-test selection, no validation set) reproduces the reported headline (blue; ≈ 0.097 at the 300–400-epoch budget, near the paper’s 0.111); under a leakage-free protocol — a proper train/validation/test split with the reported epoch chosen on held-out validation rather than on the test set — the same model and data collapse to chance (grey). The collapse corrects two coupled flaws of the released protocol at once: it has no validation set, and it selects the reported epoch on the test set. Holding the data fixed, the selection rule alone accounts for +0.04 to +0.07 MCC of the gap (the within-run inflation, Table 6).

The two model-selection rules, side by side. Same model, same data; only the rule for choosing the reported checkpoint differs.

Best-on-test (MSGCA’s released code, as published):

1. split data into **train** / **test** (no validation set);
2. every epoch: train, then record MCC on the **test** set;
3. report the **maximum** test MCC over all epochs.

Best-on-validation (this study’s re-run: same code, same data, only the selection rule changed):

1. split chronologically into **train** / **validation** / **test**;
2. every epoch: train, then record MCC on the **validation** set;
3. keep the best-validation epoch; report its **test** MCC once.

The test set is read $N+1$ times on the left (N times to choose the epoch, once to report it) and once on the right. That single difference moves the reported MCC by +0.04 to +0.07 within a run.

Figure 8: The selection rule that separates a reproducible result from an inflated one. Both columns run the identical MSGCA model on the identical BigData22 data; only the checkpoint-selection criterion differs. Reading the test set during selection (left) lets test-set noise inflate the reported maximum; reserving the test set for a single final measurement (right) does not.

releases its code.⁶ Its checkpoint selection is sound (the released loop keeps the model that minimises validation MSE, not a test-set maximum), so the concern is not selection but the split, and the audit runs in three steps. *The headline is real:* the authors’ released weights give MSE 0.000431 through their own `test.py`, matching the reported 0.000489, so unlike MSGCA the number is genuine. *But the documented recipe does not produce it:* a from-scratch retrain at the stated default of five epochs reaches only MSE 0.00249, about $5.8\times$ worse, so the released weights must come from a longer, undocumented training budget. *And the result depends on a non-temporal split:* the paper gives the split only as a 6:2:2 ratio, and whether it respects time is governed by a flag, `--split-by-time`, off by default and absent from every documented command; under the default, $\approx 99.7\%$ (913 of 916) of training events fall after the earliest test event, so the model trains on the test period’s future. Flipping the flag to chronological, all else held identical, makes the authors’ own recipe diverge — the training

⁶<https://github.com/lakebodhi/CAMEF> at commit `bd31b0c`, retrieved 2026-06-11. We ran the SP500 series at forecasting length 35 (the paper’s reported configuration); the only change to the authors’ files was the data-loader worker count (`num_workers` 15 $\rightarrow$ 2, an environment fit), with `train.py`, `test.py`, and the model byte-identical. Our harness and audit script are available from the authors on request.

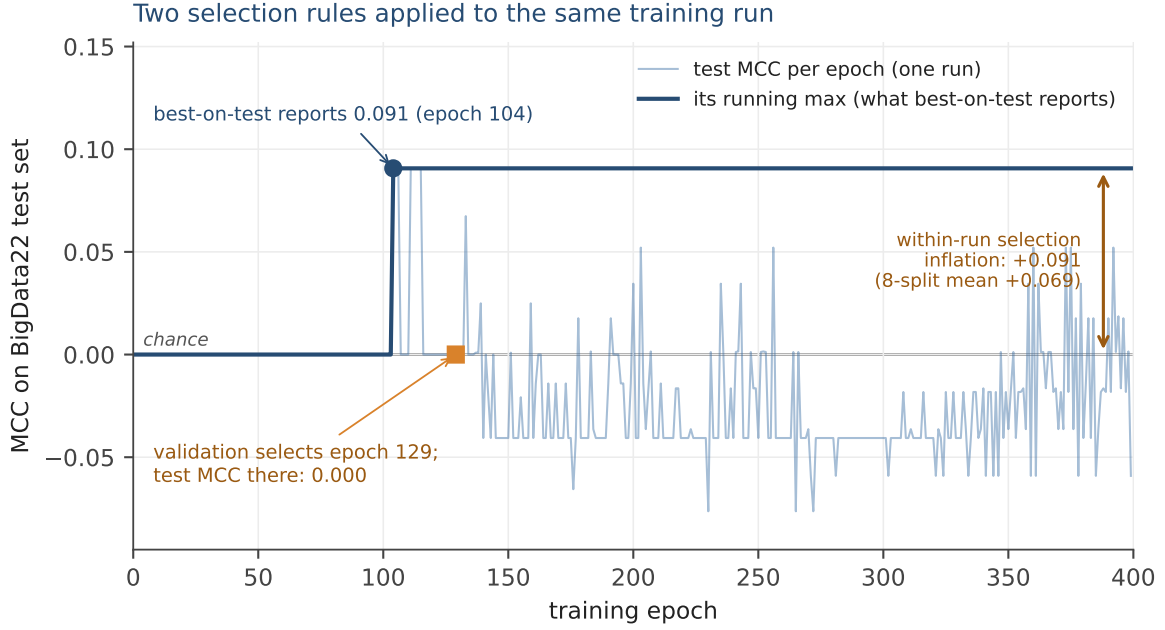


Figure 9: Both selection rules applied to one and the same training run (one of the eight validation-arm splits, whose history records test and validation MCC at every epoch). The light line is the per-epoch test MCC; the dark line its running maximum — what best-on-test selection reports. The square marks the epoch held-out validation selects, and the test MCC there, on the same curve. The vertical gap between the two reported numbers is therefore pure selection effect, training data and length held fixed: $+0.091$ on this run and $+0.069$ averaged over the eight splits at the 400-epoch budget shown (three of the eight never train above chance, so their gap is zero and the mean includes them), and $+0.04$ to $+0.07$ across budgets (§4.2).

loss rises and the test MSE blows up to about 15. We cannot retrain at the released model’s true (larger) budget, but a collapse this severe at the documented budget makes it implausible that more training would rescue a chronological split: the published headline rests on a non-temporal evaluation, of doubtful validity for a time-series forecast.

These are two papers, but the dimensions on which their evaluations differ from ours are not peculiar to them — they recur across the landmark systems. Table 7 sets the evaluation protocol of this study against the multi-source systems it is positioned against, on the axes that determine whether a reported number is an honest estimate of generalisation. The pattern is consistent: the published systems test on one temporal split rather than several regimes, most report a single best run rather than a seed-averaged mean with dispersion, none applies multiple-testing correction, and where the model-selection rule is documented it is not always insulated from the test set. None of this means the systems are wrong in their own terms; it means a lower absolute MCC reported under the protocol in the first column is not evidence of a worse model, and a higher one reported under the others is not evidence of a better one. This comparison is why the null here should be read as honest rather than weak.

4.3. Fusion strategies are equivalent

Within the classifier-shaped family, the three published fusion strategies (concatenation, gated cross-attention, and multi-head self-attention) are statistically indistinguishable. A Friedman test [6], the rank-based analogue of a repeated-measures ANOVA that asks whether any of three or more matched treatments consistently outranks the others, gives $\chi_F^2 = 1.300$ (χ_F^2 is its test statistic) and $p = 0.522$ across 240 matched subjects: the three strategies’ within-subject rankings are indistinguishable from random ordering. The per-fusion mean MCCs cluster within 0.0015 of one another. The result reproduces on the volatility target ($\chi_F^2 = 1.575$, $p = 0.455$). The architecture of the fusion module does not buy signal in this universe — which is what we should expect if MSGCA’s apparent fusion advantage (Section 4.2) was an evaluation effect, not an architectural one.

4.4. Two bounded findings inside the null

Two effects survive inside the overall null, neither of them a multi-source benefit over price.

Table 7: Evaluation protocol of this study against representative multi-source systems — the two audited directly, MSGCA and CAMEF, are named where they differ — on the axes that determine whether a reported metric estimates generalisation. The point is not that the other protocols are invalid but that absolute metrics across them are not comparable; this study fixes every axis to its conservative setting.

Dimension	This study	Compared systems (MSGCA, CAMEF, ...)
Train/test split	expanding temporal window, 5 chronological folds, no look-ahead	one fixed split; MSGCA a <i>random</i> stock-level split, CAMEF non-temporal by default
Validation set	yes — early stopping on validation MCC, test set seen once	varies; MSGCA none (scores the test set every epoch), CAMEF validation-MSE selection
Model selection	best-on-validation epoch; test set held out	MSGCA best-on-test; CAMEF best-on-validation (sound); often unspecified
Seeds	3 (mean $\pm$ dispersion reported)	1 single run, or a best-of reported
Temporal folds	5 (five distinct market regimes)	1 fixed period
Significance tests	paired t , Bonferroni at $k = 8$, Cohen’s d	none in the compared set
Reported metric	mean over seeds $\times$ folds	single best run, or best epoch averaged over seeds
Universe	55 sector-balanced large-caps, 8 years	1–5 stocks or an index, 1–2 years (varies)

Backward-looking news degrades prediction. The one source with a non-null marginal effect is news, and the effect is to degrade; we read it as a consistent directional effect rather than a high-precision significant one. The evidence is clearest at the fold level: the P+N+M-versus-P+M contrast is negative in four of the five folds, in both architectures, and largest in the two highest-news-flow regimes (the 2022 bear and the 2023 rally; Figure 10). Averaged to one value per fold the effect is large ($d = -0.80$, $n = 5$, with a fold-cluster bootstrap interval below zero, $[-0.020, -0.001]$), though with only five folds a strict t -test has little power ($p = 0.15$).⁷ The effect is also specifically *marginal over macro*: news added to price alone (P+N vs P) is merely null rather than harmful, and removing news from the full P+N+M+S model leaves performance unchanged (leave-one-out $d = -0.04$, ns). News hurts only where it is asked to add something macro has not already supplied, and there it adds nothing.

The mechanism is timing. Aggregated daily news sentiment correlates with *same-day* returns at $r = 0.143$ but with *next-day* returns at only $r = 0.005$: by the time a published article is in the dataset, the price has already moved on its content, so the feature the model receives is a lagged echo of a move that has happened, and adding it dilutes the price signal with noise. This is the forward-versus-backward distinction of Koval et al. [12] — forward-looking text carries signal, backward-looking news of past events does not — and it matches the decaying news-sentiment edge of Lopez-Lira and Tang [14], whose finance-tuned encoders post negative drift. The encoder is not at fault: across four encoders spanning a 5.5-fold size range (FinBERT, BGE, FinLang, Qwen3) the six pairwise contrasts are all indistinguishable (smallest $p = 0.26$, $n = 30$; §4.5), so a better reader of the same backward-looking text extracts the same absence of forward signal. The per-fold breakdown (Figure 10) concentrates the degradation in the heaviest-news windows, the 2022 bear (-0.014) and the 2023 rally (-0.029), and leaves it near zero in the calm folds.

Social sentiment helps an LSTM, but through the price model. The one positive interaction in the programme is between StockTwits sentiment and recurrence, and it does not qualify the headline null: neither social configuration beats the price-only baseline (Table 5). The contrast is between the two *architectures* given the same data, so it says something about the estimator, not about a new information source. On P+M+S the LSTM variant beats the feedforward variant by $\Delta = +0.0111$ MCC ($d = +0.439$, $p = 0.006$); on P+S the same comparison gives $\Delta = +0.0074$ ($d = +0.419$, $p = 0.008$), similar but weaker. P+M+S is the more defensible of the two: it survives removing the 2023 fold ($p = 0.028$), whereas P+S weakens to non-significance under the same restriction ($p = 0.222$), so we rest the finding on P+M+S.⁸ Taken alone this looks like social sentiment carrying signal that only a sequence model can exploit. A supplementary experiment rules that reading out: we

⁷The fold is the natural unit because the 90 (fold, seed, fusion, architecture) cells share only five test windows. The pooled cell-level figure usually quoted for such a contrast ($\Delta = -0.010$ MCC, $d = -0.355$, $p_{\text{bonf}} = 0.010$) treats those correlated cells as independent and overstates the certainty (§5.1).

⁸These are pre-specified secondary contrasts — a single planned feedforward-versus-recurrent comparison per configuration, outside the eight-way modality family — so reported at raw p rather than under the modality-grid Bonferroni correction. As with the news contrast they pool (seed, fusion) cells that share five folds; at the fold level the interaction is large and consistent ($d \approx +1.0$, positive in four of five folds, with a fold-cluster bootstrap interval above zero) but does not reach significance under a strict five-fold t -test ($p = 0.083$), a consistent but modestly powered pattern (§5.1).

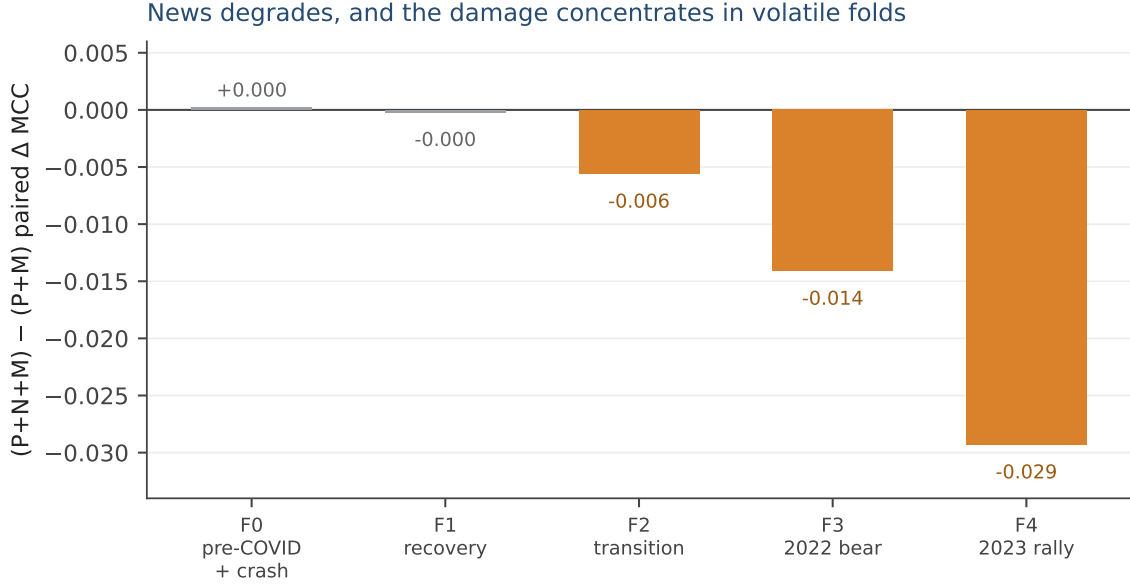


Figure 10: Per-fold news degradation: the paired MCC delta of $P+N+M$ minus $P+M$, $n = 18$ per fold (3 seeds $\times$ 3 fusions $\times$ 2 architectures). Negative in four of the five folds and largest in the 2022 bear and 2023 rally windows. The five fold means are the independent units of the effect: their mean is -0.010 ($d = -0.80$ at the fold level), with a fold-cluster bootstrap interval below zero, though five folds give a strict fold-level t -test little power ($p = 0.15$; §4.4). This is the direct paired contrast; it does not equal the difference of the corresponding rows of Table 9, which are each paired against the baseline and pooled differently.

gave the social stream its own 20-day temporal sequence, the same treatment price gets, so that any temporal structure *within* social sentiment (momentum in bullishness, bursts of posting) would be available to the model, and performance did not improve. The interaction is therefore not the LSTM finding structure in the social posts; it is the LSTM extracting more from the *price* sequence when same-day social context is present as a static side feature. The help comes from how the model processes price, not from a new information source. That is why we report it as a bounded finding rather than a multi-source gain, and why it does not survive the move to the forecaster-shaped family, whose single architectural class offers no feedforward-versus-recurrent contrast to reproduce it.

Macro and graph: the cleanest nulls. Macroeconomic indicators are null from every angle the grid offers: added to price alone and serving as the anchor for each other source, all four macro-containing deltas are slightly negative (Table 5). The design explains why. Macro features are shared by all 55 stocks on a given day, so they can express only market-level timing — and that market state is already carried by the price-and-technical block, whose volatility and trend indicators move with it. The inter-stock graph is the most precisely null of all, and informative because it is the modality with the most architectural latitude and still finds nothing: a two-layer graph attention network over a GICS-plus-correlation graph adds 0.0000 MCC over price, with the threshold and encoder choices behind it (§3.2) calibrated rather than left to chance. Relational structure between large-cap US equities is real, but it already lives in their correlated price histories; encoding it as an explicit graph re-represents information the price sequence carries rather than adding information it lacks.

4.5. Robustness of the null

A single null is weak evidence; a null that survives every reasonable way of trying to break it is strong evidence. The headline result was therefore re-tested along four independent axes, each chosen to target a specific “but you didn’t try ...” objection. None of the four reverses it.

A different prediction target. The same nine-configuration grid was re-run in the classifier-shaped family (the family whose multi-task head carries a regression output) with next-day volatility, the absolute return, as the target, on the reasoning that news arrival is a textbook predictor of volatility even when direction is unpredictable. Every configuration, including the price-only baseline, yields negative R^2 , and no configuration beats the baseline after correction: the nearest miss is the social configuration ($p_{\text{bonf}} = 0.69$, $|d| = 0.32$), and

the macro configurations, re-run leakage-free under the publication-lag protocol of §3.2, likewise fail to clear ($p_{\text{bonf}} = 0.47$, every macro R^2 negative). The per-fold breakdown is in Appendix A.8 (Table 15). Fusion equivalence reproduces on this target too ($\chi_F^2 = 1.575$, $p = 0.455$). The null is not specific to the direction target.

A different text encoder. Because news is the most-claimed channel, the news pathway was re-run with four encoders spanning a 5.5-fold size range: FinBERT, BGE-base, FinLang, and Qwen3-0.6B. All six pairwise contrasts are statistically indistinguishable (smallest $p = 0.26$, largest $|d| = 0.21$, on the 30 P+N cells; the price-only baseline is excluded as a structural zero (Appendix A.9); no encoder lifts news above price-only, and the four MCC distributions sit on top of one another and on the baseline (Figure 13). The bottleneck is the information content of backward-looking news, not the quality of the encoder reading it.

A different fusion strategy. Already shown equivalent in Section 4.3 (Friedman $\chi_F^2 = 1.300$, $p = 0.522$ on direction; $\chi_F^2 = 1.575$, $p = 0.455$ on volatility): however the five sources are combined before the head, the verdict is the same, and the architecture of the fusion module does not convert absent signal into present signal.

An order of magnitude more capacity. The strongest objection to a null from a light architecture is that a properly-sized forecaster would find what the light one missed. The forecaster-shaped Temporal Fusion Transformer carries roughly ten times the parameters, concentrated in a real forecasting body, and still produces the null: across the eighteen cross-family contrasts not one clears Bonferroni correction in either direction. The closest positive is $d = +0.593$ at $p_{\text{bonf}} = 0.339$, well short of the bar (Figure 11; full per-configuration contrasts in Table 14, Appendix A.7). The largest contrast of any sign is negative, the forecaster’s P+S falling below the classifier-shaped LSTM ($d = -0.916$, $p_{\text{bonf}} \approx 0.058$, borderline), which bounds the social-LSTM finding (§4.4) as a property of the classifier-shaped LSTM rather than a transferable gain. A model with an order of magnitude more forecasting capacity recovers no stock-specific multi-source signal in either direction, so the null is a property of the information, not of under-capacity.

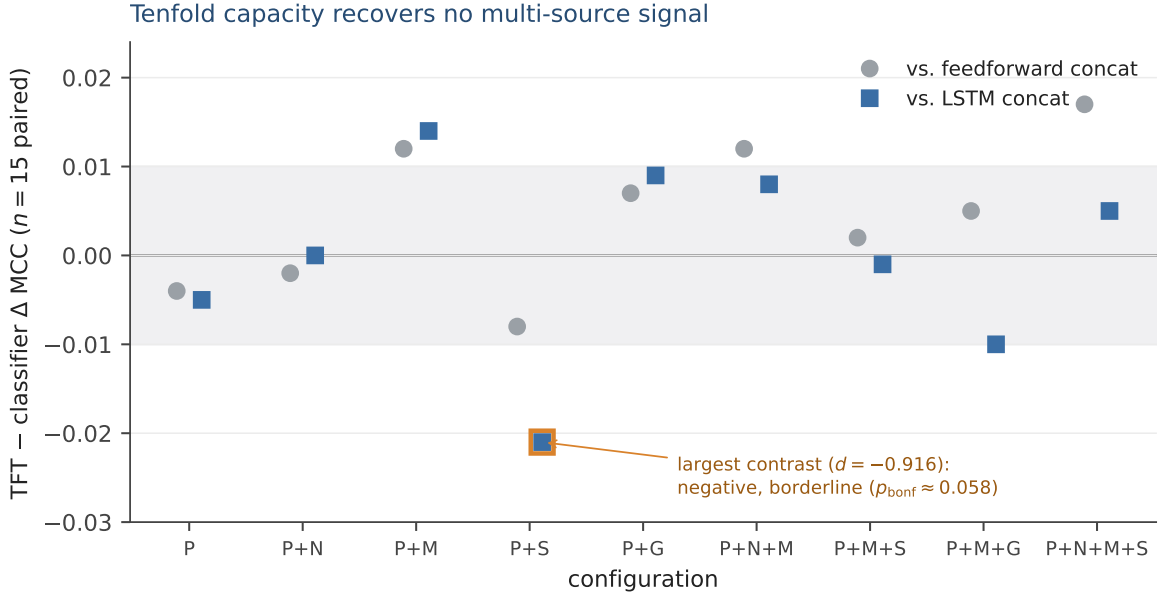


Figure 11: Forecaster-shaped TFT minus each classifier-shaped concatenation baseline, per configuration ($n = 15$ paired). No contrast clears Bonferroni correction in either direction across the eighteen-test grid; the largest is P+S against the LSTM reference, negative and borderline ($p_{\text{bonf}} \approx 0.058$), and every other contrast sits in the band indistinguishable from zero at this resolution. A tenfold parameter increase recovers no positive multi-source signal. Full statistics in Table 14, Appendix A.7.

4.6. Synthesis

Table 8 consolidates the findings across both architectural families and the robustness axes. Read as a whole it is the article’s central claim in one view: every row that asks “does this data source help?” returns null; the only non-null rows are the two bounded findings inside the null (news degrades, social helps an LSTM) and

the methodological audits (MSGCA’s headline is a selection artefact; CAMEF’s headline reproduces but holds only under a non-temporal split). The result holds whatever we vary: prediction target, text encoder, fusion strategy, or a tenfold change in model capacity.

Table 8: Findings synthesis across both architectural families and the robustness axes. “null” = no significant effect after Bonferroni at the relevant family size; “equivalent” = no within-family difference detectable; “bounded” = a real but conditional effect that is not a multi-source improvement over price. [†]The social-LSTM row reports raw p : it is a single pre-specified architecture-interaction contrast, outside the modality-grid Bonferroni family (§4.4). [‡]For the two within-null rows the d and p shown are the fold-level values ($n = 5$ independent test windows): both effects are large and consistent (4/5 folds; fold-cluster bootstrap interval excluding zero) but, with only five folds, not significant under a strict paired t -test. The cell-level statistics that treat (seed, fusion, architecture) as independent ($d = -0.35$, $p_{\text{bonf}} = 0.010$ for news; $d = +0.44$, $p = 0.006$ for social) overstate the certainty (§4.4, §5.1).

Question / finding	RQ	d	p_{bonf}	Verdict
Any multi-source config beats price (classifier)	RQ1	≤ 0.46	≥ 0.14	null
Any multi-source config beats price (forecaster/TFT)	RQ1	—	—	null
Marginal value of each source (add-one-in)	RQ2	≤ 0.25	—	null
Fusion strategy (concat / gated / MHSA)	RQ1	—	$p = 0.522$	equivalent
Volatility target instead of direction	RQ1	≤ 0.32	≥ 0.69	null (extends)
Text encoder choice (4 encoders)	RQ1	≤ 0.21	> 0.25	null
Capacity: TFT (10× params) vs classifier family	RQ1	+0.59	0.339	null (closest)
Backward-looking news added to price+macro	RQ3	−0.80	0.15 [‡]	bounded (degrades, 4/5 folds)
Social sentiment under an LSTM (P+M+S)	RQ3	+1.0	0.083 ^{‡†}	bounded (estimator, not source)
MSGCA: best-on-test vs validation-best selection	—	—	—	selection inflation +0.04 to +0.07
CAMEF: headline reproduces; non-temporal default split	—	—	—	diverges under a chronological split (§4.2)

The null is economic as well as statistical. In trading terms these values are immaterial. The best directional accuracy anywhere in the grid is about 52%, and the per-configuration mean Sharpe ratios sit near zero (overall mean 0.17); the price-only baseline is in fact the single highest-Sharpe configuration, with multi-source configurations no better (Table 13, Appendix A.6). None of this is net of transaction costs, and a round-trip cost of even a few basis points on daily rebalancing would erase the gross figures. The gap any additional source would need to close is therefore economically as well as statistically immaterial: a paired MCC delta of +0.003 (the $d = 0.2$ floor the design cannot resolve) is roughly 50.15% directional accuracy, far below any plausible cost threshold. The null is not “an effect there but too small to confirm”; on this universe and horizon there is no effect of a size that would matter even if confirmed.

5. Discussion

What the null means. For daily next-day direction prediction of liquid US large-caps, heterogeneous data sources do not add signal over price and technical indicators, and the result is stable across targets, encoders, fusion strategies, and a tenfold capacity range. This is the expected outcome under the replication prior of Section 2.3: a heavily-analysed, highly-liquid universe absorbs public information into price quickly, leaving little residual for downstream sources to exploit. The corroborating evidence runs in the same direction: the cross-domain null of Zhang et al. [27], the decaying news edge of Lopez-Lira and Tang [14], and the complexity-scepticism of Tan et al. [19] and Zeng et al. [26].

Reconciling with the positive literature. The gap between the field’s positive record and this null is not only a model-quality gap; part of it is an evaluation gap. Three mechanisms, all visible in Table 1, account for it. The first is model selection and its relatives: Section 4.2 shows one prominent positive resting on

test-set selection, an inflation this study’s protocol rules out by design, and a second audit there finds another such degree of freedom, a temporal-ordering flag left off by default. The same pattern, an evaluation-affecting choice the protocol leaves unspecified, is easy to introduce anywhere. The second is incomparable baselines: the landmark results are reported on different markets, horizons, and label definitions, against baselines that are rarely a well-tuned price-only model, so a reported “gain over baseline” need not be a gain over price. The third is the nature of the signal itself: the channel the field leans on hardest, news, is backward-looking and carries little forward information at the daily horizon [12, 14], which is why adding it degrades rather than helps. Held to one universe, one baseline, one protocol, and one set of corrected statistics, the advantage does not survive. These systems need not be wrong in their own settings; the point is that “multi-source beats price” is a claim about a comparison, and the comparison has rarely been run with the baseline, the protocol, and the statistics all held honest at once.

What it means in practice. The result is not that text, social, macro, and relational data are worthless, but that for next-day direction on liquid large-caps they carry no *marginal* signal beyond what price already reflects. For a practitioner the implication is to spend effort on the estimator and on honest out-of-sample evaluation before spending it on acquiring and fusing more data; for a reviewer it is to treat a multi-source “gain” as unverified until it is shown over a tuned price-only baseline under write-once-test evaluation.

5.1. Threats to validity

A null result is a claim about what the data does *not* contain, so its credibility rests on ruling out the mundane explanations for finding nothing. We organise these by the standard validity taxonomy [22].

Internal validity — could the apparatus have missed a real effect? The principal threat is statistical power: the design’s detectable floor is $|d| \geq 0.70$ at the primary $n = 30$ resolution (the power analysis of §3.4). Smaller effects (a d of 0.2, say, which on this metric is a paired MCC delta of roughly $+0.003$) could exist and go undetected on any single contrast. Two things bound this threat. First, such an effect is at or below the level of economic relevance (§4.6), so a positive too small to detect here is also too small to trade. Second, and more decisively, the same null recurs across the four independent robustness axes of §4.5 (target, encoder, fusion, capacity); an effect that hides from one underpowered contrast is plausible, but one that hides from all of them while never once clearing Bonferroni is not. The promotion checklist is deliberately one-sided, guarding against false positives rather than false negatives. That is the correct asymmetry for a study whose claim is a null, but it means small true effects remain formally possible.

A related point concerns the *unit* of analysis. The paired tests pool the (seed, fusion, architecture) cells within a fold, and those cells share a single test window, so they are not independent draws: the effective number of independent units is the five folds, not the 30 or 90 cells. For the headline nulls this only sharpens the conclusion: treating correlated cells as independent would, if anything, have manufactured significant differences, and none appeared, so the null is conservative under it. For the two within-null effects (news degradation, the social-LSTM interaction) it cuts the other way: the cell-level p -values overstate the certainty, which is why §4.4 reports those effects at the fold level ($n = 5$) and under a fold-cluster bootstrap alongside the cell-level figure, and characterises them as consistent but not strictly significant on five windows. A mixed-effects model with fold random effects would use the within-fold replication more efficiently than fold-level aggregation; we report the conservative fold-level analysis instead, and note that with only five folds either estimator is power-limited.

Construct validity — do the operationalisations measure what we intend? Three choices bound the claim. The target is per-stock daily direction (with volatility as a second target), not cross-sectional ranking, so the study speaks to “does this source predict a stock’s next move” and not to “does it help rank stocks against each other”, a related but distinct question. Encoders are frozen rather than fine-tuned end-to-end; the four-encoder screen shows the news result is not an artefact of a weak encoder, but it does not exclude that a jointly-trained encoder could extract something the frozen ones cannot, nor does it rule out the entity-specific biases Glasserman and Lin [7] document in pretrained sentiment models on heavily-covered stocks, a concern that a universe of the most-covered large-caps sharpens rather than relieves. And in the forecaster family the graph modality is encoded by the same two-layer GAT *frozen at its random initialisation*: a fixed, non-trainable structural feature map. That choice keeps the graph operator out of the per-sample training graph (for compute) and off the leakage surface; it is not an information-preserving reduction, so the forecaster-side graph result is a lower bound on what relational structure could contribute. The trained-GAT test exists in the classifier family, and is null (§4.4).

External validity — how far does the null generalise? The boundaries here are deliberate rather than accidental. The universe is 55 large-cap US equities over 2015–2023 at a daily horizon with backward-looking text — the most demanding setting for the claim (§3.1), so the null is a tight upper bound rather than a weak average. It does not speak to less efficient markets (small-caps, emerging markets), to intraday horizons, or to forward-looking text, each of which could behave differently and each of which is a stated future direction below rather than a claim made here. The one bias the design does not eliminate is survivorship, since the constituents are ranked at end of sample. As Section 3.1 argues, it acts on the price-only baseline and on every multi-source configuration alike and so cancels from the paired marginal contrasts that carry the result; it is unlikely to mask a true multi-source signal. A point-in-time universe is the clean future test.

Future work: information timing. The most informative extensions share one variable: *when* information reaches the model relative to when price has discovered it. Backward-looking news at the daily horizon, the case tested here, arrives after the price has moved; the same diagnostic that explains the news degradation ($r = 0.143$ same-day versus $r = 0.005$ next-day) predicts that two changes should behave differently. The first is forward-looking text (earnings-call transcripts, management guidance, analyst revisions), which carries information not yet absorbed when it reaches the model, and which Koval et al. [12] find helps where backward text does not. The second is the intraday horizon, which compresses the absorption window so that even same-day news retains a within-window component. Both are the natural next tests of the same hypothesis; a positive on either would carve out the conditions under which multi-source integration begins to pay, without contradicting the daily-horizon null established here. Figure 12 places the present null and the two extensions on the same map: the one axis that organises them is how much of the information has *not* yet been absorbed by price at the moment the model reads it.

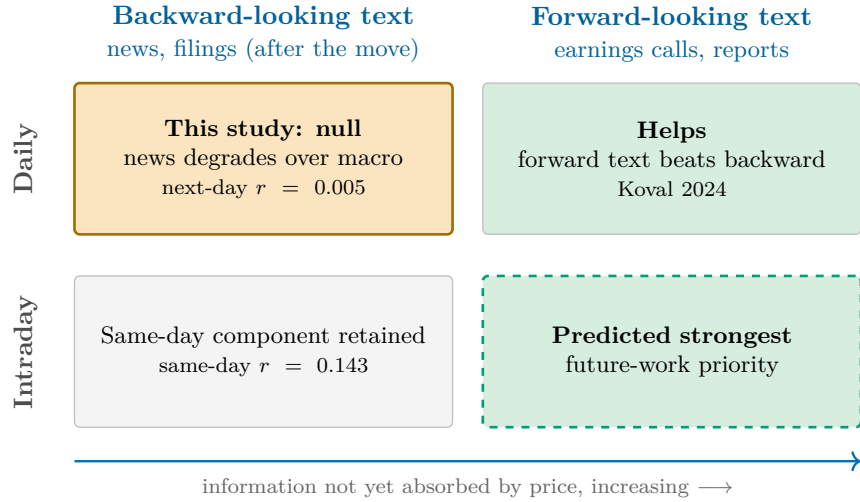


Figure 12: The information-timing map. The single axis that organises the news result and its two natural extensions is how much of the textual information has *not* yet been absorbed into price when the model reads it. The case this study tests — backward-looking news at the daily horizon (orange) — sits where almost all the information is already in the price, and the result is the null. Moving toward forward-looking text (which Koval et al. [12] find helps) or toward the intraday horizon (which compresses the absorption window, so the same-day $r = 0.143$ signal is not yet arbitrated away) both increase the unabsorbed component; the forward×intraday cell is the predicted strongest case and the priority for future work. The map is a hypothesis about *when* multi-source data should begin to pay, not a result.

6. Conclusions

This study tested whether integrating heterogeneous data sources improves daily stock-direction prediction over a price-only baseline, on 55 US large-caps under a single controlled ablation of 2,468 experiments. The answer to RQ1 is no: no combination of news, social, macroeconomic, or graph signals significantly beats price alone. The answer to RQ2 is that no source contributes positive marginal value, and that adding backward-looking news degrades direction prediction in four of the five folds. The answer to RQ3 is that the only conditions under which a data source changes direction prediction are adverse or estimator-bound: backward-looking news *degrades* it, most in the high-news-flow 2022 and 2023 folds, and social sentiment helps only a recurrent estimator, through how price is processed rather than through new information. No condition makes integration help. The factors that do move the reported number are methodological, not informational: the fusion architecture is irrelevant,

a tenfold capacity increase changes nothing, the one prominent published positive we reproduced reports a headline that does not survive leakage-free model selection in its released code, and a second whose headline does reproduce holds it only under a non-temporal split, diverging once the split is made chronological.

The two contributions follow. First, a robust null on multi-source financial prediction for liquid US equities, established across prediction targets, text encoders, fusion strategies, and architectural families. Second, a reproducibility audit of two reported positives that gives concrete, reproducible reasons — test-set model selection in one, a non-temporal data split in the other — for the gap between the field’s reported gains and this null. A single reordering of priorities connects them: on this problem, getting the estimator and its evaluation right matters more than adding data sources, and a study that does not separate the two effects can mistake an evaluation artefact for an information gain.

The most informative next step lies on the information-timing axis: forward-looking text (earnings calls, guidance, analyst revisions) and intraday horizons both change *when* information reaches the model relative to price discovery, and both are predicted to behave differently from the backward-looking, daily-horizon case tested here. Neither would overturn the present null; both would sharpen its boundary, locating the conditions under which heterogeneous data begins to pay. Until such a boundary is drawn, the disciplined default for next-day prediction on liquid equities is the one this study could not beat: price, evaluated honestly.

A. Per-fold breakdowns and experiment inventory

This appendix reports the per-fold detail behind the pooled results in Section 4, and reconciles the experiment count. The body reports pooled or overall statistics because the per-configuration claim is a claim about the average over regimes; the per-fold tables here let a reader confirm that no configuration’s null is hiding a strong single-regime effect cancelled by the others.

A.1. Multi-source delta by configuration and fold

Table 9 gives the per-fold MCC delta of each configuration against the price-only baseline P. The pattern that matters for the headline null is that the column of overall paired p -values is uniformly non-significant: where a configuration shows a starred single-fold cell (for example P+G in F1 and F4, or P+M in F0 and F4), the sign is inconsistent across folds and the effect washes out in the average. The one systematically negative column is F0 (pre-COVID and the COVID crash), where almost every multi-source configuration underperforms price alone — the addition of noisy alternative data is most harmful in the most turbulent training window.

Table 9: Per-fold MCC delta versus P by configuration (classifier-shaped family, pooled over architecture and fusion). Significance markers are uncorrected per-fold and exploratory: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The final columns are the pooled mean delta and the overall paired p -value, none of which is significant after Bonferroni at $k = 8$.

Config	F0	F1	F2	F3	F4	Mean	p
P+N	−0.0013	−0.0119	+0.0048	−0.0089	+0.0004	−0.0034	0.313
P+M	−0.0364***	−0.0047	−0.0023	−0.0069	+0.0188**	−0.0063	0.180
P+S	−0.0022	−0.0078	+0.0085	+0.0028	+0.0011	+0.0005	0.872
P+G	−0.0118	−0.0272**	+0.0200*	−0.0056	+0.0248*	+0.0000	0.996
P+N+M	−0.0288***	−0.0012	−0.0049	−0.0144**	+0.0004	−0.0097	0.018
P+M+S	−0.0277***	+0.0004	+0.0040	−0.0056	+0.0030	−0.0052	0.184
P+M+G	−0.0365***	+0.0040	+0.0071	−0.0163**	+0.0130*	−0.0057	0.165
P+N+M+S	−0.0202*	+0.0055	−0.0029	−0.0091*	+0.0027	−0.0048	0.197

A.2. Architecture by fold

Table 10 gives the per-fold LSTM-minus-feedforward comparison pooled over all configurations. The overall effect ($d = 0.276$, $p < 0.0001$) is small but reliable; the per-fold cells are reported for completeness only and are not interpreted as a regime prescription, since with five folds the per-regime estimates are single observations of their regime type and were not pre-specified as hypotheses. The defensible claim is the pooled one: a recurrent estimator extracts marginally more from the price sequence than a one-step feedforward one, on a metric where both remain close to chance.

Table 10: Per-fold LSTM–FF MCC delta, pooled over all nine configurations ($n = 75$ per fold; $n = 375$ pooled). The pooled effect is $d = 0.276$, $p < 0.0001$. Per-fold cells are exploratory.

Fold	Regime	FF MCC	LSTM MCC	Δ (d)
F0	pre-COVID + crash	+0.0012	+0.0128	+0.0116 (0.57)
F1	2020–2021 recovery	+0.0306	+0.0266	−0.0039 (−0.14)
F2	2021–2022 transition	−0.0072	−0.0060	+0.0013 (0.07)
F3	2022 bear	+0.0063	+0.0257	+0.0194 (0.83)
F4	2023 AI rally	−0.0096	−0.0023	+0.0073 (0.24)
Pooled		+0.0042	+0.0114	+0.0071 (0.28)

A.3. Baseline behaviour by fold

Table 11 reports the price-only baseline P per fold and architecture. It is included because the headline null is a comparison against this baseline, and the baseline is itself regime-dependent: P is solidly positive in the 2020–2021 recovery (F1) and weak or negative in the 2021–2022 transition (F2), so the per-configuration deltas in Table 9 are differences against a moving reference. The pooled P mean of 0.0095 is marginally above chance; the spread across folds (from -0.024 to $+0.046$) is an order of magnitude larger than any multi-source delta, which is the practical reason every comparison in the body is paired on the (fold, seed) cell rather than computed from raw means.

Table 11: Price-only baseline P mean MCC by fold and architecture ($n = 3$ seeds per cell). P runs only the concatenation path (single modality, no cross-modal fusion to perform). The fold-to-fold spread dwarfs the per-configuration multi-source deltas, motivating the paired design.

Fold	Regime	FF MCC	LSTM MCC
F0	pre-COVID + crash	+0.0263	+0.0302
F1	2020–2021 recovery	+0.0459	+0.0261
F2	2021–2022 transition	−0.0085	−0.0239
F3	2022 bear	−0.0093	+0.0260
F4	2023 AI rally	−0.0191	+0.0015
Pooled		+0.0071	+0.0120

A.4. Experiment inventory

The 2,468 valid experiments reconcile as follows. The classifier-shaped family contributes 2,268: the 750-run core direction ablation (eight multi-source configurations $\times$ three fusion strategies, plus the price-only baseline under one fusion, all $\times 5$ folds $\times 3$ seeds $\times 2$ architectures = 750), the matched 750-run volatility ablation, 240 encoder-screening runs (4 encoders on the news pathway), 360 supplementary runs (the price+news+social configuration, the Social-LSTM control, and the LayerNorm-corrected graph re-runs), and the remaining calibration and screening runs that fixed the dead-zone, the news-aggregation pipeline, and the training hyperparameters before the main grid. The forecaster-shaped family contributes 200: the 135-run nine-configuration TFT ablation, a baseline sanity check against the classifier family, and the extension-seed P+M retest. All counts are reproduced from the result CSVs (available from the authors on request), which identify configurations by archive identifiers: A1 (P+N+M+S), A2 (P+N+M), A3 (P+N), A4 (P+M), A5 (P+M+S), A6 (P+S), A7 (P), A8 (P+G), A9 (P+M+G).

On top of this grid, the macro-containing configurations and the price-only baseline were re-run under the publication-lag macro protocol of §3.2 (1,010 runs: on the direction target, the five macro configurations and the baseline across both classifier architectures and all three fusion strategies, plus the forecaster-shaped extension-seed runs, 530 in all; and a matched 480-run re-run on the volatility target); every macro-versus-baseline contrast reported in the body and the macro rows of the volatility table (Table 15) come from these matched batches. Contrasts in which the same macro features sit on both sides (the fusion comparison, the cross-family contrasts, the news-over-macro and architecture-interaction contrasts) are invariant to the macro timing convention, since any macro term enters both sides identically and cancels from the paired difference; those retain the original grid, as do the non-macro rows of the volatility table.

A.5. Training dynamics

Table 12 reports per-configuration training behaviour for the classifier-shaped family. The early-stop rate and the small, stable validation-to-test MCC gap that §3.4 cites hold across every configuration; news-containing configurations (P+N+M+S, P+N+M) train longest and show the widest gaps, consistent with the news pathway adding parameters that fit noise.

Table 12: Per-configuration training dynamics (classifier-shaped family, pooled over architecture, fusion, fold, and seed). Early-stop rate is the fraction of runs halted by the patience criterion before the epoch budget; the val-test gap is the mean MCC difference between the validation and test folds.

Config	Mean epochs	Early-stop rate	Val-test gap	Mean params
P	54.1	100%	0.0182	35,267
P+N	54.9	100%	0.0198	48,686
P+M	67.8	96%	0.0227	48,430
P+S	62.0	99%	0.0203	49,134
P+G	62.7	99%	0.0207	53,038
P+N+M	66.9	93%	0.0294	50,691
P+M+S	63.9	98%	0.0291	51,139
P+M+G	66.0	98%	0.0280	55,043
P+N+M+S	71.6	97%	0.0287	53,400

A.6. Financial metrics

Table 13 gives per-configuration directional accuracy and a naive long-short Sharpe ratio, the detail behind the economic-significance argument of §4.6. No configuration exceeds 52.3% directional accuracy; the per-configuration Sharpe ratios are near zero and unstable in sign (overall mean 0.17, range $[-1.38, 1.47]$), with the price-only baseline P among the strongest rows.

Table 13: Per-configuration financial metrics ($n = 45$ per architecture, except P with $n = 15$). Sharpe is a naive daily long-short ratio gross of costs; directional accuracy is the fraction of correctly-signed predictions. A statistically positive MCC does not imply profitable trading; these figures are reported for scale, not as a strategy.

Config	FF Sharpe	LSTM Sharpe	FF dir. acc.	LSTM dir. acc.
P	+0.427	+0.512	0.5225	0.5201
P+N	+0.333	+0.259	0.5178	0.5144
P+M	−0.007	+0.031	0.5053	0.5072
P+S	+0.377	+0.403	0.5195	0.5207
P+G	+0.303	+0.240	0.5143	0.5112
P+N+M	+0.038	−0.115	0.5101	0.5044
P+M+S	+0.105	+0.049	0.5088	0.5083
P+M+G	+0.117	+0.116	0.5092	0.5111
P+N+M+S	+0.006	+0.221	0.5057	0.5147

A.7. Cross-family architecture detail

Table 14 gives the full per-configuration contrasts behind the capacity-robustness claim of §4.5: the forecaster-shaped TFT minus the classifier-shaped model at each matching (configuration, fold, seed), paired at $n = 15$ against both the feedforward and LSTM concatenation references. We correct at the natural family size $k = 18$ (one hypothesis per contrast across two references and nine configurations); under it no contrast clears in either direction, as §4.5 reports. The bookkeeping does not drive that: under the more generous $k = 9$ the positive direction is unchanged, still with nothing clearing, and only the borderline negative P+S-versus-LSTM contrast would tip to $p_{\text{bonf}} = 0.029$; we report the conservative $k = 18$. One caveat applies to the graph rows: the forecaster encodes the inter-stock graph with the GAT frozen at its random initialisation (§5.1), so its P+G and P+M+G cross-family cells are corroborating rather than capacity-clean — but the fair graph test already exists and is null (the trained-GAT classifier nulls graph at an add-one-in $d = 0.001$, §4.1), so the graph conclusion does not rest on the frozen encoder.

Table 14: Forecaster-shaped TFT against the classifier-shaped feedforward and LSTM concatenation baselines, paired on (fold, seed), $n = 15$ per row, Bonferroni at $k = 18$. No contrast clears correction in either direction; the bold row (P+S vs LSTM) is the largest of any sign — negative and borderline.

Config	vs. feedforward concat			vs. LSTM concat		
	Δ MCC	d	p_{bonf}	Δ MCC	d	p_{bonf}
P	−0.004	−0.158	1.000	−0.005	−0.198	1.000
P+N	−0.002	−0.077	1.000	+0.000	+0.016	1.000
P+M	+0.012	+0.315	1.000	+0.014	+0.369	1.000
P+S	−0.008	−0.406	1.000	−0.021	−0.916	0.058
P+G	+0.007	+0.323	1.000	+0.009	+0.384	1.000
P+N+M	+0.012	+0.466	0.831	+0.008	+0.314	1.000
P+M+S	+0.002	+0.058	1.000	−0.001	−0.046	1.000
P+M+G	+0.005	+0.138	1.000	−0.010	−0.249	1.000
P+N+M+S	+0.017	+0.593	0.339	+0.005	+0.174	1.000

A.8. Volatility cross-target detail

Table 15 gives the per-fold paired delta in volatility R^2 (configuration minus P) behind the cross-target null of §4.5: the classifier-shaped volatility ablation, concatenation fusion only, pooled across two architectures and three seeds at $n = 6$ paired observations per cell. The five macro rows are the leakage-free publication-lag re-run (§3.2) paired against the re-run baseline; the three non-macro rows (P+N, P+S, P+G) are leak-invariant and retain the original ablation. The per-fold detail confirms the aggregate null is not masking regime structure: the deltas change sign across regimes, and the largest departures anywhere are the *negative* ones in the 2023 rally (P+N −0.0627, P+M −0.0591). Every configuration’s aggregate R^2 , P included, is negative, so each delta is a difference between models that both underperform the unconditional mean.

Table 15: Per-fold paired delta in volatility R^2 (configuration minus P; classifier-shaped family). Concatenation fusion only, pooled across two architectures and three seeds, $n = 6$ paired observations per cell. The five macro rows are the leakage-free publication-lag re-run; the three non-macro rows are leak-invariant and retain the original ablation. The aggregate R^2 is negative for every configuration; per-fold deltas are exploratory and uncorrected.

Config	F0	F1	F2	F3	F4
P+N	−0.0085	−0.0236	+0.0146	+0.0171	−0.0627
P+M	−0.0089	+0.0111	+0.0093	−0.0280	−0.0591
P+S	+0.0165	+0.0172	+0.0298	+0.0092	−0.0164
P+G	+0.0252	−0.0400	+0.0249	+0.0097	−0.0241
P+N+M	−0.0018	+0.0116	+0.0014	−0.0323	−0.0024
P+M+S	+0.0018	+0.0698	+0.0389	−0.0417	−0.0050
P+M+G	+0.0023	−0.0063	+0.0092	−0.0238	+0.0091
P+N+M+S	+0.0229	+0.0267	+0.0381	−0.0333	+0.0084

A.9. Encoder-screening detail

Table 16 gives the six pairwise contrasts behind the encoder-independence claim of §4.5. Four encoders spanning a 5.5-fold parameter range (FinBERT, BGE-base, FinLang, and Qwen3-0.6B) were run on the news pathway across 240 experiments. Each contrast is computed on the P+N cells only — 30 paired observations per pair (5 folds $\times$ 3 seeds $\times$ 2 architectures) — and deliberately excludes the price-only P cells: P carries no news pathway, so its outputs are identical across encoders, and every P-versus-P pair is a structural zero that would only shrink the apparent differences and inflate the sample size toward the very null being tested. On the P+N cells every difference sits within ± 0.0029 MCC and none is significant even uncorrected (smallest $p = 0.255$, largest $|d| = 0.212$, both on BGE-versus-FinLang). A 5.5-fold change in encoder capacity moves the news result by nothing, which locates the bottleneck at the information content of backward-looking news rather than the quality of the encoder reading it.

Table 16: Encoder pairwise paired-difference tests on the P+N news pathway, $n = 30$ paired observations per row (5 folds $\times$ 3 seeds $\times$ 2 architectures). The price-only P cells are excluded as structural zeros (see text). All six differences are within ± 0.0029 MCC; none is significant even uncorrected.

Pair (encoder A vs B)	Δ MCC (A–B)	$ d $	p_{paired}
FinBERT vs BGE	+0.0010	0.056	0.760
FinBERT vs FinLang	−0.0019	0.132	0.474
FinBERT vs Qwen3	−0.0006	0.037	0.842
BGE vs FinLang	−0.0029	0.212	0.255
BGE vs Qwen3	−0.0016	0.124	0.502
FinLang vs Qwen3	+0.0013	0.082	0.657

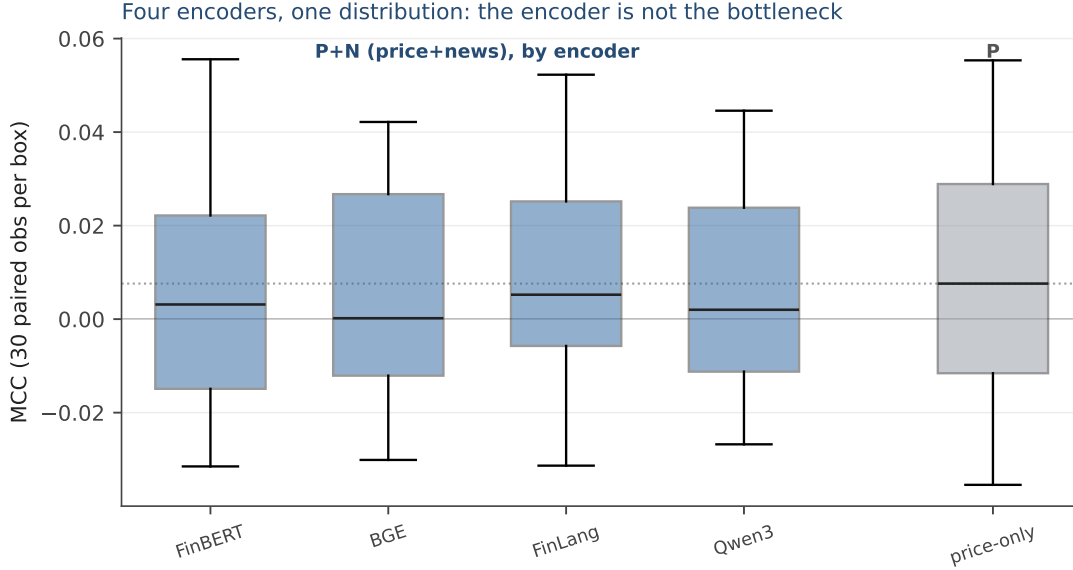


Figure 13: MCC distribution on P+N under each of the four text encoders, with the price-only P baseline shown once for reference (dotted median); each encoder box is 30 paired observations (5 folds $\times$ 3 seeds $\times$ 2 architectures). P is shown for context but excluded from the pairwise contrasts of Table 16 (a structural zero; see text). The four encoder distributions are mutually indistinguishable and overlap the baseline: news does not beat price under any encoder.

A.10. Computational cost

All experiments ran on a single Apple M1 Max (CPU, four runs in parallel). The figures below are wall-clock compute on that machine, reported for reproducibility and scale rather than as a hardware benchmark: the whole programme is re-runnable on one laptop-class machine, with no cluster required. The full direction grid (750 runs) took 216.8 hours: 101.1 for the feedforward family (mean 24,542 parameters per model) and 115.7 for the LSTM family (mean 76,574). News-containing configurations dominate the cost — P+N+M alone took 47.2 hours, against 3.7 for the price-only baseline — because the FNSPID encoding pathway is the heaviest per-sample operation in the pipeline. The forecaster-shaped programme adds to this; run counts reconcile in §A.4.

References

- [1] Dogu Tan Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019. URL <https://arxiv.org/abs/1908.10063>.
- [2] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988. ISBN 978-0805802832.
- [3] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. *arXiv preprint arXiv:2402.06698*, 2024. URL <https://arxiv.org/abs/2402.06698>.
- [4] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.
- [5] Antonino Ferraro and Giancarlo Sperli. How does user-generated content on social media affect stock predictions? a case study on GameStop. *Online Social Networks and Media*, 43–44:100293, 2024. doi: 10.1016/j.osnem.2024.100293.
- [6] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522.
- [7] Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *arXiv preprint arXiv:2309.17322*, 2023. URL <https://arxiv.org/abs/2309.17322>.
- [8] Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.
- [9] Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020. doi: 10.1093/rfs/hhy131.
- [10] Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. Multi-modal time series analysis: A tutorial and survey. *arXiv preprint arXiv:2503.13709*, 2025. URL <https://arxiv.org/abs/2503.13709>.
- [11] Sarthak Khanna, Armin Berger, David Berghaus, Tobias Deusser, Lorenz Sparrenberg, and Rafet Sifa. Towards unified multimodal financial forecasting: Integrating sentiment embeddings and market indicators via cross-modal attention. *arXiv preprint arXiv:2508.13327*, 2025. URL <https://arxiv.org/abs/2508.13327>.
- [12] Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300, 2024. URL <https://aclanthology.org/2024.findings-emnlp.485/>. Code: https://github.com/rosskoval/multimodal_ts_ff.
- [13] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.
- [14] Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023. URL <https://arxiv.org/abs/2304.07619>.
- [15] Brian W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. doi: 10.1016/0005-2795(75)90109-9.
- [16] R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016. doi: 10.1111/jofi.12365.
- [17] Yunhua Pei, John Cartlidge, Anandadeep Mandal, Daniel Gold, Enrique Marcilio, and Riccardo Mazzon. Cross-modal temporal fusion for financial market forecasting. *arXiv preprint arXiv:2504.13522*, 2025. URL <https://arxiv.org/abs/2504.13522>.

- [18] Jiahao Qin. MSMF: Multi-scale multi-modal fusion for enhanced stock market prediction. *arXiv preprint arXiv:2409.07855*, 2024. URL <https://arxiv.org/abs/2409.07855>.
- [19] Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL <https://arxiv.org/abs/2406.16964>.
- [20] Ankit Thakkar and Kinjal Chaudhari. Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions. *Information Fusion*, 65:95–107, 2020. doi: 10.1016/j.inffus.2020.08.019.
- [21] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018. URL <https://arxiv.org/abs/1710.10903>.
- [22] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering*. Springer, Berlin, Heidelberg, 2nd edition, 2012. ISBN 978-3-642-29044-2. doi: 10.1007/978-3-642-29044-2.
- [23] Sheng Xiang, Dawei Cheng, Chencheng Shang, Ying Zhang, and Yuqi Liang. Temporal and heterogeneous graph neural network for financial time series prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, 2022. doi: 10.1145/3511808.3557089.
- [24] Wenyan Xu, Dawei Xiang, Yue Liu, Xiyu Wang, Yanxiang Ma, Liang Zhang, Shu Hu, Chang Xu, and Jiaheng Zhang. FinMultiTime: A four-modal bilingual dataset for financial time-series analysis. *arXiv preprint arXiv:2506.05019*, 2025. URL <https://arxiv.org/abs/2506.05019>.
- [25] Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 2018. doi: 10.18653/v1/P18-1183.
- [26] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)*, pages 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.
- [27] Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C. Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W. Mahoney, Hao Wang, Yan Liu, Huzefa Rangwala, George Karypis, and Bernie Wang. When does multimodality lead to better time series forecasting? *arXiv preprint arXiv:2506.21611*, 2025. URL <https://arxiv.org/abs/2506.21611>.
- [28] Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. CAMEF: Causal-augmented multi-modality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*, 2025. doi: 10.1145/3711896.3736872.
- [29] Chang Zong and Hang Zhou. Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *arXiv preprint arXiv:2406.06594*, 2024. URL <https://arxiv.org/abs/2406.06594>. Published 2025 in *Complex & Intelligent Systems*, doi:10.1007/s40747-025-02023-3.

Abstract in lingua italiana

Nella letteratura sulla previsione finanziaria multi-sorgente i risultati positivi si susseguono, studio dopo studio: notizie, social media, indicatori macroeconomici e grafi di relazione tra titoli, combinati attraverso architetture sempre più elaborate, batterebbero il solo prezzo. La letteratura di *empirical asset pricing* restituisce però il quadro opposto: la maggior parte dei predittori di rendimento pubblicati non è replicabile e decade rapidamente dopo la pubblicazione, il che fissa un *prior* basso per qualunque nuova affermazione di prevedibilità sui titoli statunitensi ad alta liquidità. Questo lavoro prende sul serio tale prior e mette alla prova l'ipotesi multi-sorgente sulla previsione della direzione del rendimento del giorno successivo per 55 titoli statunitensi a grande capitalizzazione (2015–2023). Lo fa con un unico studio di ablazione controllato, composto da 2.468 esperimenti, che in ogni confronto mantiene costanti l'universo dei titoli, i *fold* temporali, i *seed* e le statistiche appaiate con correzione di Bonferroni. *Nessuna* combinazione di notizie, social media, indicatori macroeconomici o grafi supera in modo significativo il modello basato sul solo prezzo, e il risultato nullo si conferma lungo ogni asse considerato: tre strategie di fusione, due *target* di previsione, quattro *encoder* testuali e una variazione di un ordine di grandezza nel numero di parametri tra due architetture strutturalmente distinte. All'interno di questo risultato nullo, l'aggiunta di notizie rivolte al passato *peggiora* la previsione della direzione in quattro *fold* su cinque. Riproduciamo poi un noto risultato positivo della letteratura (MSGCA) e mostriamo, a partire dal codice rilasciato dagli stessi autori, che il guadagno riportato non sopravvive a una valutazione priva di *leakage*: il modello riportato è selezionato sull'insieme di test e, con una selezione basata sull'insieme di validazione e priva di *leakage*, a parità di dati e di modello, il vantaggio svanisce. Un secondo sistema (CAMEF) segue una via diversa: il suo risultato di punta si riproduce a partire dai pesi rilasciati dagli autori, ma crolla con una suddivisione train/test cronologica, che nel codice rilasciato è governata da un'opzione disattivata di default. Su questo problema contano più lo stimatore e il rigore della sua valutazione che non le sorgenti di dati aggiunte.

Parole chiave: previsione dei mercati finanziari, integrazione di dati multi-sorgente, studio di ablazione, distorsione nella selezione del modello, riproducibilità

Acknowledgements

I thank Prof. Fabrizio Pittorino and Francesco Puoti at Politecnico di Milano, and Prof. Molnár Bálint at Eötvös Loránd University, for their supervision.