# EVENTALIGN: AUDITING INPUT EXPOSURE IN SPARSE-RECORD FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Financial forecasts can be made daily, whereas earnings-call transcripts arrive only occasionally and may be reused for weeks. However, evaluating models under a single reuse rule makes it difficult to separate how well a learner uses text from how that text is assigned across prediction samples. Our key observation is that transcript assignment changes the information supplied to the predictor, making it part of the evaluation rather than a neutral preprocessing step. To isolate this effect, we introduce EventAlign, which varies transcript assignment while keeping prediction samples, labels, numerical histories, and training and model-selection procedures fixed within each exposure comparison. On our financial case study, whose primary close-to-close target includes the initial overnight and same-day intraday movements on first-eligible earnings-call rows, one-day use yields higher mean AP than 30-day reuse across all ten learners. With matched numerical histories and company information, the GRU–logistic-regression gap is 1.08 AP points under one-day use but 3.76 under prolonged reuse; fresh/stale channels change this interaction, while a separate background study shows that stronger company information improves the no-content baseline more than the text model. The main learner-gap interaction is much smaller and uncertain on subsequent return windows. Together, these results show that measured text gains and model advantages depend on transcript assignment, what the predictor already knows, and the target window.

## 1 INTRODUCTION

Our financial task is to predict whether a stock's daily return is unusually large in magnitude, using numerical histories and earnings-call transcripts. The numerical history changes with each prediction date, but the same transcript may be reused for weeks. A pipeline can use a transcript only on its first eligible trading day or carry it forward for 30 trading days. Both choices use the same transcript collection but assign text to different prediction samples. We call this assignment an *input exposure rule*. **We study how much the measured gap between two learners changes across these rules, with prediction samples and labels held fixed.**

Prior work links textual earnings disclosures to short-window market responses (Siano, 2025) and models earnings calls through verbal–vocal (Qin & Yang, 2019) and dialogue-structure (Sang & Bao, 2022) representations, long-document modeling (Koval et al., 2023), and text–numerical fusion (Koval et al., 2024). Measured text gains also depend on company information (Yu et al., 2025) and the predictive strength of the numerical model (Zhang et al., 2026); temporal methods study recency weighting (Chang et al., 2025). Recent evaluations examine forecasting sensitivity through controlled context replacement and also study ranking stability (Liu et al., 2026). Related work varies last-value shortcuts and context quality to test the contribution of contextual fusion (Zhou et al., 2026). We ask a complementary question about the comparison itself: when the same records are reassigned across fixed prediction samples, how much does the gap between learners change? Appendix K details the related comparisons.

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

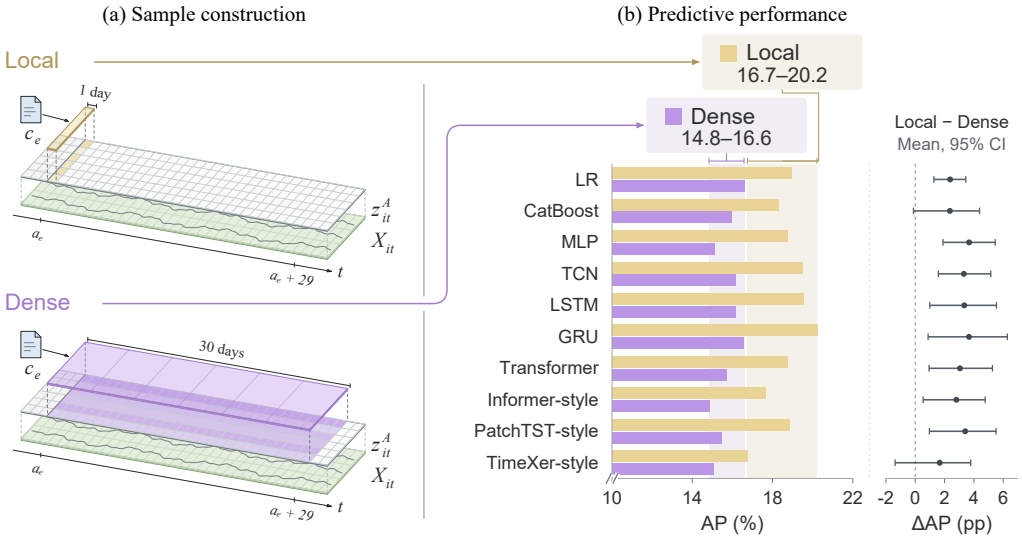



Figure 1: **One transcript, different inputs and scores.** Left: Local exposure lasts one eligible trading day; Dense lasts 30. History and outcomes stay fixed. Right: mean AP over six 2019–2021 starts for ten common-task adaptations. Paired Local-minus-Dense 95% intervals use 10,000 rolling-start bootstrap draws; two cross zero. Badges (16.7–20.2 and 14.8–16.6) give ranges of learner means, not confidence intervals.

Our key distinction is between a change in scores and a change in the model comparison. Figure 1 illustrates the former; an equal AP shift for a given pair of learners would leave their gap unchanged. **EventAlign** tests the latter through a paired learner-by-exposure interaction: the change in the between-learner AP gap across assignment rules (Section 2.3). Prediction samples, labels, numerical histories, the record collection, and training and model-selection procedures are held fixed within each comparison, with models retrained under each rule. Separate comparisons examine record-age encoding and background information.

With matched histories and company information, the primary close-to-close GRU–LR AP gap is $1.08\,\mathrm{pp}$ under one-day use and $3.76\,\mathrm{pp}$ under prolonged reuse (Figure 3). Fresh/stale channels change the interaction's sign and yield a smaller absolute point estimate. Separately, adding company information improves the no-content baseline more than the text model (Figure 5). The corresponding GRU–LR interaction is much smaller and uncertain on subsequent future-return windows. The primary label on first-eligible rows includes overnight and intraday movements around the stored call date. This interval can overlap the call and its initial market response, so the primary result is interpreted retrospectively.

Our contributions are:

- **An evaluation finding:** on the primary close-to-close target, transcript assignment changes the GRU–LR AP gap despite matched histories, company information, and configuration selection shared across input conditions within each learner.
- **A paired diagnostic protocol** that distinguishes individual score sensitivity from exposure-dependent model advantages, with separate contrasts for text-model margins and output allocation.
- **Evidence on conditions and scope:** input encoding and background information change the measured comparisons; target-window checks, controlled simulations, and external audits identify where related sensitivity persists or weakens.

## 2 DIAGNOSTIC PROTOCOL

EventAlign proceeds in three steps. Specify prediction rows, the history cutoff, target window, and record-availability evidence. Fit learners under paired exposure rules, fixing samples, records, histories, and model-selection procedures. Then test whether encoding or background information changes the comparison. Report absolute scores, paired gaps, uncertainty, and availability assumptions. Table 1 maps the questions to studies; Appendix B.4 gives release requirements.

Table 1: **Study map.** Does exposure change learner gaps (Model-gap), and does company information change text margins (Background)? Supporting studies test score sensitivity and alternative explanations. Samples, labels, and splits are paired within each comparison.

| | Comparison design | | |
|---|---|---|---|
| **Study** | **Fixed inputs** | **Varied factor** | **Selection / evidence** |
| **Model-gap study** 2022–24 | Histories, records, company information | Learner, exposure, encoding; five targets | Shared over three inputs per learner/target/period. Figure 3. |
| **Background study** 2022–24 | Dense exposure; $C_{t-1} \to C_t$ and $C_t \to C_{t+1}$ returns | Identity / past reactions; reference variants | Shared over backgrounds and variants per target/period. Figure 5. |
| **Breadth study** 2019–21 | Primary task, records, learner settings | Exposure for ten adapted learners | Fixed settings. Figure 1. |
| **Encoding study** 2022–24 | Primary target, histories | Exposure, channels; four learners, two representations | Shared over exposures per learner/encoding. Figure 4. |
| **Routing/age** Shared routing | Primary target, Dense content, expert capacity | Constant / Event / random; common decay | Routing-shared configuration. Appendix E. |
| **Routing/age** Learned age | Primary target, Dense content, expert backbone | Age-only gate (+13 parameters); common decay | Separate selection per condition. Table 2; Appendix F.1. |
| **Coverage audit** 2022–24 | Primary task, background | Sum / mean / coverage | Separate supporting protocol. Appendix A.5. |

*Note.* Shading identifies the main studies. Routing/age uses 2022–24. The separate four-feature clock gate (Appendix D.4) has unequal capacity. Scores are not pooled across protocols.

### 2.1 PREDICTION SAMPLES AND ELIGIBLE RECORDS

For entity $i$ and prediction row $t$, let $X_{it}$ be numerical history and $y_{it}$ the outcome. Record $e$ has entity $i_e$, content $c_e$, and first eligible time $a_e$. For earnings calls, $a_e$ is the next trading date after the stored call date, under the availability assumption in Section 3. An exposure rule $A$ selects records for each sample:

$$z_{it}^{A} = \sum_{e} A(e,i,t)\phi(c_e), \qquad A(e,i,t) \in \{0,1\}, \quad A(e,i,t) = 0 \text{ if } a_e > t. \tag{1}$$

For earnings calls, the raw record vector contains text coordinates, presence, and log length. Summing these vectors gives content, active count, and total log length; fitted transforms and channel layouts are study-specific (Table 5, Appendix B.2). Local exposure selects the first eligible trading day; Dense retains the same-firm record for 30 trading days, jointly changing coverage and record age. Delayed exposure shifts a one-day assignment by 30 or 60 days. A wrong-entity control breaks identity under the stated daily eligibility rule (Appendix Figure 8). Paired comparisons retain rows, labels, history source, record collection, and splits, but reconstruct inputs and retrain the learner.

### 2.2 Exposure, Encoding, and Background Information

*Eligibility* sets when a record may enter; *exposure* selects its recipient rows; *encoding* specifies its input representation; and *background* specifies other supplied information. Let $F$ contain first-eligible (fresh) records, $S$ older eligible (stale) records, and $C$ common timing features. Fresh and stale are record-age categories under this eligibility rule; older content may still be informative. We compare pooled content, pooled content with timing, and fresh/stale channels (Figure 2). The common clock gives event presence, latest-call age, 30-day call count, and absence of a previous call, independently of exposure.

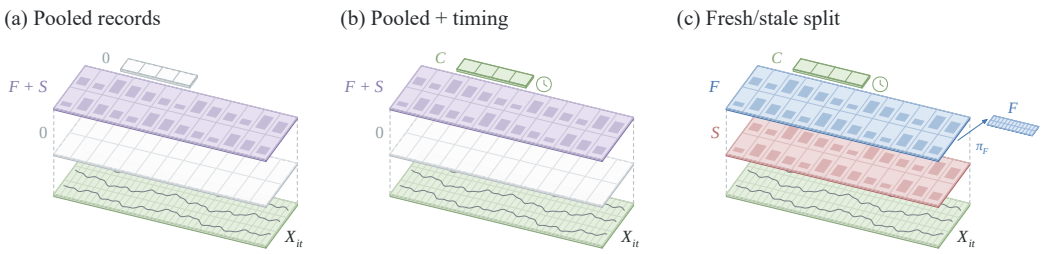



Figure 2: **Timing features add age information to pooled content; separate channels additionally allow different weights for fresh and stale content.** Pooled input supplies $F + S$; pooled + timing adds $C$; splitting supplies $F, S, C$ separately. 0 denotes padding; $\pi_F$ is the fixed projection onto the fresh channel. The schematic permits overlap, but fresh and stale channels do not co-occur in the six 2022–2024 test periods.

For a linear score, omitting the history and intercept terms gives

$$\underbrace{w^\top(F+S)+\gamma^\top C}_{\text{pooled + timing}} \quad \text{versus} \quad \underbrace{w_F^\top F + w_S^\top S + \gamma^\top C}_{\text{fresh/stale split}}. \tag{2}$$

For this additive linear score, separate channels permit different fresh/stale content weights; pooled + timing ties them together. Nonlinear learners may learn such interactions, and zero padding only matches storage dimensions. The Encoding study compares metadata and random partitions; the Routing/age studies compare shared-capacity routing and a separate learned-age gate. The Background study tests how adding firm identity and past-call return summaries, observed by the prediction cutoff, changes the text-model margin.

### 2.3 Paired Evaluation Contrasts

Let $P_s(A, g)$ be test average precision (AP) on start $s$ for fitted learner protocol $g$ under exposure $A$; encoding and background are implicit. We report AP levels in percent and differences in AP percentage points (pp). Define

$$\begin{aligned} E_{s,g} &= P_s(A_l, g) - P_s(A_d, g), \\ I_s &= \big[P_s(A_l, g_1) - P_s(A_l, g_0)\big] - \big[P_s(A_d, g_1) - P_s(A_d, g_0)\big]. \end{aligned} \tag{3}$$

$E$ measures one learner's exposure sensitivity; $I = E_{s,g_1} - E_{s,g_0}$ measures the change in the gap between learners, primarily GRU and LR. Both compare global held-out AP after fitting under each rule. Equal exposure effects give $I = 0$; a nonzero $I$ changes the relative advantage, even if the ranking stays the same.

The Routing/age and Background studies keep 30-day Dense exposure fixed: both text experts read the same pooled content. *Output routing* combines their scores. Constant averages them; Event selects one on age-zero days and the other otherwise, retaining older content. Routing therefore differs from changing exposure or splitting inputs. A no-content reference omits the text contribution while retaining numerical history and the specified metadata; informative sampling patterns in healthcare motivate retaining that information (Che et al., 2018; Chung et al., 2026). Section 4.3 introduces the background margins, with full definitions in Appendix A.4. EventGain and the separate history study's log-score definitions are in Appendix G.

# 3 EXPERIMENTAL SETUP

## 3.1 DATA, TASKS, AND TEXT COVERAGE

We classify large absolute returns on a daily firm panel, including rows without earnings events. The Breadth study has 1,114 transcripts for 57 firms in 2017–2021 (Kurry, 2025). For every row $t$, numerical inputs end at the preceding trading close. The primary label is $y_{it} = \mathbf{1}\{|C_{it}/C_{i,t-1}-1| \geq q_{i,s}^{95}\}$, where $C$ is adjusted close and $q$ a training-prefix quantile. A call is eligible on the next trading date after its stored date. On that first-eligible row, the label includes the initial overnight and intraday movements around the stored call date; other daily rows remain in the evaluation. Six expanding-window tests start semiannually in 2019–2021 and span 126 calendar days each.

Public earnings-call webcasts provide access during the event (Appendix B.4). We assume calls have ended and usable transcriptions are available by the next trading open, using archived transcripts as proxies. The primary interval can overlap the call and is evaluated retrospectively. Target checks use $|C_{i,t+h}/C_{it} - 1|$ for $h = 1, 3, 5$ trading days and $|C_{it}/O_{it} - 1|$ for open-to-close return, retaining previous-close histories (Table 4). Future close-to-close labels start at day-$t$ close. Timing assumptions are distinct from performance sensitivity (Kapoor & Narayanan, 2023). After observing the primary results, we added overnight-only and matched-endpoint comparisons (Appendix A.8).

The 2022–2024 tests retain the same firms and use a cached 2017–2024 corpus of 1,734 records (1,703 unique texts). The Model-gap and Encoding studies supply learners with the same $20 \times 18$ tensor of lagged returns, log volume, and rolling summaries within each study. LR flattens it; neural models process the sequence. Time-ordered partitions are $T, V_h, V_s, T_e$, with outcome endpoints purged at boundaries. In the Model-gap and Background studies, $V_h$ selects configurations and $V_s$ supplies additional final-refit data; labels and transforms stay fitted on $T$. Appendix B lists the other studies' uses of $V_s$, field definitions, and data examples.

Text representations are train-fitted TF-IDF/SVD24 and frozen FinBERT (Araci, 2019) followed by train-fitted SVD24. Full-text FinBERT uses a token-count-weighted average of chunk embeddings, without cross-chunk attention. The Encoding study includes a first-256-token alternative, covering a median 2.18% of content. Transforms use training data, and metadata-only conditions remove text coordinates while retaining counts, length, and the specified clock.

## 3.2 MATCHED STUDIES AND INFERENCE

The Breadth study adapts ten learners to the same classification task (Appendix C). The Encoding study compares LR, GRU (Cho et al., 2014), TCN (Bai et al., 2018), and Transformer (Vaswani et al., 2017) through a common 56-dimensional record-input interface, with separately fitted network parameters. For each learner and encoding, it selects one configuration shared across both exposure rules. The Routing/age study uses two nonlinear text experts and a common GRU history encoder; routing conditions share initializations and available content.

The Model-gap and Background studies use the same rows, restricted to observations valid for all five targets; the two flagged text records are excluded. The Model-gap study evaluates all five targets and averages validation AP across Local + timing, Dense + timing, and Dense split per learner/target/period. The Background study evaluates the $C_{t-1} \to C_t$ and $C_t \to C_{t+1}$ targets, selecting across four backgrounds and five reference variants per target/period. Budgets are matched *within a learner*; neural fits use two tuning and five final seeds. This separates configuration selection from testing (Cawley & Talbot, 2010); LR and GRU retain different fitting budgets. Appendix A.2 specifies their architectures and Table 5 their settings; Appendix B.5 preserves historical protocols.

We compute non-interpolated AP for each fitted model, then average first over seeds and then over periods. Breadth-study intervals resample six starts; the Encoding study uses paired date blocks. The Model-gap and Background studies use 20,000 linked-firm draws and two six-comparison families, with 2,000 five-/20-date-block checks (Appendix A). Supplementary pointwise intervals are labeled. Encoding, routing, and learned-age families are specified separately in Appendices D.5, E, and F.1. These intervals condition on the fitted predictions and selected configurations; they exclude variation from repeating model selection or refitting text representations (Bouthillier et al., 2021). External and simulation effects are reported separately; intervals crossing zero indicate uncertainty, not equivalence.

# 4 RESULTS

## 4.1 EXPOSURE CHANGES THE MEASURED MODEL ADVANTAGE

**GRU's AP advantage over LR depends on exposure.** The Model-gap study compares LR and GRU under Local and Dense exposure. Both learners receive identical numerical histories, full-text FinBERT, firm identity, and completed past-call reactions; each learner shares a selected configuration across input conditions within each target and period. On the primary close-to-close target, GRU leads by 1.08 pp under Local but 3.76 pp under Dense exposure (Figure 3). The interaction is $I = -2.68$ pp $[-4.74, -0.58]$, using six-comparison-adjusted linked-firm intervals. Assignment changes the size of GRU's advantage while preserving the ranking.

**The pooled interaction depends on the return window.** At a common day-$t$ close endpoint, the close-to-close interaction is more negative than the open-to-close interaction ($\Delta I = -2.33$ pp; three-comparison-adjusted 95% CI $[-4.37, -0.33]$); the next-close comparison remains uncertain (Appendix A.8). For 1-, 3-, and 5-day returns starting at day-$t$ close, the primary pooled interactions are $+0.21$, $+0.20$, and $+0.21$ pp, with all six-comparison-adjusted intervals crossing zero (Appendix A.3). First-eligible rows average 1.74% of evaluation rows and account for most of the primary interaction in the descriptive global-AP decomposition (Appendix A.6).

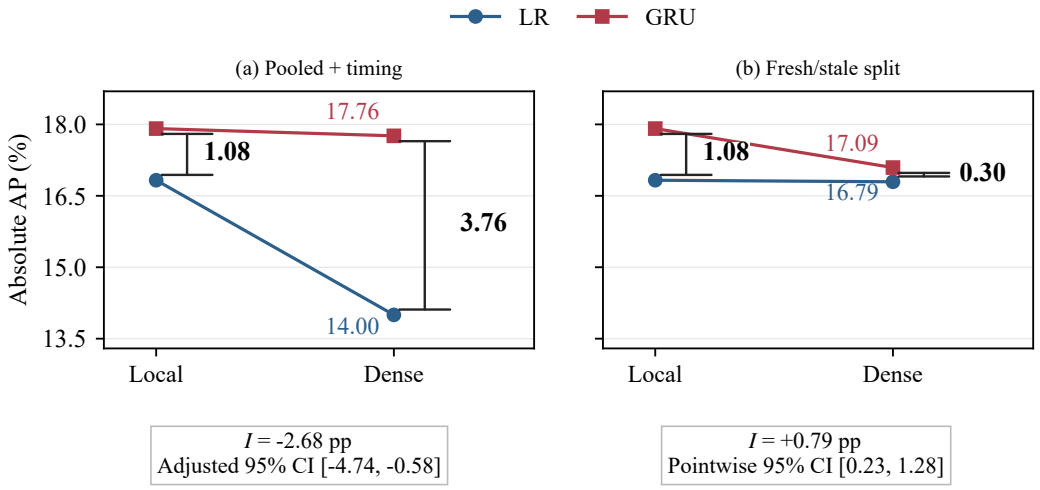



Figure 3: **The model gap is 1.08 or 3.76 pp, depending on exposure.** Model-gap study: primary target, full-text FinBERT with company information. Both panels reuse the same Local fits; only Dense encoding changes. Within each period, each learner uses one selected configuration across the three input conditions. Brackets: GRU-minus-LR AP; $I$: Local gap minus Dense gap. Boxes: linked-firm intervals, six-comparison-adjusted for pooled and pointwise for split. Scores average seeds, then six periods; intervals condition on fitted predictions.

**The negative interaction persists across seeds and period omissions.** Both date-block intervals lie below zero. Refitting models with configurations selected on the close-to-close target gives $I = -2.92$ pp (Table 12); all five seed means and all leave-one-period-out means remain negative, although individual periods differ (Appendix A.6).

In the separate Coverage audit, matching the number of content-exposed rows within each date retains a negative interaction ($-3.51$ pp; pointwise 95% CI $[-4.75, -2.18]$; Appendix A.5).

**Exposure dependence extends beyond GRU–LR.** All ten Breadth-study means favor Local, with two intervals crossing zero (Figure 1). Encoding-study score changes appear with both representations, without company augmentation (Figure 4); delay and wrong-entity controls appear in Appendix C. In the Model-gap study, TCN–LR and Transformer–LR close-to-close interactions are $-2.32$ and $-2.13$ pp, with twelve-comparison-adjusted intervals below zero. Interactions among the three neural learners, and all six pairs on $C_t \to C_{t+1}$, remain uncertain (Appendix A.3).

With inherited configurations, LR gains $+2.97$ pp (Local) and $+0.15$ pp (Dense) over no content on close-to-close returns; only Local has a thirty-comparison-adjusted interval above zero. GRU's margins are $-0.93$ and $-1.08$ pp, with adjusted intervals crossing zero (Appendix A.3). Absolute AP and text benefit therefore differ; Section 4.3 examines background-dependent margins under a separate two-expert protocol.

### 4.2 Age-Aware Inputs Change the Model Comparison

**Fresh/stale channels change the two learners differently on the primary close-to-close target.** Under Dense exposure, LR rises from 14.00% to 16.79%, while GRU falls from 17.76% to 17.09%. The Dense GRU–LR gap therefore falls from 3.76 to 0.30 pp, while the Local gap remains 1.08 pp. The interaction's point estimate changes sign, from $-2.68$ to $+0.79$ pp; the split estimate has a supplementary pointwise interval [0.23, 1.28].

The Encoding study tests the pattern across four learners without adding company identity or past-call reactions (Figure 4). Timing features reduce exposure gaps more for neural learners than for LR; splitting brings all eight mean gaps between $-0.28$ and $+0.12$ pp. The pattern also appears when FinBERT encodes the full transcript rather than only its opening tokens, whereas metadata-only and date-activity-matched random partitions reproduce only part of it. Appendix D reports absolute scores and paired uncertainty.

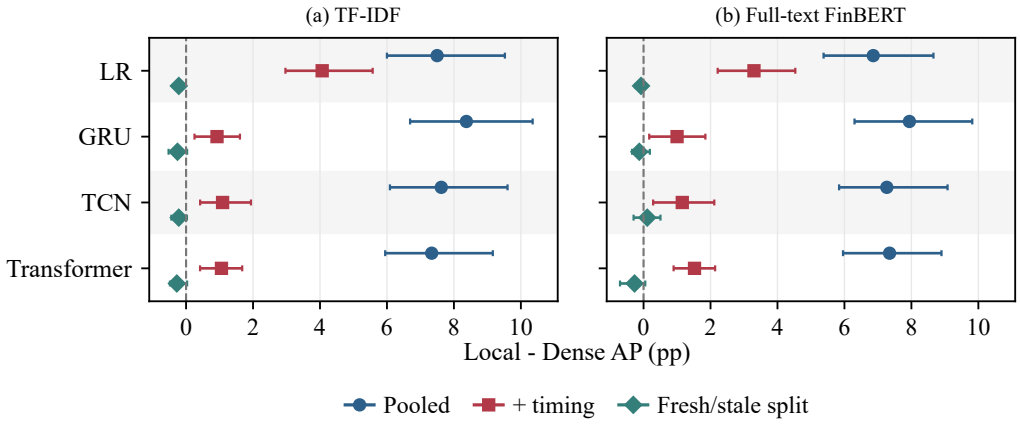



Figure 4: **Input encoding changes exposure sensitivity.** Encoding study: Local-minus-Dense AP ($E$), four matched-history learners, six 2022–2024 starts. Bars: pointwise 95% circular 20-date-block intervals (10,000 fixed-prediction resamples). Paired tests: Tables 40–41. This Encoding study matches numerical histories without company augmentation; the Model-gap study in Figure 3 supplies company information to both learners.

**Event routing outperforms equal-parameter alternatives.** The shared-configuration Routing/age comparison holds Dense content fixed and varies how two experts' outputs are combined. With FinBERT, Event routing exceeds Constant and date-matched random routing by 2.52 and 2.39 pp, respectively, with adjusted intervals excluding zero; these shared-routing comparisons use equal parameter counts. Common decay reduces the Event-minus-Constant margin by 1.66 pp [0.85, 2.53] (pointwise), leaving 0.86 pp [0.13, 1.55] (adjusted). A separate age-only gate adds 13 parameters and selects configurations separately for each condition; its comparison with Event routing is uncertain under common decay (Table 2).

Table 2: **Common decay makes the advantage over learned age uncertain.** Full-text FinBERT, fixed Dense inputs, common expert backbone and search budget; the age-only gate adds 13 parameters.

| Common decay | Event minus learned age AP difference (pp) | Adjusted 95% CI |
|---|---|---|
| None | 2.52 | [1.19, 3.99] |
| Exponential | 0.60 | [−0.10, 1.32] |

*Note.* Each cell selects its configuration separately. Linked-firm intervals adjust four comparisons. Full protocol: Appendix F.1.

### 4.3 Company Information Narrows the Text-Model Margin

**Company information improves the no-content baseline more than the text model.** We now fix Dense exposure and vary company information in the separate two-expert Background study, with configuration selection shared across backgrounds and reference variants. Adding identity alone raises no-content AP from 13.82% to 15.34%; adding completed past-call reactions alone raises it to 16.33%. Together they raise it to 17.15%, while Event text rises from 15.95% to 17.51% (Figure 5).

Define the text-model margin $V$ as Event minus no-content AP, and the allocation margin $R$ as Event minus Constant AP.

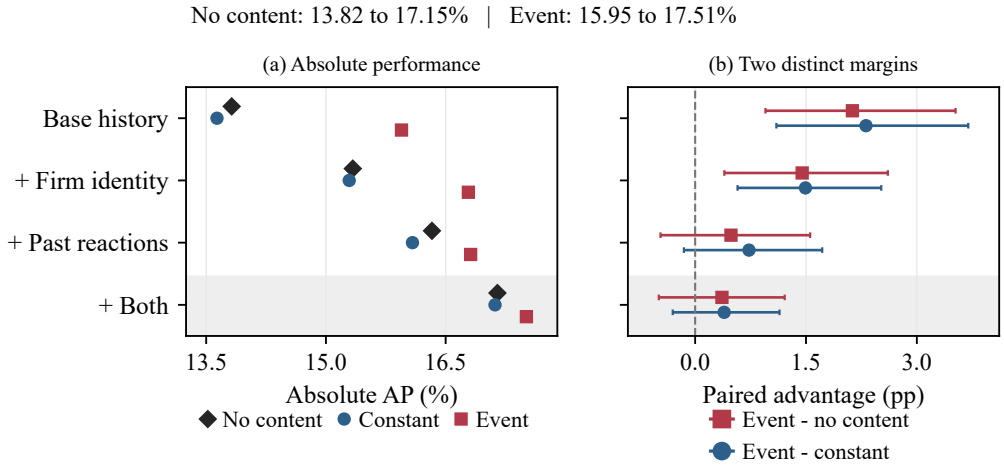



Figure 5: **The baseline catches up; measured margins shrink.** Background study: close-to-close target, full-text FinBERT, fixed Dense exposure and two experts. No content retains metadata; Constant averages expert scores; Event routes by record age. Right: event-minus-no-content ($V$) and event-minus-constant ($R$), with pointwise linked-firm intervals. Grey marks the joint background. The direct change in $V$ uses six-comparison adjustment.

$V$ falls from 2.13 to 0.36 pp; its direct change is $-1.76$ pp $[-2.93, -0.80]$, adjusted over six comparisons. $R$ falls descriptively from 2.31 to 0.39 pp. The base-background margin over equal-parameter numerical experts and the matched current–past content contrast remain uncertain (Appendix A.4, Table 15). For the $C_t \to C_{t+1}$ target, the change in $V$ is $-0.05$ pp $[-0.35, 0.18]$ (Appendix A).

The separate history study finds both increasing and decreasing text margins, with less conclusive full-text and log-score evidence (Appendix G).

### 4.4 Scope Across Controlled and External Studies

External audits distinguish changes in gain magnitude from changes in its sign (Table 3). In a frozen-model VoT audit on Energy, increasing cache age from 0 to 12 weeks lowers the event-module MSE gain from $+0.0680$ to $-0.0515$, even with target dates aligned. The direct change is $-0.1195$ (38-comparison simultaneous 95% CI $[-0.1882, -0.0508]$); the older-cache gain itself remains uncertain. In Aurora, all 72 mean text gains remain positive across 18 record positions and four forecast horizons.

VoT retains batch-wide normalization, which couples predictions across forecast origins; the cache audit is retrospective (Appendix I.1).

The simulations pair four combinations of signal persistence and state-specific coefficients over 50 independent replicates. With slow decay and different state coefficients, state-sensitive treatment helps beyond continuous-age controls: the adjusted lower bound exceeds the generator's $0.002$-nat reference. Appendix H retains all conditions and eight contrasts, including small effects.

Table 3: **External interface case studies at three intervention levels.** Each system retains its own metric and uncertainty analysis.

| System / intervention | Finding | Comparison scope |
|---|---|---|
| **VoT** (Wang et al., 2026)<br>Cache age; frozen models | Energy module gain falls by **0.1195 MSE** from age 0 to 12 weeks; its simultaneous change interval is below zero. | Target-aligned caches with models and batch composition fixed (Table 52; Appendix I.1). |
| **Aurora** (Wu et al., 2026)<br>Record selection; fixed checkpoint | All **72 mean text gains are positive** across 18 positions and four horizons, with varying magnitudes. | Record selection changes scores without a mean sign reversal in this grid (Table 53). |
| **XTraffic** (Gou et al., 2025)<br>Event encoding and recent-history visibility | No consistent splitting benefit. Recent speed reduces GRU's exploratory incident-over-metadata margin. | GRU-specific margin over metadata; absolute AP relative to no-event is reported separately (Appendix J.1). |

*Note.* VoT gain is MSE without event inputs minus MSE with them. Effects are not pooled across systems; screened cases remain in Appendix I.3.

## 5 Conclusion and Limitations

EventAlign tests whether transcript assignment changes a model comparison. On the primary close-to-close target, the learner gap changes despite matched histories and company information. Fresh/stale channels change the interaction's sign and yield a smaller absolute point estimate, with a residual pointwise effect. Separately, company information improves the no-content baseline more than the text model. Reporting assignment rules, target windows, reference predictors, and paired interactions alongside absolute scores helps distinguish model advantage from incremental text benefit.

The primary window can overlap the call and its initial market response, so these comparisons are retrospective under the stated transcript-availability assumption. The corresponding pooled interactions on subsequent targets are smaller and uncertain. Inference is conditional on fitted predictions and excludes full-workflow refitting uncertainty. Simulations and external audits, including retained null cases (Appendix I.3), show where related sensitivity recurs or weakens.

## AI USE STATEMENT

Generative AI tools assisted with literature discovery, conceptual framing and hypothesis refinement, experimental design, code implementation, result analysis and interpretation, manuscript drafting and translation, and figure preparation. The authors reviewed and verified all AI-assisted material incorporated into the submission and take responsibility for the final content.

## ETHICS STATEMENT

The financial experiments concern retrospective predictive evaluation; causal market effects and trading profitability after costs are outside their scope. External audits address the specified released interfaces; use and distribution of source data and model artifacts remain subject to their original terms.

## REPRODUCIBILITY STATEMENT

Tables 5–7 specify current inputs, partitions, and selected configurations; Appendices A–B give target and inference definitions. A prepared aggregate reproduction package reconstructs Figures 3 and 5 from saved exports, checks their arithmetic and source hashes, and includes the plotting code and protocol manifests. This path has been tested locally; clean-environment reproduction from raw data through training remains unvalidated. The package excludes raw transcripts and row-level artifacts. Public release remains subject to dependency and redistribution checks (Appendix B.4).

## REFERENCES

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

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wei Wang, Wen-Chih Peng, and Tien-Fu Chen. Time-IMM: A dataset and benchmark for irregular multimodal multivariate time series. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Peng Chen, Zhihao Zhuang, Hongzhou Chen, Junhao Huang, Aiping Yang, Mengsen Wu, Yiding Liu, Xilin Dai, and Zewei Dong. Beyond numerical time series: A unified benchmark for multimodal forecasting with heterogeneous context. arXiv preprint arXiv:2609.15087, 2026.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

Hsing-Huan Chung, Shijun Li, Yoav Wald, Xing Han, Suchi Saria, and Joydeep Ghosh. MILM: Large language models for multimodal irregular time series with informative sampling. arXiv preprint arXiv:2605.13711, 2026.

Lixiang Fan, Bohao Li, Tao Zou, Junchen Ye, and Bowen Du. Incident-guided spatiotemporal traffic forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 243–254, 2026.

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Zhishuai Li, Bingyu Zhao, Chen Zhang, Di Wang, and Xiangliang Zhang. TraffiDent: A dataset for understanding the interplay between traffic dynamics and incidents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. Dataset released as XTraffic.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. UniCA: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, 2026.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Benjamin Irving and Annika Marie Schoene. MEANT: Multimodal encoder for antecedent information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8579–8600, 2024.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Ross Koval, Nicholas Andrews, and Xifeng Yan. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8197–8209, 2023.

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*, pp. 8289–8300, 2024.

Kurry. S&P 500 earnings transcripts dataset. Hugging Face dataset, `https://huggingface.co/datasets/kurry/sp500_earnings_transcripts`, 2025. Revision 71db6d3.

Seunghan Lee, Jun Seo, Jaehoon Lee, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, SoonYoung Lee, and Wonbin Ahn. Rethinking multimodal fusion for time series: Text modalities need constrained fusion. arXiv preprint arXiv:2603.22372, 2026.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77888–77933, 2024.

Haoxin Liu, Yichen Zhou, Rajat Sen, B. Aditya Prakash, and Abhimanyu Das. Rethinking multimodal time-series forecasting evaluation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yu Qin and Yi Yang. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 390–401, 2019.

Yunxin Sang and Yang Bao. DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1623–1633, 2022.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8001–8013, 2020.

Federico Siano. The news in earnings announcement disclosures: Capturing word context using LLM methods. *Management Science*, 71(11):9831–9855, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *International Conference on Learning Representations*, 2026.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498, 2024.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 66887–66944, 2025.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, 2026.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

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 (Volume 1: Long Papers)*, pp. 1970–1979, 2018.

Zhijian Xu, Wanxu Cai, Xilin Dai, Zhaorong Deng, and Qiang Xu. Fidel-TS: A high-fidelity multimodal benchmark for time series forecasting. arXiv preprint arXiv:2509.24789, 2025.

Ding Yu, Zhuo Liu, and Hangfeng He. Same company, same signal: The role of identity in earnings call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18403–18422, 2025.

Zixuan Yuan, Yada Zhu, Wei Zhang, and Hui Xiong. Earnings call analysis using a sparse attention based encoder and multi-source counterfactual augmentation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 331–339, 2023.

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, Michael Bohlke-Schneider, Huzefa Rangwala, George Karypis, and Bernie Wang. When does multimodality lead to better time series forecasting? *Transactions on Machine Learning Research*, 2026.

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 V.2*, pp. 3867–3878, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Ruizhe Zhou, Gaoyuan Du, Xiaoyang Liu, Haoqi Yao, Deepayan Chakrabarti, Jiating Lin, and Yixuan Shen. When does context routing help? A systematic study of multi-modal fusion in time series forecasting. arXiv preprint arXiv:2608.25128, 2026.

## GUIDE TO THE APPENDIX

Appendix A reports the Model-gap and Background results, all return-window checks, and their inference. Data and implementation details follow in Appendix B. The supporting evidence covers breadth and encoding (Appendices C–D), output routing and learned age (Appendices E–F), additional history and simulated conditions (Appendices G–H), and external scope including traffic (Appendices I–J). Appendix K compares the nearest work. Each study retains its own selection and uncertainty definitions.

## A MAIN COMPARISONS AND RETURN-WINDOW SENSITIVITY

This section gives the Model-gap and Background protocols for Figures 3 and 5 and reports every planned return window. Here and below, "predeclared" and "registered" refer to local analysis plans fixed after these retrospective periods had already been inspected.

### A.1 RETURN WINDOWS AND INFORMATION TIMING

Let $t$ index a daily prediction row, and let $C_t$ and $O_t$ denote adjusted close and open. All valid daily rows are retained, including rows without a call. For a record $e$, only its first-eligible row has $t = a_e$, the first trading date strictly after its stored call date. The primary target uses $|C_t/C_{t-1} - 1|$, combining the overnight movement into $t$ and its intraday movement. On a first-eligible row, a call after the preceding close may already affect this interval before the transcript is accessible.

Table 4: Predeclared return targets. Each absolute return is converted to a binary label using its own firm-specific 95th percentile on valid common training rows. All horizons count trading days.

| Return window | Absolute return | Window interpretation |
|---|---|---|
| $C_{t-1} \to C_t$ | $\lvert C_t/C_{t-1} - 1\rvert$ | Initial overnight plus day-$t$ intraday movement |
| $C_t \to C_{t+1}$ | $\lvert C_{t+1}/C_t - 1\rvert$ | Following close-to-close day |
| $C_t \to C_{t+3}$ | $\lvert C_{t+3}/C_t - 1\rvert$ | Three-day return beginning at day-$t$ close |
| $C_t \to C_{t+5}$ | $\lvert C_{t+5}/C_t - 1\rvert$ | Five-day return beginning at day-$t$ close |
| $O_t \to C_t$ | $\lvert C_t/O_t - 1\rvert$ | Day-$t$ intraday return only |

We assume completed-call content can be transcribed by the next trading open and use archived transcripts as its proxy (Appendix B.4). Numerical histories still end at the previous close for every target. Future close-to-close labels begin at day-$t$ close, after the assumed information cutoff; they may contain later overnight returns, but exclude the movement from $C_{t-1}$ to $O_t$. Open–close is the separately registered intraday check. Because the target comparison changes both the outcome and its window, it probes a different return interval rather than isolating the initial overnight component.

### A.2 COMMON ROWS AND SHARED SELECTION

All 57 firms and six 2022–2024 starts are retained. Two records are quarantined: AMD dated 2017-01-25 contains Xilinx text, and WMT dated 2021-02-20 has a conflicting stored date. Their price rows remain. Text projections, event-time features, counts, and completed past-call reaction summaries are computed after excluding these records. The intersection of valid masks across all five targets defines common eligible rows; label endpoints crossing training, validation, or test boundaries are purged. Thresholds and transforms use initial $T$ and stay fixed for the final pre-test refit. All fits reuse frozen token-weighted FinBERT embeddings.

The Model-gap study supplies the same $20 \times 18$ numerical history and joint firm-identity/past-reaction background to all learners. For each target, learner, and period, a shared configuration maximizes mean validation AP over three distinct input conditions: Local with timing, Dense with timing, and Dense fresh/stale split. Local split is identical to Local timing and is not counted twice. LR searches $C \in \{.01, .1, 1, 10, 100\}$; neural models use the six-setting grid in Table 32, two tuning seeds, and five final seeds. No-content models use the same selected architecture configuration but are not included in its selection objective. The Model-gap LR fits use L-BFGS with tolerance $10^{-4}$ and a 5,000-iteration cap, adopted after an initial convergence problem.

The Background study uses a common 84-coordinate context interface and selects one configuration per target/period by equally averaging validation AP over four backgrounds and five variants: no content, text constant/event allocation, and numerical constant/event allocation. No-content models share this selection. Numerical experts receive existing history states and the same dense-presence gating; they match active parameter counts, not function classes. Historical-content controls replace each record by the previous distinct same-firm transcript available strictly earlier, while retaining recipient metadata and routing. If no donor exists, both matched arms mask content. This tests record correspondence under a shared mask.

The Model-gap and Background protocols differ from the historical fits in quarantine, common valid rows, transforms, boundary purging, and shared selection. Cross-protocol score changes therefore combine these differences. A configuration-transfer check refits weights for every target using shared configurations selected on close-to-close validation AP, with their source and exact values given in Table 12.

**Main-study protocol.** Tables 5–7 specify the settings for Figures 3 and 5. Calendar boundaries follow Table 30; main-study row counts use the common five-target masks after sequence filtering and endpoint purging (Table 6).

Table 5: Effective protocol for the current main figures. All fitting prefixes use the common valid masks after sequence filtering and endpoint purging.

| Component | Current rule |
|---|---|
| Rows and time | For each start, use the preceding 126 trading dates to define two 63-date validation blocks. $T$ precedes $V_h$; $V_s$ precedes the test start. Test ends 126 calendar days after the start. In each partition, intersect finite 20-row histories and valid labels across all five targets; every label endpoint must precede that partition's next boundary. |
| Labels and fitting | Each firm's target-specific 95th percentile uses common $T$ rows with linear interpolation. Tune weights on $T$, select on $V_h$, refit on $T \cup V_h \cup V_s$, test on $T_e$. Labels and transformations stay frozen at initial $T$; $V_s$ is additional refit data, not a second model-selection stage here. |
| Shared transformations | Numerical sequences use means and population standard deviations pooled over common-$T$ sequence positions, then clip to $[-12, 12]$. Completed-reaction coordinates use common-$T$ affine scaling without clipping. FinBERT is frozen; SVD24 uses records eligible by the final common-$T$ row. Identity uses the $T$ vocabulary plus an unknown coordinate. |
| Figure 3 inputs | Each condition forms 48 content coordinates: Local/pooled content plus zero padding, or fresh/stale content. Each condition fits its own coordinate means and standard deviations on the same $T$ rows, without clipping. The 74 shared context coordinates contain eight timing/count/length fields, 58 identity fields, and eight reaction fields. No content zeros all 48 content coordinates but retains context and history. |
| Figure 5 inputs | All variants reuse the same dense text sum, scaled by unique training-record standard deviations without mean subtraction. Context width is 84: the same eight metadata fields, ten fixed random-routing indicators, 58 identity fields, and eight reaction fields. Disabled background blocks are zeroed after shared scaling, not restandardized per condition. |
| Selection | Figure 3: maximize equally weighted mean validation AP across three input conditions, separately for each learner, target, and period. Its no-content variant inherits that choice. Figure 5: maximize equally weighted mean validation AP across four backgrounds and five reference variants for each target and period. Ties select the first candidate; no test score enters selection. |

**Model-gap architectures.** In Figure 3, the record input has 122 coordinates: 48 content coordinates and the 74 shared context coordinates in Table 5. LR concatenates the flattened $20 \times 18$ history with this record input, yielding 482 predictors for an L2-regularized logistic model with an intercept. GRU uses a single unidirectional layer with 18 inputs and 24 hidden units; its final hidden state represents the history. A linear layer maps the record input from 122 to 16 coordinates, followed by

Table 6: Current common-row counts. The final fit unions the three earlier partitions without restoring purged boundary rows. Masks match across representations, conditions, and all five targets.

| Start | $T$ | $V_h$ | $V_s$ | Final fit | $T_e$ |
|---|---|---|---|---|---|
| 2022-01-01 | 62586 | 3306 | 3306 | 69198 | 4674 |
| 2022-07-01 | 69654 | 3306 | 3306 | 76266 | 4731 |
| 2023-01-01 | 76893 | 3306 | 3306 | 83505 | 4617 |
| 2023-07-01 | 83961 | 3306 | 3306 | 90573 | 4731 |
| 2024-01-01 | 91143 | 3306 | 3306 | 97755 | 4617 |
| 2024-07-01 | 98211 | 3306 | 3306 | 104823 | 4731 |

Table 7: Actual shared configurations for the close-to-close, full-text main figures; IDs are keys to the numerical settings below, not model names. LR uses balanced L-BFGS, tolerance $10^{-4}$ and a 5,000-iteration cap. Neural batch size is 1,024; tuning seeds are 419/631 and final seeds are 419/631/853/1061/1297.

| Start | Figure 3 LR $C$ | Figure 3 GRU | Figure 5 |
|---|---|---|---|
| 2022-01-01 | 100 | N4 | N4 |
| 2022-07-01 | 100 | N1 | N0 |
| 2023-01-01 | 100 | N2 | N3 |
| 2023-07-01 | 0.01 | N2 | N3 |
| 2024-01-01 | 100 | N2 | N4 |
| 2024-07-01 | 100 | N3 | N4 |

| ID | Learning rate | Weight decay | Epochs |
|---|---|---|---|
| N0 | 0.0003 | 1e-05 | 16 |
| N1 | 0.0003 | 0.0001 | 16 |
| N2 | 0.001 | 1e-05 | 16 |
| N3 | 0.001 | 0.0001 | 16 |
| N4 | 0.001 | 0.0001 | 32 |
| N5 | 0.001 | 0.001 | 32 |

ReLU. The concatenated 40-dimensional vector passes through a 32-unit ReLU layer, dropout 0.10, and a one-unit linear output; sigmoid gives the probability. Each exposure/encoding condition fits its own weights under the shared configuration. The TCN/Transformer extensions use the encoders in Appendix D, with the same 122-coordinate record input and the selection rules above.

**Background architecture.** For Figure 5, the shared GRU produces a 24-dimensional history state $h$. The base head maps $[h, c]$ through 32 ReLU units and dropout 0.1 to one logit. Each expert $u, v$ maps 24 inputs through 16 ReLU units to one bias-free scalar output. Text routing adds $q[u(P)-u(0)]+(1-q)[v(P)-v(0)]$ to the base logit; $q = 1/2$ for Constant and $q = \mathbf{1}\{\text{fresh record}\}$ for Event. No content omits these experts but keeps the base head and metadata. Capacity-matched numerical routing instead adds $d\{q[u(h) - u(0)] + (1 - q)[v(h) - v(0)]\}$, where $d$ is 30-day record presence. It uses the same active parameter count, not the same function class or additional numerical observations.

### A.3 Model Comparisons Across Every Target

Tables 8–9 give all architecture means. AP is calculated for each fit and then averaged over seeds and periods; predictions are never ensembled before AP is computed. Each target defines a different label distribution and prevalence, so absolute AP levels are not directly comparable across targets. Supplementary paired interactions cover all six model pairs on the two targets available for all four learners (Table S2).

**No-content references and paired margins.** In the Model-gap study, no-content GRU averages 18.84% AP, compared with 17.91% under Local and 17.76% under Dense; no-content LR averages 13.85%, compared with Local LR's 16.83% (Table 8). These references separate absolute model

performance from incremental text value. They inherit selected configurations; Table S1 reports paired text-minus-no-content intervals. The two-expert Background study below estimates text-model margins under its own protocol. In the notation of Section 2, the within-exposure gap can also be written $D_s(A) = P_s(A, g_1) - P_s(A, g_0)$, giving $I_s = D_s(A_l) - D_s(A_d)$.

Table 8: Model-gap study with full-text FinBERT, mean AP (%). Local and Dense include common timing features. All models receive joint company background. No content retains metadata.

| Target | Learner | No content | Local | Dense | Dense split |
|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | LR | 13.85 | 16.83 | 14.00 | 16.79 |
| $C_{t-1} \to C_t$ | GRU | 18.84 | 17.91 | 17.76 | 17.09 |
| $C_{t-1} \to C_t$ | TCN | 17.10 | 17.27 | 16.76 | 17.11 |
| $C_{t-1} \to C_t$ | Transformer | 16.36 | 16.73 | 16.04 | 16.30 |
| $C_t \to C_{t+1}$ | LR | 7.98 | 7.93 | 8.01 | 7.95 |
| $C_t \to C_{t+1}$ | GRU | 7.85 | 7.72 | 7.59 | 7.55 |
| $C_t \to C_{t+1}$ | TCN | 7.39 | 7.36 | 7.33 | 7.29 |
| $C_t \to C_{t+1}$ | Transformer | 7.42 | 7.22 | 7.17 | 7.06 |
| $C_t \to C_{t+3}$ | LR | 8.79 | 8.79 | 8.78 | 8.78 |
| $C_t \to C_{t+3}$ | GRU | 9.51 | 9.43 | 9.22 | 9.05 |
| $C_t \to C_{t+5}$ | LR | 8.81 | 8.81 | 8.71 | 8.73 |
| $C_t \to C_{t+5}$ | GRU | 9.42 | 9.43 | 9.12 | 9.23 |
| $O_t \to C_t$ | LR | 10.14 | 10.39 | 10.11 | 10.35 |
| $O_t \to C_t$ | GRU | 9.92 | 9.87 | 9.94 | 9.80 |

Table 9: Corresponding TF-IDF grid, mean AP (%). Same target set and selection design as Table 8; each representation fits its own training-only transform.

| Target | Learner | No content | Local | Dense | Dense split |
|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | LR | 13.84 | 17.46 | 13.90 | 17.54 |
| $C_{t-1} \to C_t$ | GRU | 18.88 | 18.46 | 18.58 | 18.29 |
| $C_{t-1} \to C_t$ | TCN | 16.73 | 17.17 | 16.37 | 17.02 |
| $C_{t-1} \to C_t$ | Transformer | 17.14 | 16.94 | 16.85 | 16.80 |
| $C_t \to C_{t+1}$ | LR | 7.96 | 7.93 | 8.01 | 7.95 |
| $C_t \to C_{t+1}$ | GRU | 8.05 | 7.95 | 7.77 | 7.70 |
| $C_t \to C_{t+1}$ | TCN | 7.30 | 7.30 | 7.23 | 7.29 |
| $C_t \to C_{t+1}$ | Transformer | 7.47 | 7.45 | 7.30 | 7.38 |
| $C_t \to C_{t+3}$ | LR | 8.79 | 8.75 | 8.82 | 8.78 |
| $C_t \to C_{t+3}$ | GRU | 9.56 | 9.45 | 9.28 | 9.29 |
| $C_t \to C_{t+5}$ | LR | 8.75 | 8.71 | 8.81 | 8.81 |
| $C_t \to C_{t+5}$ | GRU | 9.36 | 9.15 | 9.26 | 9.14 |
| $O_t \to C_t$ | LR | 10.36 | 10.52 | 10.32 | 10.46 |
| $O_t \to C_t$ | GRU | 9.73 | 9.83 | 9.69 | 9.81 |

The primary full-text architecture family contains pooled $I$ for all five targets and the $C_t \to C_{t+1}$ split-minus-pooled change. Linked-firm intervals use 20,000 paired resamples, synchronizing firm weights across periods. Six-comparison Bonferroni percentile endpoints are used for that family; supplementary split interactions and other changes are pointwise. Date-block checks use 2,000 draws of common circular five-/20-date blocks within periods. All intervals condition on fitted predictions and selected configurations, not full-workflow refitting.

Table 10: All full-text target interactions, AP pp. $I = (\mathrm{GRU} - \mathrm{LR})_{local} - (\mathrm{GRU} - \mathrm{LR})_{dense}$. Pooled intervals adjust six comparisons; split intervals are supplementary pointwise 95% intervals.

| Target | $I$, pooled | Adjusted interval | $I$, split | Pointwise interval |
|---|---|---|---|---|
| $C_{t-1} \to C_t$ | -2.675 | [-4.740, -0.583] | +0.786 | [0.231, 1.276] |
| $C_t \to C_{t+1}$ | +0.205 | [-0.196, 0.681] | +0.182 | [-0.018, 0.453] |
| $C_t \to C_{t+3}$ | +0.201 | [-0.305, 0.818] | +0.366 | [0.065, 0.767] |
| $C_t \to C_{t+5}$ | +0.207 | [-0.212, 0.920] | +0.114 | [-0.138, 0.553] |
| $O_t \to C_t$ | -0.342 | [-1.240, 0.499] | +0.039 | [-0.197, 0.283] |

Table 11: Direct signed interaction changes, AP pp. Only $C_t \to C_{t+1}$ belongs to the primary six-comparison family. These are not intervals for $|I_{split}| - |I_{pooled}|$.

| Target | $I_{split} - I_{pooled}$ | Interval | Adjustment |
|---|---|---|---|
| $C_{t-1} \to C_t$ | +3.461 | [1.769, 5.019] | Pointwise |
| $C_t \to C_{t+1}$ | -0.023 | [-0.298, 0.328] | Six-comparison |
| $C_t \to C_{t+3}$ | +0.165 | [-0.121, 0.474] | Pointwise |
| $C_t \to C_{t+5}$ | -0.093 | [-0.420, 0.176] | Pointwise |
| $O_t \to C_t$ | +0.381 | [-0.206, 0.986] | Pointwise |

Table 12: Refitting close-to-close-selected shared configurations on each target, full-text AP pp. All intervals are supplementary pointwise linked-firm intervals. Weights are refitted on each target's common rows. The lower panel lists the transferred settings.

| Target | $I$, pooled | Pointwise interval | $I$, split | Pointwise interval |
|---|---|---|---|---|
| $C_{t-1} \to C_t$ | -2.925 | [-4.406, -1.438] | +0.530 | [0.031, 0.931] |
| $C_t \to C_{t+1}$ | +0.191 | [-0.104, 0.516] | +0.194 | [0.013, 0.433] |
| $C_t \to C_{t+3}$ | +0.295 | [-0.188, 0.787] | +0.371 | [0.024, 0.756] |
| $C_t \to C_{t+5}$ | +0.422 | [0.004, 0.993] | +0.260 | [-0.140, 0.786] |
| $O_t \to C_t$ | +0.234 | [-0.448, 0.827] | +0.202 | [-0.079, 0.414] |

| Test start | LR $C$ | GRU learning rate | Weight decay | Epochs |
|---|---|---|---|---|
| 2022-01-01 | 0.1 | 0.001 | 0.0001 | 16 |
| 2022-07-01 | 10 | 0.0003 | 0.0001 | 16 |
| 2023-01-01 | 10 | 0.001 | 0.0001 | 32 |
| 2023-07-01 | 0.01 | 0.001 | 0.0001 | 32 |
| 2024-01-01 | 10 | 0.001 | 0.00001 | 16 |
| 2024-07-01 | 100 | 0.001 | 0.0001 | 16 |

These configurations were recovered from the historical model-comparison validation fits with joint identity/reaction background. For each learner and period, selection averaged close-to-close validation AP across Local + timing, Dense + timing, and Dense split, and over the two GRU tuning seeds. The selected values stay fixed across all five target windows; final weights use the common-row fitting protocol in Table 5. This retrospective configuration-transfer check is separate from both target-specific shared selection and the historical per-cell choices in Appendix B.5.

The large negative pooled interaction persists on the primary close-to-close target; the corresponding subsequent-window interactions are smaller and uncertain. Other future contrasts retain positive intervals: the $C_t \to C_{t+3}$ split pointwise interval and the fixed-configuration $C_t \to C_{t+5}$ pooled interval. Splitting changes the individual learners' exposure gaps in opposite directions: on the primary close-to-close target, GRU's mean local-minus-dense gap rises from 0.15 to 0.82 pp, while LR's falls from 2.83 to 0.03 pp. The cross-learner interaction and within-learner exposure gaps answer different questions.

**Supplementary paired comparisons.** The following analyses use saved full-text FinBERT predictions from the Model-gap study, with joint company background, six periods, and all five final seeds for each neural learner; LR is deterministic. Content margins subtract the same learner's no-content AP under matched configurations, retaining metadata and numerical history. The no-content reference inherits the configuration selected across text conditions. Interactions use Eq. 3 for each stated model pair, with Local and Dense pooled + timing inputs.

The comparisons were specified after inspecting these retrospective results. We reuse the 20,000 linked-firm draws, synchronizing firm multiplicities across periods, targets, learners, and conditions. Tables S1 and S2 report both pointwise and Bonferroni-adjusted percentile intervals: 30 content margins form one supplementary family, and 12 interactions form another. The primary six-comparison family in Table 10 is unchanged. All intervals condition on saved fitted predictions and selected configurations; they exclude full-workflow reselection and refitting uncertainty.

**Content margins depend on the learner and input arrangement.** On $C_{t-1} \to C_t$, LR's Local and Dense fresh/stale margins are $+2.97$ and $+2.94$ pp, with adjusted intervals above zero; its Dense pooled margin is $+0.15$ pp with an interval crossing zero. GRU's three mean margins are negative. Its Dense pooled and split pointwise intervals lie below zero, while their 30-comparison-adjusted intervals cross zero. All 24 adjusted intervals on the other four targets also cross zero (Table S1). Absolute scores appear in Table 8.

Table S1: **All text-minus-no-content margins.** Model-gap study, full-text FinBERT with joint company background. Estimates are AP pp; adjusted intervals account for all 30 comparisons in this supplementary family. Each reference retains metadata and history and inherits its learner's selected configuration.

| Comparison | Estimate (pp) | Pointwise 95% CI | Adjusted 95% CI |
|---|---|---|---|
| $C_{t-1} \to C_t$ | | | |
| LR / Local + timing | 2.97 | $[1.71, 4.26]$ | $[1.09, 5.21]$ |
| LR / Dense + timing | 0.15 | $[-0.14, 0.54]$ | $[-0.34, 0.77]$ |
| LR / Dense fresh/stale | 2.94 | $[1.65, 4.28]$ | $[0.99, 5.23]$ |
| GRU / Local + timing | $-0.93$ | $[-2.00, 0.12]$ | $[-2.70, 0.74]$ |
| GRU / Dense + timing | $-1.08$ | $[-1.91, -0.31]$ | $[-2.48, 0.08]$ |
| GRU / Dense fresh/stale | $-1.75$ | $[-3.00, -0.49]$ | $[-3.85, 0.19]$ |
| $C_t \to C_{t+1}$ | | | |
| LR / Local + timing | $-0.05$ | $[-0.14, 0.03]$ | $[-0.22, 0.08]$ |
| LR / Dense + timing | 0.03 | $[-0.07, 0.14]$ | $[-0.20, 0.24]$ |
| LR / Dense fresh/stale | $-0.03$ | $[-0.15, 0.06]$ | $[-0.29, 0.14]$ |
| GRU / Local + timing | $-0.13$ | $[-0.39, 0.02]$ | $[-0.55, 0.12]$ |
| GRU / Dense + timing | $-0.26$ | $[-0.55, -0.07]$ | $[-0.72, 0.05]$ |
| GRU / Dense fresh/stale | $-0.30$ | $[-0.67, -0.09]$ | $[-0.93, 0.04]$ |
| $C_t \to C_{t+3}$ | | | |
| LR / Local + timing | 0.00 | $[-0.11, 0.07]$ | $[-0.23, 0.13]$ |
| LR / Dense + timing | 0.00 | $[-0.19, 0.28]$ | $[-0.34, 0.49]$ |
| LR / Dense fresh/stale | $-0.01$ | $[-0.22, 0.20]$ | $[-0.38, 0.37]$ |
| GRU / Local + timing | $-0.08$ | $[-0.30, 0.08]$ | $[-0.49, 0.21]$ |
| GRU / Dense + timing | $-0.29$ | $[-0.66, 0.06]$ | $[-0.91, 0.28]$ |
| GRU / Dense fresh/stale | $-0.46$ | $[-0.89, -0.15]$ | $[-1.19, 0.09]$ |
| $C_t \to C_{t+5}$ | | | |
| LR / Local + timing | 0.00 | $[-0.11, 0.05]$ | $[-0.26, 0.10]$ |
| LR / Dense + timing | $-0.10$ | $[-0.31, 0.23]$ | $[-0.46, 0.44]$ |
| LR / Dense fresh/stale | $-0.09$ | $[-0.30, 0.22]$ | $[-0.49, 0.41]$ |
| GRU / Local + timing | 0.01 | $[-0.21, 0.22]$ | $[-0.39, 0.36]$ |
| GRU / Dense + timing | $-0.30$ | $[-0.76, 0.06]$ | $[-1.09, 0.28]$ |
| GRU / Dense fresh/stale | $-0.18$ | $[-0.58, 0.12]$ | $[-0.83, 0.32]$ |
| $O_t \to C_t$ | | | |
| LR / Local + timing | 0.25 | $[-0.09, 0.69]$ | $[-0.28, 0.98]$ |
| LR / Dense + timing | $-0.03$ | $[-0.13, 0.10]$ | $[-0.21, 0.19]$ |
| LR / Dense fresh/stale | 0.21 | $[-0.09, 0.62]$ | $[-0.25, 0.87]$ |
| GRU / Local + timing | $-0.05$ | $[-0.61, 0.55]$ | $[-0.99, 1.03]$ |
| GRU / Dense + timing | 0.02 | $[-0.34, 0.31]$ | $[-0.65, 0.49]$ |
| GRU / Dense fresh/stale | $-0.13$ | $[-0.75, 0.52]$ | $[-1.22, 0.93]$ |

**Interactions extend to other comparisons with LR.** On $C_{t-1} \to C_t$, the GRU–LR, TCN–LR, and Transformer–LR interactions are $-2.68$, $-2.32$, and $-2.13$ pp, with all twelve-comparison-adjusted intervals below zero. The three neural–neural interactions have intervals crossing zero, as do all six pairs on $C_t \to C_{t+1}$ (Table S2). Thus the observed interaction extends beyond GRU–LR, primarily to comparisons between LR and the tested neural learners on the close-to-close target.

Table S2: **All model-pair interactions on the two common targets.** Full-text FinBERT, joint company background, and pooled + timing inputs. Estimates are Local-minus-Dense changes in the stated learner gap, in AP pp. Adjusted intervals account for all 12 comparisons in this supplementary family.

| Comparison | Estimate (pp) | Pointwise 95% CI | Adjusted 95% CI |
|---|---|---|---|
| $C_{t-1} \to C_t$ | | | |
| GRU–LR | $-2.68$ | $[-4.15, -1.11]$ | $[-4.97, -0.40]$ |
| TCN–LR | $-2.32$ | $[-3.73, -1.00]$ | $[-4.36, -0.43]$ |
| Transformer–LR | $-2.13$ | $[-3.70, -0.70]$ | $[-4.51, -0.08]$ |
| GRU–TCN | $-0.36$ | $[-1.06, 0.53]$ | $[-1.46, 0.95]$ |
| GRU–Transformer | $-0.54$ | $[-1.12, 0.20]$ | $[-1.45, 0.47]$ |
| TCN–Transformer | $-0.18$ | $[-1.04, 0.70]$ | $[-1.47, 1.14]$ |
| $C_t \to C_{t+1}$ | | | |
| GRU–LR | $0.21$ | $[-0.08, 0.55]$ | $[-0.23, 0.73]$ |
| TCN–LR | $0.11$ | $[-0.17, 0.36]$ | $[-0.30, 0.47]$ |
| Transformer–LR | $0.12$ | $[-0.08, 0.39]$ | $[-0.19, 0.53]$ |
| GRU–TCN | $0.10$ | $[-0.18, 0.46]$ | $[-0.32, 0.67]$ |
| GRU–Transformer | $0.08$ | $[-0.26, 0.42]$ | $[-0.46, 0.60]$ |
| TCN–Transformer | $-0.01$ | $[-0.27, 0.18]$ | $[-0.39, 0.28]$ |

### A.4 Background, Capacity, and Historical Content

For background $b$, denote no-content, Constant, and Event AP by $P_s^0(b)$, $P_s^c(b)$, and $P_s^e(b)$; $b_0$ is the base background. These variants share the stated Dense exposure and differ in content access or output routing. Define

$$H_s(b) = P_s^0(b) - P_s^0(b_0), \qquad V_s(b) = P_s^e(b) - P_s^0(b),$$
$$R_s(b) = P_s^e(b) - P_s^c(b), \qquad J_s(b) = V_s(b) - V_s(b_0). \tag{4}$$

$H$ compares no-content AP under $b$ with the base background. $V$ compares Event text with no-content under the same $b$. $R$ compares Event with Constant routing under $b$. $J$ is the change in $V$ from $b_0$ to $b$. Each margin depends on its stated reference predictor; $R$ and $V$ can change differently.

Tables 13–14 retain all four backgrounds and both predeclared targets. The ordinary no-content head and equal-parameter numerical experts are distinct references. Let $V_{cap}$ be text event allocation minus numerical event allocation. Current–past contrasts use the donor-matched content arms, not unmatched text predictions.

Table 13: Background study, AP (%). Constant/Event are text experts; Num. denotes equal-parameter numerical experts. Base uses numerical history without company augmentation.

| Target | Background | No content | Constant | Event | Num. const. | Num. event |
|---|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | Base | 13.82 | 13.64 | 15.95 | 14.09 | 14.80 |
| $C_{t-1} \to C_t$ | Identity | 15.34 | 15.29 | 16.78 | 15.80 | 15.99 |
| $C_{t-1} \to C_t$ | Reaction | 16.33 | 16.08 | 16.81 | 16.35 | 16.47 |
| $C_{t-1} \to C_t$ | Both | 17.15 | 17.12 | 17.51 | 17.42 | 17.49 |
| $C_t \to C_{t+1}$ | Base | 7.14 | 7.33 | 7.33 | 7.17 | 7.21 |
| $C_t \to C_{t+1}$ | Identity | 7.76 | 7.88 | 7.93 | 7.88 | 7.87 |
| $C_t \to C_{t+1}$ | Reaction | 7.32 | 7.47 | 7.44 | 7.34 | 7.30 |
| $C_t \to C_{t+1}$ | Both | 7.80 | 7.88 | 7.93 | 7.81 | 7.85 |

Table 14: Historical-content control, AP (%). Current and past arms share recipient metadata, routing, and the mask for records without an eligible previous donor.

| Target | Background | Current const. | Current event | Past const. | Past event |
|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | Base | 13.50 | 16.00 | 13.33 | 15.65 |
| $C_{t-1} \to C_t$ | Identity | 15.41 | 16.89 | 15.18 | 16.65 |
| $C_{t-1} \to C_t$ | Reaction | 16.24 | 16.92 | 16.11 | 16.60 |
| $C_{t-1} \to C_t$ | Both | 17.00 | 17.51 | 16.87 | 17.27 |
| $C_t \to C_{t+1}$ | Base | 7.31 | 7.30 | 7.25 | 7.25 |
| $C_t \to C_{t+1}$ | Identity | 7.91 | 7.91 | 7.86 | 7.82 |
| $C_t \to C_{t+1}$ | Reaction | 7.46 | 7.41 | 7.34 | 7.37 |
| $C_t \to C_{t+1}$ | Both | 7.90 | 7.90 | 7.81 | 7.82 |

Table 15: $C_{t-1} \to C_t$ background contrasts, AP pp with pointwise linked-firm intervals. $V$: text event minus no content; $R$: text event minus text constant; $V_{cap}$: text event minus numerical event; Current–past: matched content under event routing; Numeric routing: numerical event minus numerical constant.

| Background | Contrast | AP difference | Pointwise interval |
|---|---|---|---|
| Base | $V$ | +2.128 | [0.952, 3.524] |
| Base | $R$ | +2.312 | [1.099, 3.696] |
| Base | $V_{cap}$ | +1.143 | [-0.439, 2.849] |
| Base | Current–past | +0.355 | [-0.167, 0.966] |
| Base | Numeric routing | +0.711 | [-0.153, 1.704] |
| Identity | $V$ | +1.448 | [0.395, 2.608] |
| Identity | $R$ | +1.493 | [0.574, 2.518] |
| Identity | $V_{cap}$ | +0.793 | [-0.653, 2.344] |
| Identity | Current–past | +0.240 | [-0.261, 0.853] |
| Identity | Numeric routing | +0.188 | [-0.596, 1.104] |
| Reaction | $V$ | +0.485 | [-0.469, 1.556] |
| Reaction | $R$ | +0.729 | [-0.152, 1.719] |
| Reaction | $V_{cap}$ | +0.342 | [-0.877, 1.643] |
| Reaction | Current–past | +0.321 | [-0.135, 0.770] |
| Reaction | Numeric routing | +0.125 | [-0.609, 0.891] |
| Both | $V$ | +0.363 | [-0.491, 1.212] |
| Both | $R$ | +0.393 | [-0.303, 1.142] |
| Both | $V_{cap}$ | +0.021 | [-1.080, 1.212] |
| Both | Current–past | +0.246 | [-0.169, 0.702] |
| Both | Numeric routing | +0.068 | [-0.601, 0.681] |

Table 16: $C_t \to C_{t+1}$ background contrasts, AP pp, with the same definitions and pointwise intervals as Table 15. All backgrounds and reference contrasts are retained.

| Background | Contrast | AP difference | Pointwise interval |
|---|---|---|---|
| Base | $V$ | +0.186 | [0.038, 0.379] |
| Base | $R$ | -0.001 | [-0.121, 0.157] |
| Base | $V_{cap}$ | +0.121 | [-0.085, 0.374] |
| Base | Current–past | +0.046 | [-0.031, 0.172] |
| Base | Numeric routing | +0.043 | [-0.076, 0.176] |
| Identity | $V$ | +0.166 | [-0.015, 0.351] |
| Identity | $R$ | +0.054 | [-0.088, 0.200] |
| Identity | $V_{cap}$ | +0.059 | [-0.169, 0.279] |
| Identity | Current–past | +0.087 | [-0.002, 0.222] |
| Identity | Numeric routing | -0.010 | [-0.133, 0.109] |
| Reaction | $V$ | +0.113 | [-0.011, 0.302] |
| Reaction | $R$ | -0.030 | [-0.153, 0.156] |
| Reaction | $V_{cap}$ | +0.140 | [-0.022, 0.409] |
| Reaction | Current–past | +0.047 | [-0.027, 0.148] |
| Reaction | Numeric routing | -0.048 | [-0.178, 0.067] |
| Both | $V$ | +0.133 | [-0.029, 0.275] |
| Both | $R$ | +0.053 | [-0.082, 0.173] |
| Both | $V_{cap}$ | +0.085 | [-0.121, 0.298] |
| Both | Current–past | +0.078 | [-0.012, 0.227] |
| Both | Numeric routing | +0.042 | [-0.087, 0.196] |

Table 17: All six primary background changes, AP pp. Each subtracts the Base margin from Both and uses six-comparison-adjusted linked-firm intervals. Changes in $R$ are outside this family.

| Target | Margin changed | Both minus Base | Adjusted interval |
|---|---|---|---|
| $C_{t-1} \to C_t$ | $V$ | -1.765 | [-2.932, -0.803] |
| $C_{t-1} \to C_t$ | $V_{cap}$ | -1.122 | [-2.230, 0.009] |
| $C_{t-1} \to C_t$ | Current–past | -0.109 | [-0.636, 0.351] |
| $C_t \to C_{t+1}$ | $V$ | -0.053 | [-0.351, 0.177] |
| $C_t \to C_{t+1}$ | $V_{cap}$ | -0.035 | [-0.348, 0.300] |
| $C_t \to C_{t+1}$ | Current–past | +0.032 | [-0.127, 0.178] |

On the primary close-to-close target, the ordinary baseline improves more than the text model. Margins over equal-parameter numerical experts and historical content have intervals crossing zero, leaving the semantic source of the contraction unresolved. The adjusted change in $V_{cap}$ is sensitive to resampling: its linked-firm interval crosses zero narrowly, whereas date-block intervals are negative (Table 18).

For $C_t \to C_{t+1}$, Base has a small positive $V$ of 0.186 pp with pointwise interval [0.038, 0.379]; its margin over the equal-parameter numerical reference is uncertain. The Base-to-Both change is -0.053 pp [-0.351, 0.177], adjusted. This return-window result therefore supports a small margin over the ordinary baseline, with uncertainty about its contraction under company information.

Table 18: Conditional resampling sensitivity for core main-study contrasts. All entries are six-comparison-adjusted intervals in AP pp. Linked-firm and date-block schemes provide alternative conditional uncertainty analyses for different dependence structures.

| Contrast | Linked firm | 5-date blocks | 20-date blocks |
|---|---|---|---|
| $C_{t-1} \to C_t$: pooled $I$ | [-4.740, -0.583] | [-4.640, -0.759] | [-4.437, -0.559] |
| $C_t \to C_{t+1}$: pooled $I$ | [-0.196, 0.681] | [-0.299, 0.592] | [-0.146, 0.543] |
| $C_{t-1} \to C_t$: change in $V$ | [-2.932, -0.803] | [-2.873, -0.787] | [-3.057, -0.814] |
| $C_{t-1} \to C_t$: change in $V_{cap}$ | [-2.230, 0.009] | [-2.125, -0.271] | [-2.119, -0.318] |
| $C_t \to C_{t+1}$: change in $V$ | [-0.351, 0.177] | [-0.367, 0.210] | [-0.373, 0.196] |

### A.5 Coverage Audit: Aggregation and Matched Coverage

The Coverage audit uses a separate supporting protocol with the close-to-close target, six 2022–2024 periods, full-text FinBERT/SVD24, joint company information, and matched $20 \times 18$ numerical histories. One configuration per learner and period is selected by averaging validation AP over local timing, dense timing, and dense split. GRU uses seeds 419, 631, 853, 1061, and 1297; LR is deterministic. The audit retains its historical record collection, labels, and condition-specific training-only scaling.

Mean aggregation divides only the text coordinates by the active-record count (zero if none). Counts, lengths, clocks, and company context remain unchanged. The coverage control retains all prediction rows and metadata, but masks dense content to match local content coverage separately within each date and partition. Eligible own-firm rows are chosen without outcomes using fixed partition seeds 11, 23, and 37. AP is averaged over training seeds, coverage seeds where applicable, and periods. Two thousand paired linked-firm resamples give exploratory pointwise intervals conditional on the fitted models; full-refitting uncertainty is not included.

The saved input-count audit shows that every Local row and every Dense test row has at most one active record. Dense has four overlapping rows in the first $T$ partition and seven in each later $T$ partition; $V_h$ has none, and only the first $V_s$ has three. These counts refer to overlapping expanding prefixes, not disjoint events. Thus sum and mean give identical raw test content. They alter a few fitting rows, the condition-specific training scaler, and potentially all transformed rows and fitted predictions. The mean variant inherits the shared sum-selected configuration. The aggregation sensitivity is therefore limited to rare overlap in fitting data and its downstream effects.

Table 19: **Historical Coverage audit: paired sensitivity checks.** Full-text LR/GRU, AP pp. The final column directly compares each variant with the sum baseline. Sum and mean coincide on raw test content; their fitted transforms and models can differ.

| Supporting audit | $I$ (pp), 95% CI | Change from sum |
|---|---|---|
| Shared configuration, sum | -3.73 [-5.14, -2.21] | – |
| Text mean, same metadata | -3.63 [-5.08, -2.10] | +0.10 [-0.07, +0.21] |
| Date-matched content coverage | -3.51 [-4.75, -2.18] | +0.22 [-0.51, +0.83] |

Both variants retain a negative interaction, and the intervals for their changes include zero. The interaction persists after equalizing content-exposed row counts. Firm identity, record age, and

content remain unmatched, so the coverage control leaves an age mechanism unresolved. Rare overlap limits the sum/mean comparison's identifying power.

### A.6 Period, Seed, and Row Localization

Tables 20–21 use the full-text Model-gap study, not the Coverage audit. Omitting a period recomputes the five-period mean from the existing fitted contrasts. The negative interaction remains after every omission, although the 2024-January period contributes substantially. Seed-specific means reuse the same deterministic LR reference and selected configurations. They assess final-training variation on the same datasets; repeated hyperparameter-search variability is not included.

Table 20: Close-to-close interaction by test start and after omitting that start. AP pp, descriptive fitted-model means. The six periods do not all have the same direction.

| Omitted / evaluated start | Period $I$ (pp) | Other five periods $I$ (pp) |
|---|---|---|
| 2022-01-01 | -0.74 | -3.06 |
| 2022-07-01 | -0.12 | -3.19 |
| 2023-01-01 | -2.56 | -2.70 |
| 2023-07-01 | -3.53 | -2.50 |
| 2024-01-01 | -9.24 | -1.36 |
| 2024-07-01 | +0.15 | -3.24 |

Table 21: Interaction for each final GRU seed, averaged over the six periods. Same deterministic LR comparator. $C_t \to C_{t+1}$ seed means are small, while the adjusted interval in Table 10 remains uncertain.

| GRU seed | $C_{t-1} \to C_t$ $I$ (pp) | $C_t \to C_{t+1}$ $I$ (pp) |
|---|---|---|
| 419 | -2.00 | +0.14 |
| 631 | -1.83 | +0.17 |
| 853 | -3.30 | +0.20 |
| 1061 | -3.52 | +0.24 |
| 1297 | -2.73 | +0.28 |

For localization, prediction rows are grouped by the common 30-day event clock: fresh, stale only, or no active record. We partition each fitted model's global AP into contributions from the positive examples in these groups, retaining the global score ranking and precision at tied score thresholds. These contribution terms sum exactly to global AP and should be read separately from subgroup AP. Subtracting the contributions across inputs and learners gives Table 22.

Table 22: Descriptive localization of the close-to-close interaction, AP pp. Row shares average period-specific shares. Contributions depend on the global ranking, making the decomposition descriptive rather than causal. Rounding can affect displayed sums.

| Prediction rows | Rows (%) | LR $E$ | GRU $E$ | $I$ contribution |
|---|---|---|---|---|
| Fresh | 1.74 | +2.55 | -0.14 | -2.69 |
| Stale only | 38.85 | +0.11 | +0.23 | +0.11 |
| No active record | 59.42 | +0.16 | +0.07 | -0.09 |

Fresh-record rows occupy only 1.74% of evaluated samples but account for -2.69 pp of the -2.68 pp interaction. This is a descriptive decomposition of global AP, with no localized uncertainty estimate or mechanism inferred from it. Retraining and global ranking can change scores even on rows with no active content.

### A.7 All Predeclared Reuse Durations

This sweep varies input duration while holding the return horizon fixed. Full-text LR/GRU reuse each target's shared configuration across all durations. The 2/3/7/14/60-day cells extend the one-day

and 30-day exposure comparison. Tables 23–24 report all point estimates; paired duration intervals have not been computed. The sweep is non-monotonic and serves as sensitivity analysis, with reuse duration left unoptimized on test performance.

Table 23: All pooled-with-timing reuse durations, mean AP (%). Column numbers are trading days of exposure. Rows compare the $C_{t-1} \to C_t$ and $C_t \to C_{t+1}$ return windows.

| Target | Learner | 1 | 2 | 3 | 7 | 14 | 30 | 60 |
|---|---|---|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | LR | 16.83 | 15.51 | 14.76 | 14.35 | 14.16 | 14.00 | 14.36 |
| $C_{t-1} \to C_t$ | GRU | 17.91 | 18.10 | 18.09 | 17.82 | 17.80 | 17.76 | 18.51 |
| $C_t \to C_{t+1}$ | LR | 7.93 | 7.98 | 7.97 | 8.17 | 8.11 | 8.01 | 10.72 |
| $C_t \to C_{t+1}$ | GRU | 7.72 | 7.75 | 7.75 | 7.78 | 7.62 | 7.59 | 10.22 |

Table 24: Corresponding fresh/stale split duration sweep, mean AP (%). One-day split and local timing inputs coincide. Configurations are fixed across durations, and no duration-specific intervals are reported.

| Target | Learner | 1 | 2 | 3 | 7 | 14 | 30 | 60 |
|---|---|---|---|---|---|---|---|---|
| $C_{t-1} \to C_t$ | LR | 16.83 | 16.70 | 16.66 | 16.84 | 16.86 | 16.79 | 17.22 |
| $C_{t-1} \to C_t$ | GRU | 17.91 | 17.17 | 17.29 | 17.28 | 17.42 | 17.09 | 18.25 |
| $C_t \to C_{t+1}$ | LR | 7.93 | 7.94 | 7.98 | 8.10 | 8.04 | 7.95 | 10.65 |
| $C_t \to C_{t+1}$ | GRU | 7.72 | 7.68 | 7.56 | 7.64 | 7.50 | 7.55 | 10.07 |

An input audit checks all six saved full-text caches: event assignments remain on the same firm and successive trading rows at or after assumed eligibility; one-day and 30-day aggregates reproduce the saved local and dense tensors. At 60 days, mean test coverage reaches 92.02%, versus 40.58% at 30 days, with overlapping records on 0.54% of rows (Table 25). Common metadata and the 30-day event clock stay fixed throughout the sweep; 51.44% of rows then receive 60-day content despite no active record under that clock. The content-duration check thus retains a 30-day metadata definition even at 60-day exposure. The source of the higher future-target scores at 60 days remains unresolved.

Table 25: Saved input support across reuse durations. Equal-period averages on the common test rows. Overlap means more than one active same-firm record; the clock is always defined over 30 days.

| Days | Coverage (%) | Overlap (%) | Mean records | Outside 30-day clock (%) |
|---|---|---|---|---|
| 1 | 1.74 | 0.00 | 0.017 | 0.00 |
| 2 | 3.36 | 0.00 | 0.034 | 0.00 |
| 3 | 4.87 | 0.00 | 0.049 | 0.00 |
| 7 | 10.46 | 0.00 | 0.105 | 0.00 |
| 14 | 19.43 | 0.00 | 0.194 | 0.00 |
| 30 | 40.58 | 0.00 | 0.406 | 0.00 |
| 60 | 92.02 | 0.54 | 0.926 | 51.44 |

The 2019–2021 ten-learner Breadth study addresses the specific one-day versus 30-day comparison; this duration sweep uses the Model-gap protocol.

### A.8 Overnight-Window Follow-up

The primary-target findings motivated three additional target comparisons in the Model-gap study. These follow-ups retain the complete daily panel for the same 57 firms and six 2022–2024 test periods. Table 26 defines the absolute-return outcomes using adjusted open ($O$) and close ($C$) prices. Each outcome is classified with its firm-specific 95th-percentile threshold fitted on the common training partition. For first-eligible call rows, the initial overnight interval runs from the preceding close to the eligible day's open; the evaluation also retains all other daily rows.

**Inputs, rows, and fitting.** The three follow-up targets reuse exactly the common five-target $T, V_h, V_s, T_e$ masks. Full-text FinBERT, the $20 \times 18$ previous-close numerical history, firm identity, and completed past-call reactions remain fixed. We select one configuration per learner, target, and period by averaging $V_h$ AP across Local + timing, Dense + timing, and Dense split. Neural fits use two tuning and five final seeds; LR is deterministic. Models for these windows are refitted on $T + V_h + V_s$, with thresholds and transformations still fitted on $T$. The primary close-to-close and open-to-close comparisons reuse their existing predictions. Scores average seeds and then periods.

**Direct comparisons and inference.** Let $I(w)$ denote the learner-gap interaction in Eq. (3) for return window $w$. The matched-endpoint contrasts are

$$\Delta I_0 = I(C_{t-1} \to C_t) - I(O_t \to C_t),$$
$$\Delta I_1 = I(C_{t-1} \to C_{t+1}) - I(O_t \to C_{t+1}).$$

For each endpoint, the first target includes the initial overnight interval and the second starts at the open. The comparisons use the same 20,000 linked-firm bootstrap draws across periods and targets. The follow-up family contains three checks: overnight-only $I$, $\Delta I_0$, and $\Delta I_1$, with Bonferroni-adjusted intervals for that family. The main six-comparison family is unchanged. A subsequent exploratory contrast, overnight-only minus intraday-only $I$, uses a pointwise interval and is reported separately. All intervals condition on fitted predictions.

Table 26: Overnight-window follow-up in the Model-gap study, AP pp. (A) Window-specific interactions, with pointwise 95% intervals for pooled + timing and fresh/stale split. (B) The complete three-check follow-up family and a separately labeled exploratory contrast. $I$ is the Local GRU–LR gap minus the Dense gap. Existing five-target results remain in Table 10.

| **A. Return windows: estimate [pointwise 95% CI]** | | |
|---|---|---|
| Target and absolute return | Pooled + timing | Fresh/stale split |
| Overnight only $\lvert O_t/C_{t-1} - 1\rvert$ | $-4.61$ [$-6.34$, $-2.78$] | $-0.19$ [$-0.54$, $+0.24$] |
| Prior close to next close $\lvert C_{t+1}/C_{t-1} - 1\rvert$ | $-1.60$ [$-3.17$, $-0.06$] | $+0.03$ [$-0.39$, $+0.50$] |
| Open to next close $\lvert C_{t+1}/O_t - 1\rvert$ | $-0.07$ [$-0.54$, $+0.39$] | $+0.15$ [$-0.09$, $+0.38$] |
| **B. Direct checks and inference scope** | | |
| Contrast | Estimate | 95% CI |
| *Three-comparison-adjusted follow-up family* | | |
| Overnight-only $I$ | $-4.61$ | [$-6.79$, $-2.40$] |
| Same-day endpoint $\Delta I_0$ | $-2.33$ | [$-4.37$, $-0.33$] |
| Next-day endpoint $\Delta I_1$ | $-1.53$ | [$-3.73$, $+0.39$] |
| *Exploratory, pointwise; outside the three-check family* | | |
| Overnight-only minus intraday-only $I$ | $-4.27$ | [$-6.09$, $-2.43$] |

**Interpretation.** The same-day endpoint comparison links the larger primary interaction to inclusion of the initial overnight interval. The overnight-only target also retains a large negative pooled interaction. At the next-close endpoint, the direct interval crosses zero, leaving the extended-window contrast uncertain. These are comparisons between target-specific classification tasks and fitted models; they provide a return-window explanation for the pattern, with its market mechanism unresolved. The comparisons retain the retrospective scope in Section 3 and the transcript-proxy assumption detailed in Appendix B.4.

# B Data, Training, and Resampling Details

## B.1 Prediction Clock and Label Construction

For daily row $t$, numerical observations end at $C_{t-1}$; the primary label uses $|C_t/C_{t-1} - 1|$. Local exposure assigns a call to the next trading date after its stored date, with transcription assumed available by the open; Dense extends exposure for 30 trading days. The primary interval therefore

starts before the assumed text cutoff. Source timezone and synchronized intraday timestamps are unverified, and the adjusted-price snapshot has no historical point-in-time vintages (Appendix B.4). The firm-specific 95th percentile uses the whole pre-test prefix in the Breadth study, initial $T$ before complete-sequence filtering in the supporting Encoding and Routing/age studies, and common valid $T$ rows after intersection and purging in the Model-gap and Background studies. The separate LR/HGB history study refits thresholds and transforms on its outer pre-test prefix after inner validation. Test prevalence may differ from 5%. Table 4 gives all return-window definitions.

### B.2 Input Dictionary and Study Contracts

The base vector contains nine shifted return features and nine shifted log-volume features. The sequence for row $t$ stacks the engineered vectors at $t-19,\ldots,t$; because each vector is already shifted, its latest numerical measurement is at $t-1$. The sequence-based comparisons require complete finite histories. Each feature's mean and population standard deviation are pooled over all eligible $T$ sequence positions and shared across the 20 positions; standard deviations below $10^{-6}$ are replaced by 1, and standardized values are clipped to $[-12, 12]$. Table 27 gives group definitions; the accompanying feature dictionary specifies each coordinate.

The two-expert common context includes eight event metadata fields plus ten deterministic random-route indicators in every condition, including no-content. The augmented LR/GRU grid instead uses the eight metadata fields without those random indicators. No-content retains numerical history, the common context, and whichever identity/reaction group is supplied; it removes only the text experts' contribution. Table 28 makes this distinction explicit. The reaction group includes availability, count, and elapsed time alongside prior return values; all reactions must have completed before the prediction date.

### B.3 Data Versions and Observed Partitions

Table 29 records the available provenance, including unresolved source information. The initial candidate list has 60 symbols and the frozen study panel retains 57; absence of AMGN, BAC, and UPS is verifiable, but exact per-firm exclusion reasons are not established by the retained manifests. Event deduplication and text-embedding deduplication are separate operations. Full hashes and manifest paths are in the companion `data_provenance.json`; the local snapshot does not establish upstream redistribution permission. Table 30 is exported from the cached 2022–2024 matched-history partitions. H1/H2 identify January/July starts, not complete half-year test windows. $V_h$ and $V_s$ each contain 63 panel trading dates; $T_e$ ends 126 calendar days after the nominal start. Local/Dense coverage is the percentage of retained rows with at least one assigned record. The same rows and labels are shared within each comparison. These retrospective periods use a later-acquired adjusted-price snapshot.

### B.4 Source Verification and Artifact Release Plan

The cached 1,824,805,388-byte parquet matches the upstream large-file SHA256 at dataset revision `71db6d3f0e4b9e82980dc3384f2f9ff4d71cf6be` (Kurry, 2025); the provenance manifest retains the full checksum. This establishes object identity, not acquisition time or historical availability. The source defines `date` as call date and `year` as fiscal year. Across both financial caches there are 1,756 company-date records, including 22 Breadth-only records with calendar year 2016 and fiscal year 2017. A purposive 12-record web check found nine explicit publication times, one unresolved page, one company mismatch, and one conflicting date. The two anomalies are quarantined in the main studies; this check gives no population error rate.

**Public access and the transcript proxy.** Official announcements document public access to Apple's 27 January 2022 call, Microsoft's 25 July 2023 call, and JPMorgan Chase's 13 April 2022 call. They motivate assuming a completed call and usable transcription by the next trading open. Archived transcripts serve as proxies for that content, rather than outputs from a live transcription system. Corpus-wide public access, historical full-text publication times, and correspondence to contemporaneous transcriptions remain unverified. Eligibility follows stored call dates.

Table 27: Input fields and timing. $T$ denotes the initial fitting partition. Architecture and two-expert studies use protocol-specific text scaling.

| Group | Width | Definition | Trace |
|---|---|---|---|
| H18 | 18 per step | For r=Close/previous Close-1 and v=log(1+Volume): lags 1,2,3,5,10; mean and sample std over previous 5 and 20 trading steps (min 3). | H18 |
| Sequence | 20 x 18 | Rows t-19,...,t of the already shifted fields; latest numerical measurement t-1. Require all values finite; T per-feature scaling pooled across positions; clip [-12,12]. | SEQ, SCALE |
| Clock | 4 | Event-day flag; log(1+latest age in trading steps); event count at ages 0–29; no-prior-event flag. Raw no-history coding (0,0,0,1). | CLOCK |
| Common metadata | 8 total | Clock4 + fresh count, fresh sum log(1+stored chars), stale count, stale sum log(1+stored chars). Raw absent counts/lengths 0; T affine scaling. | CONTEXT |
| Reaction | 8 separate | Prior-reaction availability; log count; latest absolute return; mean2; mean4; sample std4; max4; log calendar days since prior outcome. Raw absent=0, std with fewer than 2 reactions=0. | REACTION |
| Identity | 58 | 57 training-symbol one-hot coordinates plus unseen bucket. Optional groups masked after the common transform. | CONTEXT |
| Random routes | 10 separate | Ten deterministic date-matched binary indicators, retained by routing, background, and learned-age models, including no-content. Excluded from the model-comparison inputs. | CONTEXT |
| Content | 24 | T-fitted SVD; two-expert studies divide by unique T-event coordinate std without centering. P sums ages 0–29; F sums age0. Empty exposure=0. | ROUTE |

Table 28: **Historical protocol: input contracts.** Study names identify the intervention. Figures 3 and 5 use the separate contracts in Table 5.

| Protocol | Input structure | Retained / changed information |
|---|---|---|
| Shared routing | 20x18; common18 | Common8 plus all ten random indicators. No-content removes the content experts; it retains event metadata. |
| Learned age | 20x18; common18 | Same common18 as shared routing in all six age/route cells; no separately fitted no-content cell. |
| Background | 20x18; padded context84 | Common18 + optional identity58 and reaction8. The no-content fit retains the same context mask as its matched content model. |
| Model comparison | 20x18; record122 | 48 content + common8 + identity/reaction66, masked for Base. No random10. Compares timing/split without a no-content cell. |
| Two-expert routing | Same pooled P in both experts | Output mixture q[f1(P)-f1(0)] + (1-q)[f2(P)-f2(0)]. Background and learned-age studies fix Dense ages 0–29; learned-age cells may decay weights. q changes output allocation, not exposure. |

Table 29: Data provenance and availability limits. Snapshot evidence is distinct from a verified historical publication-time archive.

| Item | Frozen-cache evidence / limit | Trace |
|---|---|---|
| Transcript source | kurry/sp500_earnings_transcripts; local part-0.parquet recorded as 1,824,805,388 bytes. SHA256 matches upstream revision 71db6d3. Acquisition time, historical publication and redistribution rights remain unverified. | DATASET |
| Population | 60 hand-listed candidates; the Breadth-study cache retains 57. Missing candidates: AMGN, BAC, UPS. Exact per-firm exclusion reasons unverified. | POPULATION, FROZEN57 |
| Text collections | Breadth study: 1,114 records; 2022–2024 studies: 1,734. Stored-text cap: 80,000 characters. Full recovery: 1,703 unique text embeddings, 36 recovered over-cap texts; stored length metadata unchanged. | coverage manifests |
| Deduplication | Events: first row per symbol, available date, quarter, year. Embedding cache: text-content hash; repeated-text event records remain separate. | AVAILABLE |
| Prices | Yahoo/yfinance 1.2.0; auto_adjust=true, actions=false; snapshot 2026-08-22. 115,140 rows; 2017-01-03 to 2025-01-14; each of 57 symbols has 2,020 dates. | price manifest |
| Availability | Next trading date strictly after call date. Prediction precedes date-t close; numerical data end at t-1. Intraday publication/timezone and historical vintages unverified. | AVAILABLE, TARGET |

Table 30: **Historical/supporting protocol: observed partitions.** These counts are not the current main-study masks in Table 6. Dates are inclusive; Pos. denotes positive labels, and Local/Dense denote record coverage, in percent.

| Start ID | Part. | Observed dates (inclusive) | $N$ | Pos.% | Local% | Dense% |
|---|---|---|---|---|---|---|
| 2022H1 | T | 2017-02-15 to 2021-07-02 | 62,871 | 5.138 | 1.52 | 45.51 |
| 2022H1 | Vh | 2021-07-06 to 2021-10-01 | 3,591 | 1.532 | 1.56 | 45.98 |
| 2022H1 | Vs | 2021-10-04 to 2021-12-31 | 3,591 | 3.704 | 1.56 | 47.20 |
| 2022H1 | Te | 2022-01-03 to 2022-05-06 | 4,959 | 8.732 | 1.98 | 43.50 |
| 2022H2 | T | 2017-02-15 to 2021-12-29 | 69,939 | 5.107 | 1.52 | 45.66 |
| 2022H2 | Vh | 2021-12-30 to 2022-03-30 | 3,591 | 8.159 | 1.59 | 47.67 |
| 2022H2 | Vs | 2022-03-31 to 2022-06-30 | 3,591 | 12.197 | 1.59 | 47.98 |
| 2022H2 | Te | 2022-07-01 to 2022-11-03 | 5,016 | 8.373 | 1.91 | 42.17 |
| 2023H1 | T | 2017-02-15 to 2022-07-01 | 77,178 | 5.144 | 1.53 | 45.84 |
| 2023H1 | Vh | 2022-07-05 to 2022-09-30 | 3,591 | 5.931 | 1.59 | 47.31 |
| 2023H1 | Vs | 2022-10-03 to 2022-12-30 | 3,591 | 7.909 | 1.59 | 47.70 |
| 2023H1 | Te | 2023-01-03 to 2023-05-05 | 4,902 | 3.590 | 1.98 | 43.80 |
| 2023H2 | T | 2017-02-15 to 2022-12-28 | 84,246 | 5.120 | 1.53 | 46.01 |
| 2023H2 | Vh | 2022-12-29 to 2023-03-30 | 3,591 | 3.509 | 1.59 | 48.06 |
| 2023H2 | Vs | 2023-03-31 to 2023-06-30 | 3,591 | 2.673 | 1.59 | 47.93 |
| 2023H2 | Te | 2023-07-03 to 2023-11-03 | 5,016 | 2.273 | 1.89 | 41.13 |
| 2024H1 | T | 2017-02-15 to 2023-06-30 | 91,428 | 5.091 | 1.54 | 46.17 |
| 2024H1 | Vh | 2023-07-03 to 2023-09-29 | 3,591 | 1.866 | 1.53 | 45.70 |
| 2024H1 | Vs | 2023-10-02 to 2023-12-29 | 3,591 | 2.729 | 1.59 | 47.15 |
| 2024H1 | Te | 2024-01-02 to 2024-05-03 | 4,902 | 2.713 | 2.00 | 43.39 |
| 2024H2 | T | 2017-02-15 to 2023-12-27 | 98,496 | 5.073 | 1.54 | 46.21 |
| 2024H2 | Vh | 2023-12-28 to 2024-03-28 | 3,591 | 2.534 | 1.59 | 48.20 |
| 2024H2 | Vs | 2024-04-01 to 2024-06-28 | 3,591 | 2.840 | 1.50 | 45.86 |
| 2024H2 | Te | 2024-07-01 to 2024-11-01 | 5,016 | 4.047 | 1.69 | 36.96 |

**Reproduction and release.** The aggregate package reconstructs Figures 3 and 5 from saved scores and intervals, with arithmetic/hash checks, plotting code, selected configurations, input/split definitions, and a reference code snapshot. Aggregate-to-figure reproduction has been tested; clean-environment raw-data-to-training reproduction remains unvalidated. The source card lists MIT and research/educational use, subject to underlying source terms. The prepared package excludes raw text, excerpts, embeddings, feature dumps, and sample-level artifacts; public release remains subject to dependency and redistribution checks.

### B.5 Configuration and Selection Index

This subsection indexes the historical selection protocols. The shared-selection rules and cell counts for Figures 3 and 5 are specified separately in Appendix A. Table 31 states which conditions share a selected configuration. Budgets are matched within each learner: the augmented-context study searches five LR regularizers and six GRU configurations. Tuned studies select using validation scores; Breadth study runs use fixed settings. Scaling, label fitting, and SVD remain training-only according to the study-specific rule. Candidate settings and selected configurations are listed below. Exact per-period configurations for the supporting studies are supplied in the configuration supplement. The candidate grid and selection contracts are retained here; the main-study selected values remain in Table 7.

Table 31: **Historical protocol: fitting and selection.** The separate history study refits on its outer training prefix; architecture/allocation studies retain initial-$T$ transforms and labels. Current main-figure rules are in Table 5.

| Study | Candidates / seeds | Selection | Labels / transforms |
|---|---|---|---|
| Encoding | LR: C=.003,.01,.03,.3; neural: four (lr,wd) pairs, 16 epochs | Per learner/encoding/fold, mean Vh AP shared across local/dense; 2 tuning / 5 final neural seeds | Raw-T firm q95 fixed; T transforms fixed through final T+Vh+Vs fit |
| Shared routing | Six neural configurations (four 16-epoch + two 32-epoch); four-config reference retained | Common config across equally weighted routing families; history separately; 2/5 seeds | Inherited fixed T labels/history/SVD; T unique-event scale without centering |
| Background | Same six neural configurations; 2/5 seeds | One common choice across four contexts x two content routes; separate common no-content choice | Inherited T transforms; optional identity/reaction masked after transform; dense content fixed |
| Learned age | Same six neural configurations; 2/5 seeds | Per method/cell Vh AP; decay half-life from the shared-routing decay study, not jointly retuned with gate | T labels/transforms fixed; learned age standardized on T dense-active rows; no-record input zero |
| Model comparison | LR: C=.01,.1,1,10,100 (5); GRU: six configurations; LR seed 0; neural: 2 tuning and 5 final seeds | Per context, exposure, encoding and learner; Vh AP; ties lower config ID | Frozen T labels; T affine content scaling; 8 common metadata features; optional 58 identity and 8 past-reaction features; no 10-indicator random-routing block |
| LR/HGB history | LR C=.1,1,10; HGB leaves=3,7,15 (80 iterations); seed90531296 | One choice per learner/fold minimizing mean validation log loss over 3 history x 3 record cells | 126-calendar-day inner validation; inner labels/transforms for tuning; refit on the outer training prefix before testing |
| 2019–2021 breadth | Eight architecture adaptations + separate GRU and LR; configurations in the source-package file below | Fixed settings, neural seeds29/113/271; distinct from the four- or six-config searches | Firm q95/text/scalers fit on whole pre-test prefix each start; without a T/Vh/Vs split |

Breadth-study configurations are provided in `repro/breadth_settings.json` in the accompanying LaTeX source package.

Table 32: Neural candidate grid. The Encoding study uses the first four candidates; the Routing/age and historical company-information studies use all six. Batch size is 1,024. Two tuning seeds select candidates and five final seeds evaluate neural cells.

| ID | Learning rate | Weight decay | Epochs | Used in |
|---|---|---|---|---|
| 0 | 0.0003 | 1e-05 | 16 | Encoding; routing, background, learned age, model comparison |
| 1 | 0.0003 | 0.0001 | 16 | Encoding; routing, background, learned age, model comparison |
| 2 | 0.001 | 1e-05 | 16 | Encoding; routing, background, learned age, model comparison |
| 3 | 0.001 | 0.0001 | 16 | Encoding; routing, background, learned age, model comparison |
| 4 | 0.001 | 0.0001 | 32 | Routing, background, learned age, model comparison |
| 5 | 0.001 | 0.001 | 32 | Routing, background, learned age, model comparison |

### B.6 A Stored Data Example

Figures 6 and 7 show one AAPL example from the stored data. Daily prices form the dense history; the call provides an occasional text record. Earlier completed reactions illustrate the separate past-reaction information group. The display combines observations before and after the call for illustration; experimental inputs follow the lag and eligibility rules in Appendix B.1.

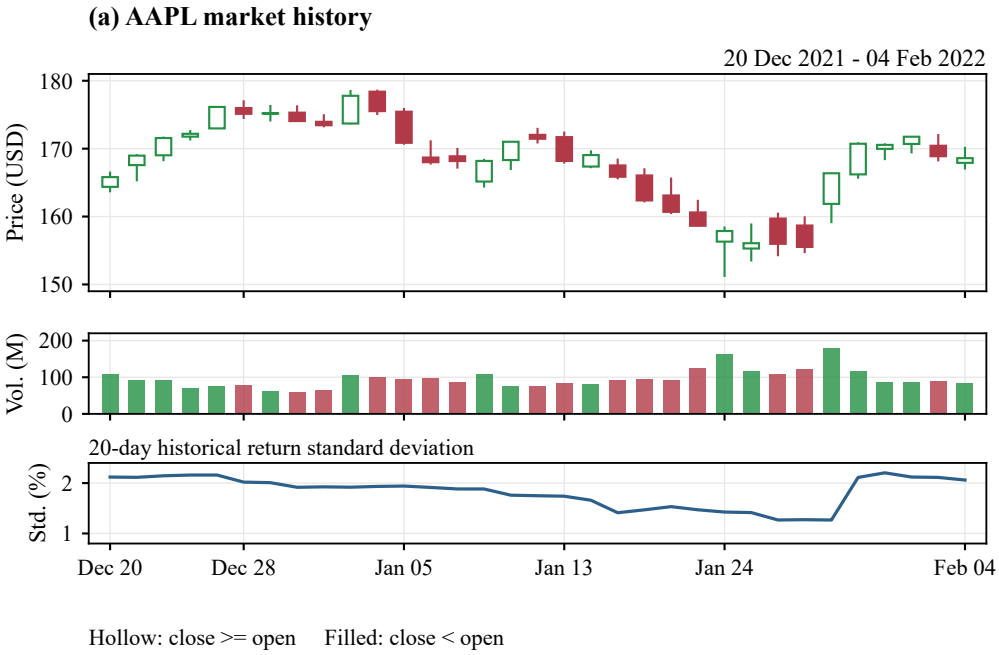



Figure 6: **Dense numerical observations around one earnings call.** Stored adjusted AAPL prices and volume from 20 December 2021 to 4 February 2022. Candles show daily open, high, low, and close; hollow bodies mark close at or above open. Volume is in millions of shares. The bottom series is the standard deviation of the preceding 20 daily returns, shifted by one trading day. Model features are specified in Table 27.

**(b) Earnings-call transcript**

**AAPL | Q1 FY2022 | 27 Jan 2022 | 44,738 characters**

**Tim Cook:**

Today, we are proud to announce Apple's biggest quarter ever. Through the busy holiday season, we set an all-time revenue record of nearly $124 billion, up 11% from last year and better than we had expected at the beginning of the quarter.

**(c) Stored price records**

| Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|
| 2022-01-26 | 159.71 | 160.58 | 154.16 | 155.98 | 108,275,300 |
| 2022-01-27 | 158.68 | 160.04 | 154.61 | 155.53 | 121,954,600 |
| 2022-01-28 | 161.86 | 166.40 | 159.02 | 166.38 | 179,935,700 |

**(d) Company / event fields**

| Field | Value |
|---|---|
| symbol | AAPL |
| available_date | 2022-01-28 |

**(e) Completed previous reactions**

| Call date | Reaction date | Abs. return |
|---|---|---|
| 2021-04-28 | 2021-04-29 | 0.07% |
| 2021-07-27 | 2021-07-28 | 1.22% |
| 2021-10-28 | 2021-10-29 | 1.82% |

Figure 7: **Sparse text, source rows, and completed previous reactions.** A shortened verbatim excerpt from the stored 27 January 2022 AAPL call is shown with three price rows and three previous reactions. The transcript has 44,738 characters. The eligible date, 28 January, follows the next-trading-day convention in Appendix B.1. Previous reactions are absolute close-to-close returns on the first trading day after each earlier call. The shown 28 January price row becomes available at that day's close.

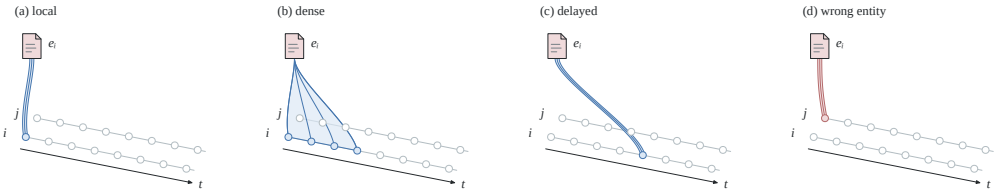



Figure 8: **Duration, timing, and entity are separate exposure controls.** Nodes are prediction times on entity tracks. Local and delayed assignments reach one node; Dense reaches multiple same-entity nodes; wrong-entity reaches track $j$. Assignments follow the daily eligibility convention in Appendix B.1. Node spacing is schematic.

### B.7 Two-Learner Coverage and Selection Protocol

Table 33: Two-learner coverage and selection protocol. Choices are shared across mappings within each learner and encoding. The four-learner Encoding study and activity-matched controls, with their separate resampling family, are specified in Appendix D.

| Component | Specification |
|---|---|
| History | Identical $20 \times 18$ tensor; flattened for LR, sequential for GRU |
| LR | L2, L-BFGS; $C \in \{0.003, 0.01, 0.03, 0.3\}$; max 2,000 iterations; tolerance $10^{-4}$ |
| GRU | Hidden size 24; 16 epochs; batch 1,024; learning rate $\{0.0003, 0.001\}$ crossed with weight decay $\{10^{-5}, 10^{-4}\}$ |
| Seeds | Tuning: 419, 631; final: 419, 631, 853, 1061, 1297 |
| Partitions | $V_h$ and $V_s$ each span 63 trading dates before the test start; $T_e$ spans 126 calendar days |
| Text | Train-fitted TF-IDF/SVD24 or frozen FinBERT/SVD24; presence and log length |
| Timing features | Event-day indicator; log age since latest available event; 30-step event count; no-prior-event flag |
| Pooled records | Exposed vector plus zero second channel; no timing features |
| Pooled + timing | Same pooled content; common timing features |
| Fresh/stale split | Local: fresh vector plus zero stale channel; dense: fresh and stale vectors in separate channels; same timing features |
| Selection | Mean validation AP; tie band $0.25$ pp favors LR |
| Inference | 2,000 circular 20-trading-day block draws within each start, preserving cross-condition pairs; five-day and ticker checks retained |

The record branch is zero-padded to 56 dimensions in all three encodings. Stale vectors equal dense minus local vectors before feature scaling. Mapping-specific scaling is fitted on training data and shared by learners. Hyperparameters use $V_h$; learner selection uses $V_s$. Final fits expand the training prefix, but test results never choose the reported learner. The TF-IDF selection grid clips standardized record features to $[-12, 12]$. In the text-coverage and random-partition comparisons, record features use affine standardization without clipping; the TF-IDF reference is refitted with the same preprocessing. Numerical-history preprocessing is unchanged. The clipped-record selection replay is retained separately in the research archive; the coverage results here use the unclipped follow-up estimates.

## C Breadth Study and Attribution Checks

In the Breadth study's representation check, frozen FinBERT/SVD32 gives mean Local-minus-Dense and Local-minus-wrong-entity AP differences of $2.02$ pp and $2.12$ pp; SVD64 gives $2.10$ pp and $2.13$ pp. Across nine representations with a fixed logistic learner, the local-versus-dense winner changes in 5/6 starts, the mean within-start Kendall correlation is 0.157, and 42.1% of representation pairs reverse order. These retrospective diagnostics describe rank sensitivity.

The model with current-call text exceeds the presence-and-length reference by $0.47$ pp [$-0.03$, 1.02]. The mean margin over past-only donor content is $0.59$ pp [0.09, 1.06] under rolling-start resampling, but ticker and date-block intervals cross zero. Donors generally change both firm identity and content. Against the previous same-firm call, the difference is $-0.01$ pp [$-0.08$, 0.06]. The Breadth study's logistic local-minus-dense and local-minus-wrong-ticker ticker-cluster intervals are [1.57, 3.72] and [1.60, 3.78] AP percentage points. Five-day block intervals are [1.60, 3.39] and [1.58, 3.48]; twenty-day intervals are [1.67, 3.29] and [1.68, 3.35]. Six-fold sign-flip enumeration gives one-sided $p = .015625$ for both contrasts, and Holm-adjusted $p = .03125$ for the two-comparison family. Exactness is conditional on joint sign exchangeability, which shared expanding-window histories do not guarantee unconditionally.

Calendar covariates encode month, quarter, weekday, year, and call intensity within the retained sample. In L2-regularized LR, ticker indicators adjust for persistent firm differences, while unrestricted firm effects and sector-by-time shocks remain uncontrolled. Volatility cutoffs are fitted on training rows. Mean Local-minus-Dense and Local-minus-wrong-entity AP differences remain positive with calendar covariates, ticker indicators, volatility covariates, and their combination. Both

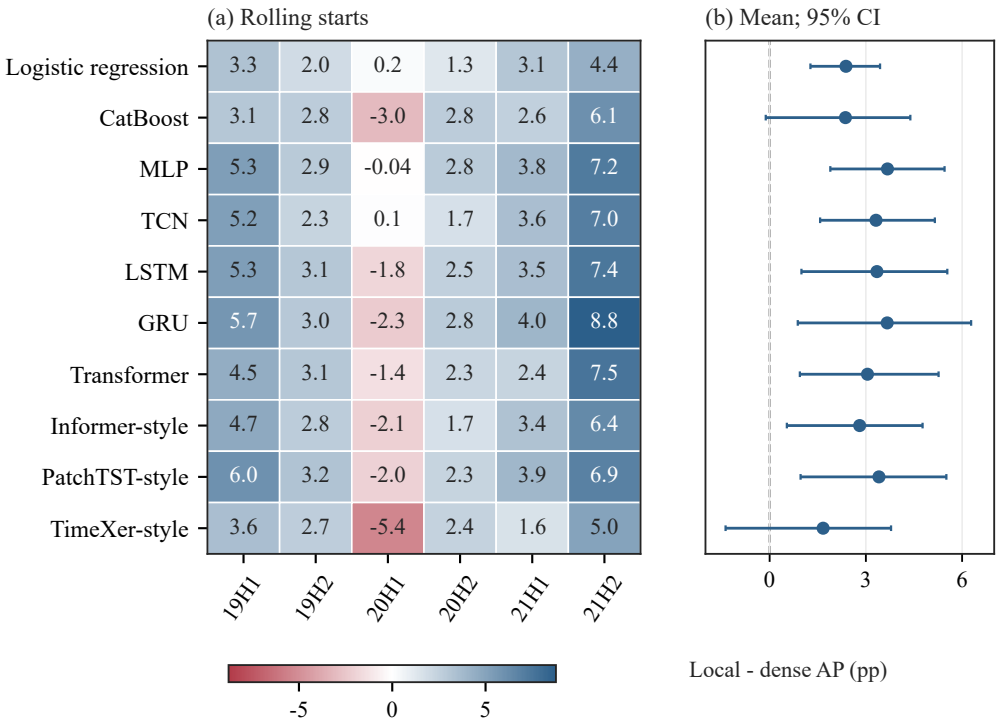



Figure 9: **Ten-learner exposure evidence in 2019–2021.** Left: local-minus-dense AP per start and learner. Right: equal-start means with 95% rolling-start intervals. Blue favors local, red favors dense. The ordering groups model families rather than sorting performance. These are common-task adaptations of the named model families.

differences are positive in five of six test periods under the volatility condition and in all six under the other conditions.

The breadth learners comprise LR, CatBoost (Prokhorenkova et al., 2018), MLP, TCN, LSTM (Hochreiter & Schmidhuber, 1997), GRU, Transformer, and three borrowed-design adaptations. Informer-style borrows self-attention distilling (Zhou et al., 2021) but uses dense attention rather than reproducing ProbSparse attention. PatchTST-style borrows channel-independent patch encoding (Nie et al., 2023). TimeXer-style borrows a global history token and cross-attention to external inputs (Wang et al., 2024). These borrowed-design adaptations solve this paper's classification task through the specified record interface.

The nine representation pipelines are word TF-IDF/SVD24, 96, and 128; character TF-IDF/SVD64; word–character SVD96; MiniLM/SVD32 and 64; and FinBERT/SVD32 and 64. The MiniLM family originates in Wang et al. (2020); our actual checkpoint is `sentence-transformers/all-MiniLM-L6-v2`, a sentence-embedding model described in its model card, not the base distillation checkpoint alone. Frozen encoders use the first 256 word-pieces, masked mean pooling, normalization, and train-fitted SVD. This interface encodes transcript prefixes. In the wrong-entity audit, the mean Kendall correlation with Local-exposure representation ranks is 0.083, 45.8% of pairs reverse, and winners differ in 6/6 starts. These are intentionally invalid entity assignments, not fair alternative deployment mappings.

Past-only content controls average ten donor seeds. Each donor predates its recipient; matching first attempts year–quarter and train-defined length bins, then relaxes the stratum when necessary. The previous-call control uses the last available same-firm transcript. The broader-history ladder and synthetic calibration remain supplementary checks: feature additions alone do not verify information growth, and intentionally label-revealing positive controls are not valid forecasts.

# D Encoding Study and Related Controls

This appendix details the Encoding study used in Sections 4.1–4.2. It concerns the six January/July starts in 2022–2024, TF-IDF and full-text FinBERT, and the unclipped-record protocol. These periods had already been inspected; the execution record fixed the comparison family for these controls before the additional training runs. The ten-partition controls below are distinct from the three-partition text-coverage experiment in Appendix D.2; the ten-partition random controls were run for both representations. The sum/mean/date-matched Coverage audit is a separate study in Appendix A.5.

## D.1 Matched Inputs, Fitting, and Metadata

All cells share prediction rows, labels, and a $20 \times 18$ numerical-history tensor. Records become eligible on the next trading day; local exposure uses age zero and dense exposure uses ages 0–29 in trading steps. Each record has 24 text coordinates, presence, and log length; pooled presence is therefore an active-record count. The 56-dimensional stored input contains two 26-dimensional channels and four common clock coordinates: event-day indicator, $\log(1 + \text{age})$ since the latest eligible event, 30-step event count, and a no-prior-event flag. Pooled records omits the clock through zero padding; pooled + timing adds it; fresh/stale split separates age zero from ages 1–29 and retains the same clock. Full-text FinBERT uses token-count-weighted chunk pooling followed by train-fitted SVD24, not cross-chunk attention.

Metadata controls zero only the 24 text coordinates in each channel. They retain numerical history, channel counts and length, and the clock specified by the encoding. Shared history-only and history-plus-clock references were also fitted. For encoding $q$ and learner $g$, define the content margin and its exposure interaction as

$$C^M_{g,q}(a) = \mathrm{AP}^{\mathrm{text}}_{g,q}(a) - \mathrm{AP}^{\mathrm{metadata}}_{g,q}(a),$$
$$I^M_{g,q} = C^M_{g,q}(\mathrm{local}) - C^M_{g,q}(\mathrm{dense}). \tag{5}$$

These differences use the same rows and resampling weights in all terms and measure paired content-margin and exposure interactions. Table 34 retains all twelve interactions; their intervals are supplementary pointwise intervals, not members of the primary family below.

Table 34: All metadata exposure interactions $I^M$ in AP percentage points. Positive means a larger local-minus-dense gap with content than with metadata. Intervals use 10,000 paired date-block draws; "Positive" counts starts after training-seed AP averaging.

| Text | Learner | Encoding | $I^M$ | Date 95% interval | Positive |
|---|---|---|---|---|---|
| TF-IDF | LR | Pooled records | 3.390 | $[2.405, 4.657]$ | 6/6 |
| TF-IDF | LR | Pooled + timing | 3.459 | $[2.349, 4.717]$ | 5/6 |
| TF-IDF | LR | Fresh/stale split | -0.190 | $[-0.355, -0.029]$ | 0/6 |
| TF-IDF | GRU | Pooled records | 2.809 | $[1.639, 3.983]$ | 6/6 |
| TF-IDF | GRU | Pooled + timing | 1.005 | $[0.340, 1.680]$ | 6/6 |
| TF-IDF | GRU | Fresh/stale split | -0.311 | $[-0.563, -0.004]$ | 2/6 |
| FinBERT (full) | LR | Pooled records | 2.752 | $[1.778, 3.799]$ | 5/6 |
| FinBERT (full) | LR | Pooled + timing | 2.692 | $[1.597, 3.684]$ | 5/6 |
| FinBERT (full) | LR | Fresh/stale split | -0.048 | $[-0.221, 0.068]$ | 3/6 |
| FinBERT (full) | GRU | Pooled records | 2.377 | $[1.291, 3.404]$ | 6/6 |
| FinBERT (full) | GRU | Pooled + timing | 1.082 | $[0.251, 1.910]$ | 5/6 |
| FinBERT (full) | GRU | Fresh/stale split | -0.183 | $[-0.398, 0.161]$ | 1/6 |

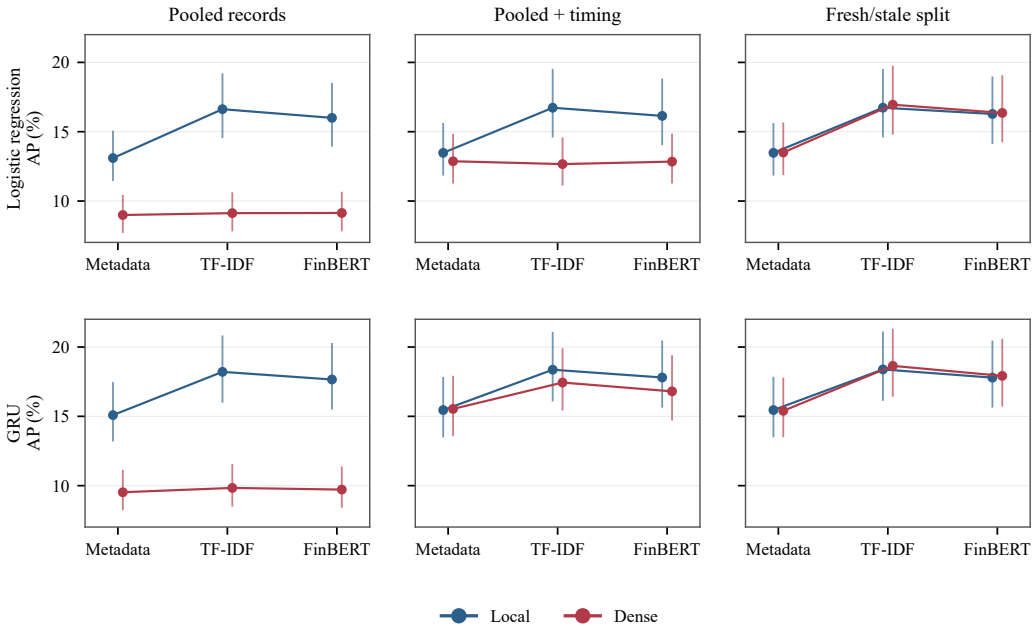



Figure 10: Absolute metadata and text AP under identical rows and numerical histories. Bars are marginal conditional 95% date-block intervals; their overlap is not a paired content test. The full-text label refers to chunk-pooled FinBERT.

**Chronological fitting and aggregation.** The two validation partitions each contain 63 panel trading dates; testing ends 126 calendar days after the start. Outcome endpoints are purged at boundaries. Fits expand from $T$ for hyperparameter validation to $T + V_h$ for learner selection, then $T + V_h + V_s$ for testing; transforms stay fitted on $T$. Record scaling is condition-specific affine standardization without clipping. Neural candidates cross rates $\{0.0003, 0.001\}$ and decay $\{10^{-5}, 10^{-4}\}$, using tuning seeds 419/631 and final seeds 419/631/853/1061/1297. Each learner/encoding shares its selected configuration across exposures; random-control sharing is specified below. Neural training uses 16 epochs, batch 1,024, AdamW, positive class weight $n_-/n_+$, and gradient clipping at 5. LR searches $C \in \{0.003, 0.01, 0.03, 0.3\}$ once per final cell. AP is computed per fit, averaged within each start, then equally across six starts.

### D.2 Text Coverage and Random-Split Follow-up

The coverage follow-up crosses LR/GRU, Local/Dense exposure, and three encodings for TF-IDF, prefix FinBERT, and full-text FinBERT on the same six retrospective periods. Three additional TF-IDF random partitions supply a separate splitting control.

**Frozen encoding.** We encode 1,703 unique transcript texts with frozen ProsusAI/finbert. Content-hash matching recovers full source text for 36 records beyond the stored 80,000-character cap. Both FinBERT variants exclude special tokens from masked mean pooling and apply final L2 normalization. The first-256 input contains at most 254 content tokens and two special tokens. Full-text encoding partitions all content into non-overlapping chunks of at most 254 tokens and averages their embeddings by content-token count. Chunk aggregation covers the full content while processing chunks independently. The stored presence and log-length metadata are unchanged. Train-fitted SVD24 gives equal representation dimensions. Figure 11 retains the full length and coverage distributions.

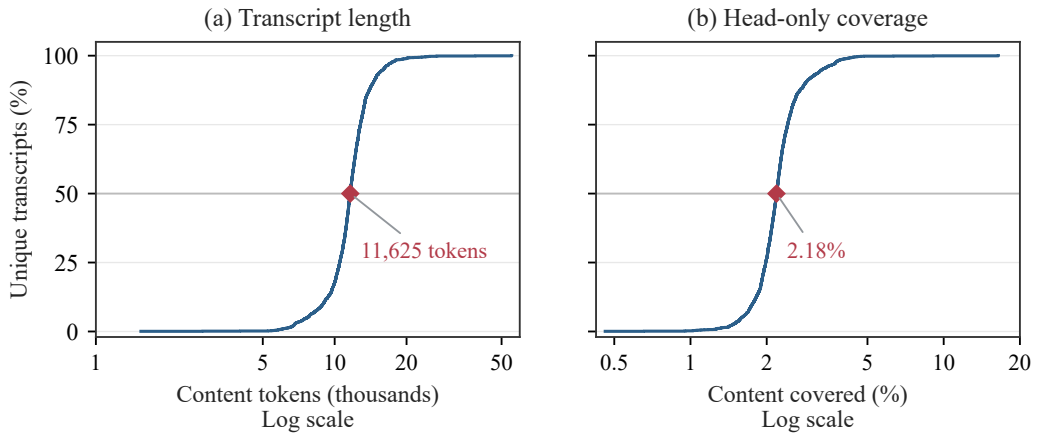



Figure 11: **The first-256 input covers a median 2.18% of transcript content.** Empirical cumulative distributions over all 1,703 unique transcripts, with logarithmic horizontal axes and no tail observations removed. Full-text encoding covers all content tokens by chunk aggregation.

**Random partitions.** Three fixed hashes of record identity and partition seeds 11, 23, and 37 assign records to one of two groups without using labels or record age. Their channels sum to the original exposed vector before scaling and retain the same timing features as pooled + timing and fresh/stale split. Affine scaling without clipping preserves recoverability. Channel activity and sparsity are recorded but not forced to match the age split. Table 35 reports the paired split-minus-random contrast.

**Paired follow-up uncertainty.** All 2,000 circular 20-trading-day block draws preserve ticker rows within sampled dates and use identical weights across conditions. AP is averaged across training

Table 35: Fresh/stale-split exposure gap minus mean random-split exposure gap, in AP percentage points. Negative values indicate a lower signed gap. The 99.375% intervals belong to the eight-comparison follow-up family, not a new two-comparison family.

| Learner | Difference (pp) | 99.375% CI |
|---|---|---|
| Logistic regression | -4.08 | [-5.96, -2.59] |
| GRU | -1.25 | [-2.29, -0.42] |

seeds before equal-start averaging; 10,000 rolling-start draws are also retained. Intervals condition on fitted models and do not jointly cover every source of cross-period dependence. Eight primary contrasts comprise four pooled-to-split exposure reductions (two learners by two FinBERT coverage settings), two full-minus-prefix split-gap differences, and two TF-IDF split-minus-random gap differences. Their Bonferroni intervals use 99.375% coverage; other 95% intervals, including the model-comparison interactions, are descriptive. These three-partition controls use TF-IDF; the ten-partition activity-matched controls cover both TF-IDF and full-text FinBERT (Appendix D).

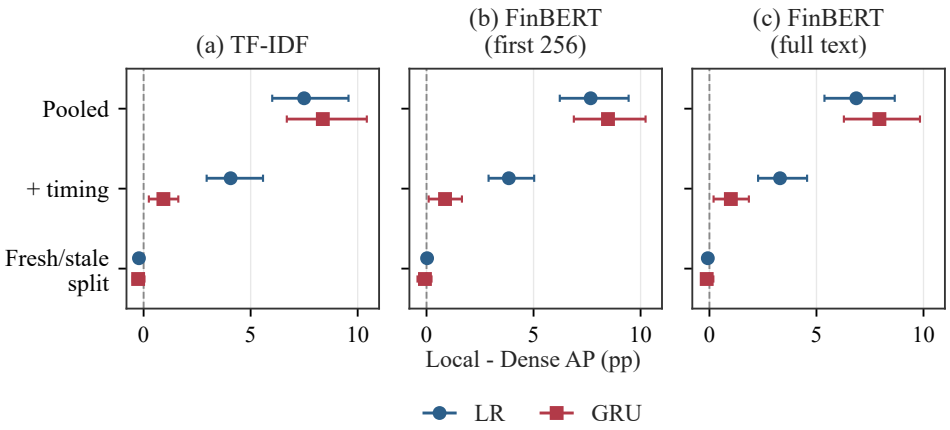



Figure 12: The two-learner text-coverage grid, including prefix FinBERT. All three representations retain the pooled-to-split reduction. Paired 95% date-block intervals condition on fitted predictions; these use 2,000 draws, whereas the four-learner Encoding study uses 10,000.

Table 36: Two-learner text-coverage study: direct model interaction, AP percentage points. These descriptive paired 95% date-block intervals use 2,000 draws. Its selection protocol is specified in Appendix D.2; the Model-gap study in Appendix A also supplies company information.

| Text representation | Pooled records | Pooled + timing | Fresh/stale split |
|---|---|---|---|
| TF-IDF | +0.87 [-0.15, +1.67] | -3.14 [-4.53, -2.17] | -0.04 [-0.30, +0.22] |
| FinBERT (first 256) | +0.81 [-0.32, +1.69] | -2.98 [-4.36, -1.79] | -0.10 [-0.42, +0.19] |
| FinBERT (full text) | +1.08 [-0.04, +2.10] | -2.30 [-3.81, -0.91] | -0.05 [-0.24, +0.26] |

### D.3 Random Routing and Realized Activity Matching

The ten-partition random comparisons are *dense-only* AP contrasts. Each partition preserves the sum of the two content channels and the common clock before affine scaling, with no exchange of text between firms or prediction rows. All ten partition seeds are retained: 11, 23, 37, 53, 71, 89, 107, 131, 157, and 181. The same partitions serve both representations.

**Training-calibrated controls.** Let $p$ be the ratio of fresh-active to dense-active training rows. Rare-record routing assigns an eligible record to the first channel with probability $p$ and keeps this

assignment over its dense lifetime. Random-position routing draws $k = \lfloor 30p \rfloor + \text{Bernoulli}(30p - \lfloor 30p \rfloor)$ distinct positions from its 30-step lifetime; only those positions enter the first channel. Identity-hashed randomness fixes each partition without using labels. Their held-out first-channel rates are 1.339072% and 1.378936%, below the actual fresh rate of 1.909428% (Table 37). They are training-calibrated *approximate* controls.

**Date-matched routing.** For each current prediction date, the rule counts fresh-active rows and randomly selects that many eligible dense-active rows for the first channel; the remainder enter the second. Each selected row retains its own pooled record. Rows with simultaneous fresh and stale content retain their original split and are excluded from rerouting. This matches daily channel-activity counts while leaving row identities and content unmatched. The rule requires current-date panel-wide availability and uses neither labels nor future text. Across these held-out periods both-channel activity is zero, and date-matched first/second rates equal the fresh/stale rates. The date control was added after inspecting input activity, before examining this control's prediction results; the two approximate controls were retained.

One hyperparameter choice per learner and start is shared across all random-control conditions, including the pooled + timing and fresh/stale references: validation AP is first averaged within each of the five families, then equally across families. This prevents the ten-partition families from receiving ten times the tuning weight. Each random contrast is fresh/stale dense AP minus the mean of ten control APs, averaging training seeds within each partition. Partitions and training seeds measure variation on the same market periods; the analysis therefore supports an age-sensitive comparison without identifying a causal age mechanism.

Table 37: Realized held-out channel activity in percent, averaged over six starts after within-start partition averaging. Activity is determined by positive count, not text-vector norm. Both and either refer to simultaneous and union activity.

| Encoding / control | First (%) | Second (%) | Both (%) | Either (%) |
|---|---|---|---|---|
| Pooled + timing | 41.823460 | 0.000000 | 0.000000 | 41.823460 |
| Fresh/stale split | 1.909428 | 39.914032 | 0.000000 | 41.823460 |
| Rare records | 1.339072 | 40.484388 | 0.000000 | 41.823460 |
| Random positions | 1.378936 | 40.444524 | 0.000000 | 41.823460 |
| Date matched | 1.909428 | 39.914032 | 0.000000 | 41.823460 |

### D.4 Encoding Architectures and Clock Gates

TCN and Transformer receive 20 history steps with 18 features per step and a 56-dimensional record input. These are matched-task adaptations. The TCN projects 18 channels to 32 with a width-one convolution, then applies three residual blocks with dilations 1, 2, and 4. Each block contains two width-three, left-padded causal convolutions, ReLU and dropout 0.10; the final history position is read out. The Transformer projects to width 32, adds a learned $1 \times 20 \times 32$ positional tensor, and uses two pre-layer-normalized encoder layers with four attention heads, feed-forward width 128, ReLU, and dropout 0.10. It applies final layer normalization and reads the last position. Attention spans the supplied 20-step history, whose observations all precede the target. Both models use a $56 \to 16$ ReLU record branch and a concatenated $48 \to 32 \to 1$ head with ReLU and dropout 0.10. The LR/GRU references use the results with record features standardized without clipping; the clipped-record selection grid is reported separately.

**Gate construction and unequal capacity.** The three gates share a 24-unit GRU history encoder. Its final state plus the four clock features enters a $28 \to 32 \to 1$ history head, with ReLU and dropout 0.10. Each text expert is $26 \to 16 \to 1$ with ReLU and receives the same pooled first-channel record. The logit is $b(X, C) + q(C)u(z) + (1 - q(C))v(z)$. Shared ties $u = v$; event-day uses the recovered, unstandardized event flag for $q$ with untied experts; learned uses a $4 \to 8 \to 1$ ReLU/sigmoid clock network with untied experts. Counts in Table 38 come from saved fit metadata and are constant across verified test fits. The shared expert is counted once; clock mean/std buffers are not parameters. These gates share input information but have unequal capacity. Stored zero-

padding can also leave registered parameters without effective variation, so these counts are not effective degrees of freedom.

Table 38: Parameter counts for the Encoding-study architectures and clock gates. Counts include registered trainable parameters; zero-padded inputs can leave some parameters inactive.

| Model | Registered parameters | Verified test fits |
|---|---|---|
| TCN | 21,745 | 360 |
| Transformer | 29,233 | 360 |
| Shared gate experts | 4,578 | 120 |
| Event-day gate | 5,027 | 120 |
| Learned clock gate | 5,076 | 120 |

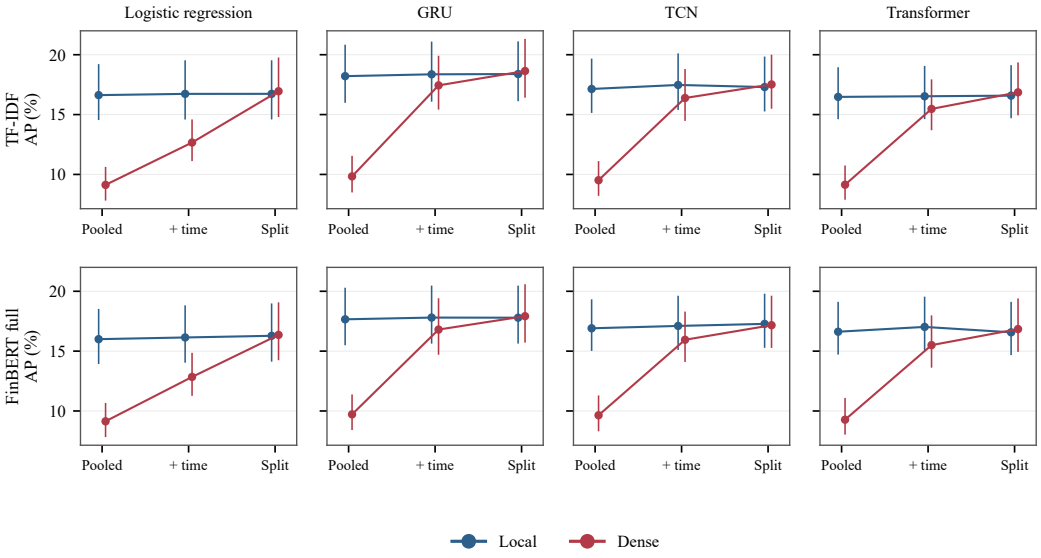



Figure 13: Matched-history absolute AP for both exposures and all three encodings. Marginal conditional intervals describe each score; cross-model inference requires the paired contrasts.

Table 39: Event-day and learned gates minus shared experts, in AP percentage points, for both exposures. These are pointwise paired date-block intervals; only the two learned-minus-shared dense contrasts enter the 24-comparison family.

| Text | Gate vs. shared | Exposure | AP change | Date 95% interval |
|---|---|---|---|---|
| TF-IDF | Event-day | Local | 0.008 | $[-0.088, 0.094]$ |
| TF-IDF | Event-day | Dense | 4.530 | $[3.293, 5.703]$ |
| TF-IDF | Learned | Local | 0.188 | $[0.037, 0.305]$ |
| TF-IDF | Learned | Dense | 2.913 | $[2.150, 3.654]$ |
| FinBERT (full) | Event-day | Local | 0.190 | $[-0.046, 0.411]$ |
| FinBERT (full) | Event-day | Dense | 4.305 | $[3.226, 5.423]$ |
| FinBERT (full) | Learned | Local | 0.314 | $[0.003, 0.535]$ |
| FinBERT (full) | Learned | Dense | 2.081 | $[1.521, 2.710]$ |

### D.5 Primary Family and Conditional Resampling

Write $E_{g,q} = \mathrm{AP}_{g,q}(\mathrm{local}) - \mathrm{AP}_{g,q}(\mathrm{dense})$ and

$$R_g = E_{g,\mathrm{pooled+timing}} - E_{g,\mathrm{fresh/stale}}. \tag{6}$$

Here $R_g$ is local Encoding-study notation for an exposure-gap reduction, distinct from the Background study's allocation margin $R_s(b)$ in Eq. 4. The comparison family contains exactly 24 contrasts across both representations: twelve dense fresh/stale-minus-random contrasts (two representations, two learners, three random families); four reductions $R_g$ (two representations, TCN/Transformer); six reduction differences $R_g - R_{\mathrm{LR}}$ (two representations, GRU/TCN/Transformer); and two dense learned-minus-shared gate contrasts. Positive $R_g$ means a reduction in the *signed* exposure gap from pooled + timing, not from pooled records without timing. Negative $R_g - R_{\mathrm{LR}}$ means less reduction than LR. Metadata, event-day gates, local gate contrasts, normalization, coarse grouping, and individual exposure gaps are outside this family. Tables 40 and 41 split the display, not the adjustment family.

**Paired resampling.** For each start, 10,000 circular 20-date-block draws retain all firms on sampled dates and truncate to the original date count (seed $20260909+s$ for zero-based start $s$). All compared conditions and fitted seeds receive the same date weights. The complementary linked-firm analysis draws 57 firms with replacement 1,000 times (seed 20260913), reusing firm multiplicities across periods, conditions, partitions, and seeds. All replicates are valid jointly. Both schemes recompute weighted AP per fitted seed before seed/start averaging; pointwise intervals use the 2.5th/97.5th percentiles. Date draws preserve within-date dependence; firm draws preserve repeated-company dependence. Neither scheme repeats training or selection.

**Simultaneous intervals.** For each resampling scheme separately, let $\widehat{\theta}_j$ be the observed contrast, $\theta^*_{jb}$ its paired bootstrap draw, and $\widehat{\sigma}_j$ the sample standard deviation of its bootstrap draws. With $j$ ranging over all 24 nondegenerate contrasts, use

$$c_{.95} = \mathrm{quantile}_{.95}\left\{ \max_j \left| \frac{\theta^*_{jb} - \widehat{\theta}_j}{\widehat{\sigma}_j} \right| \right\}, \qquad \widehat{\theta}_j \pm c_{.95}\widehat{\sigma}_j. \tag{7}$$

The date critical value is 2.8757698764693633. The ticker critical value is calculated from its own joint draws. These are max-standardized bootstrap intervals with fixed bootstrap standard deviations; Bonferroni and nested bootstrap-$t$ constructions are not used. We report the date-block and linked-ticker schemes as complementary sensitivity analyses rather than combine them into one bootstrap. All twelve random contrasts exclude zero under both simultaneous schemes. For $R_g$, date intervals exclude zero in three of four cases, except FinBERT/TCN; linked-ticker intervals exclude zero only for FinBERT/Transformer. Linked-ticker reduction differences exclude zero for the three TF-IDF comparisons but not the three FinBERT comparisons.

Separate descriptive checks resample the six period-level effects 10,000 times, omit one period at a time, and report the minimum/maximum equal-period contrast across the five paired final seeds. For deterministic LR, repeated seed entries add no independent fits. Neither these checks nor the prediction bootstraps include uncertainty from refitting thresholds, representations, hyperparameters, or model weights. The analysis reuses retrospective periods with overlapping expanding training prefixes.

### D.6 Text Normalization and Coarse Age Grouping

The normalization check divides each channel's 24 text coordinates by $\max(\|z\|_2, 10^{-8})$ before affine standardization. Counts, lengths, the common clock, and numerical history are unchanged. Both pooled + timing and fresh/stale encodings are refitted. The coarse check replaces dense ages 0 versus 1–29 with ages 0–4 versus 5–29 while preserving the pooled record and clock. Local exposure retains its original age-zero input. Configurations are retuned under the same paired protocol, so a local score change can arise despite unchanged local raw inputs.

Table 42 subtracts reference AP from AP under normalization or coarse grouping, for Local and Dense separately, and reports the signed exposure-gap change. None of the eight normalization gap-change intervals excludes zero. Coarse grouping raises the LR gap by about $3.03\,\mathrm{pp}$ (TF-IDF) and $3.01\,\mathrm{pp}$ (FinBERT), mainly through dense AP losses; the GRU gap changes are less certain. Interpret gap changes alongside both absolute scores; these supplementary intervals are pointwise.

Table 40: All twelve TF-IDF members of the joint 24-comparison family. AP effects and simultaneous 95% intervals are in percentage points; date uses 10,000 draws and linked ticker uses 1,000. "Split - control" compares dense AP.

| Contrast | Learner | AP change | Date simultaneous | Ticker simultaneous |
|---|---|---|---|---|
| Split - rare records | LR | 4.377 | $[2.483, 6.272]$ | $[2.332, 6.422]$ |
| Split - random positions | LR | 4.287 | $[2.424, 6.150]$ | $[2.261, 6.313]$ |
| Split - date matched | LR | 4.100 | $[2.281, 5.918]$ | $[2.157, 6.042]$ |
| Split - rare records | GRU | 1.417 | $[0.405, 2.428]$ | $[0.334, 2.499]$ |
| Split - random positions | GRU | 1.264 | $[0.266, 2.263]$ | $[0.241, 2.287]$ |
| Split - date matched | GRU | 1.499 | $[0.549, 2.450]$ | $[0.496, 2.503]$ |
| $R_g$ | TCN | 1.309 | $[0.224, 2.393]$ | $[-0.097, 2.714]$ |
| $R_g$ | Transformer | 1.333 | $[0.438, 2.229]$ | $[-0.113, 2.780]$ |
| $R_g - R_{\mathrm{LR}}$ | GRU | -3.102 | $[-4.848, -1.356]$ | $[-5.171, -1.033]$ |
| $R_g - R_{\mathrm{LR}}$ | TCN | -2.974 | $[-4.756, -1.192]$ | $[-5.008, -0.939]$ |
| $R_g - R_{\mathrm{LR}}$ | Transformer | -2.949 | $[-4.833, -1.065]$ | $[-5.024, -0.874]$ |
| Learned - shared (dense) | GRU gate | 2.913 | $[1.813, 4.013]$ | $[1.487, 4.339]$ |

Table 41: All twelve full-text FinBERT members of the same 24-comparison family, using the same definitions and joint adjustment as Table 40. The family is not reset by representation.

| Contrast | Learner | AP change | Date simultaneous | Ticker simultaneous |
|---|---|---|---|---|
| Split - rare records | LR | 3.516 | $[1.876, 5.156]$ | $[1.559, 5.473]$ |
| Split - random positions | LR | 3.385 | $[1.743, 5.027]$ | $[1.449, 5.322]$ |
| Split - date matched | LR | 3.265 | $[1.682, 4.848]$ | $[1.400, 5.130]$ |
| Split - rare records | GRU | 1.648 | $[0.659, 2.637]$ | $[0.167, 3.128]$ |
| Split - random positions | GRU | 1.702 | $[0.754, 2.651]$ | $[0.243, 3.161]$ |
| Split - date matched | GRU | 1.898 | $[0.970, 2.825]$ | $[0.393, 3.402]$ |
| $R_g$ | TCN | 1.046 | $[-0.260, 2.352]$ | $[-0.497, 2.589]$ |
| $R_g$ | Transformer | 1.791 | $[0.758, 2.823]$ | $[0.362, 3.220]$ |
| $R_g - R_{\mathrm{LR}}$ | GRU | -2.245 | $[-4.470, -0.020]$ | $[-4.707, 0.216]$ |
| $R_g - R_{\mathrm{LR}}$ | TCN | -2.327 | $[-4.485, -0.170]$ | $[-4.865, 0.211]$ |
| $R_g - R_{\mathrm{LR}}$ | Transformer | -1.583 | $[-3.240, 0.075]$ | $[-4.023, 0.858]$ |
| Learned - shared (dense) | GRU gate | 2.081 | $[1.201, 2.961]$ | $[0.811, 3.351]$ |

Figures show marginal score intervals where indicated; paired and simultaneous comparisons use the families defined above. The accompanying reproduction notes document source files and numerical checks.

Table 42: Supplementary sensitivity effects in AP percentage points. L2 rows compare against the same encoding without L2 normalization; coarse rows compare against fresh/stale splitting. The final interval is pointwise 95% paired date-block uncertainty for $\Delta E = \Delta AP_l - \Delta AP_d$.

| Text | Learner | Modification | $\Delta AP_l$ | $\Delta AP_d$ | $\Delta E$ | Date 95% interval |
|---|---|---|---|---|---|---|
| TF-IDF | LR | L2: pooled + timing | 0.592 | 0.022 | 0.570 | $[-0.177, 1.018]$ |
| TF-IDF | LR | L2: fresh/stale | 0.591 | 0.582 | 0.009 | $[-0.134, 0.104]$ |
| TF-IDF | LR | Ages 0–4 / 5–29 | 0.013 | -3.019 | 3.033 | $[2.091, 4.192]$ |
| TF-IDF | GRU | L2: pooled + timing | 0.252 | 0.302 | -0.050 | $[-0.324, 0.218]$ |
| TF-IDF | GRU | L2: fresh/stale | 0.236 | 0.259 | -0.023 | $[-0.240, 0.127]$ |
| TF-IDF | GRU | Ages 0–4 / 5–29 | -0.218 | -0.594 | 0.375 | $[-0.198, 0.931]$ |
| FinBERT (full) | LR | L2: pooled + timing | -0.003 | 0.002 | -0.005 | $[-0.043, 0.041]$ |
| FinBERT (full) | LR | L2: fresh/stale | -0.141 | -0.125 | -0.017 | $[-0.061, 0.061]$ |
| FinBERT (full) | LR | Ages 0–4 / 5–29 | -0.143 | -3.156 | 3.013 | $[2.140, 4.093]$ |
| FinBERT (full) | GRU | L2: pooled + timing | 0.053 | 0.021 | 0.032 | $[-0.095, 0.158]$ |
| FinBERT (full) | GRU | L2: fresh/stale | -0.002 | 0.087 | -0.089 | $[-0.254, 0.051]$ |
| FinBERT (full) | GRU | Ages 0–4 / 5–29 | 0.020 | -0.060 | 0.079 | $[-0.662, 0.851]$ |

# E Routing/Age Study: Equal-Parameter and Context Controls

The Routing/age comparisons reuse the six retrospective 2022–2024 starts and prediction rows from the matched-history earnings-call study. Here the history encoder is a 24-unit GRU and the record contains 24 TF-IDF or full-text FinBERT coordinates. Each expert is a $24 \to 16 \to 1$ ReLU network with no final bias. With $\widetilde{u}(P) = u(P) - u(0)$ and the corresponding $\widetilde{v}$, the score is

$$f(X, P, C) = b(X, C) + q\widetilde{u}(P) + (1 - q)\widetilde{v}(P). \tag{8}$$

Constant allocation sets $q = 1/2$; event allocation uses the event-day indicator; random allocation uses ten date-activity-matched assignments. Both experts see the same pooled content $P$. The no-content model retains $b(X, C)$, including metadata. Common context contains the four clocks, fresh/stale counts and length, and all ten random indicators for every routing variant. No text is exchanged between rows. Parameter counts are 5,409 for each two-expert variant, 4,993 for tied experts, and 4,577 without content. The equal-parameter controls match registered parameter count while leaving function class and optimization dynamics different.

**Availability, transforms, and fitting.** Local availability and dense ages 0–29 follow the main protocol. Text coordinates are divided by standard deviations of unique training records, floored at $10^{-6}$, without mean subtraction; an absent record remains zero. History and context transforms use initial training data. All modes start from a common two-expert initialization; shared or absent experts are then tied or removed. AdamW uses class-weighted binary cross-entropy, batch size 1,024, and gradient clipping at 5. The common search has four 16-epoch settings (learning rate $0.0003/0.001$ crossed with weight decay $10^{-5}/10^{-4}$) plus two 32-epoch settings (rate 0.001 and decay $10^{-4}/10^{-3}$). Tuning uses seeds 419 and 631; final fitting uses seeds 419, 631, 853, 1061, and 1297. Validation chooses a common configuration by first averaging within routing families and then equally across families; the no-content model is tuned independently. The four-setting, 16-epoch reference search is also retained in the saved outputs. Final fitting uses all pre-test partitions, without refitting initial transforms.

**Decay and company context.** The decay control multiplies each eligible record by $2^{-a/h}$ at age $a$ before summing, without renormalizing the weights. Both allocations share the validation-selected half-life from $\{1, 3, 5, 10, 30, \infty\}$ and the uniform-weight routing study's selected network configuration. The company-context control adds 57 training-vocabulary ticker indicators plus an unknown coordinate and eight past-call summaries to the common context. These include prior-event availability/count, latest absolute reaction, means over two/four reactions, standard deviation and maximum over four, and elapsed time. A reaction enters only after its outcome has completed, never on the current event's outcome day. The context grows from 18 to 84 coordinates; two-expert and no-content counts become 7,521 and 6,689. The cross-context comparison therefore augments information and changes capacity in a refitted pipeline.

Table 43: Complete primary routing comparisons, AP percentage points. Intervals use 5,000 linked-firm bootstrap draws, with the same firm weights across periods. The uniform-weight, decay, and identity/history comparisons use separate adjustment families of four, two, and two comparisons, respectively.

| Context | Text | Paired contrast | AP pp | Adjusted interval |
|---|---|---|---|---|
| Uniform weights | FinBERT | Event – constant | 2.516 | [1.115, 3.833] |
| Uniform weights | FinBERT | Event – random mean | 2.390 | [1.009, 3.765] |
| Uniform weights | TF-IDF | Event – constant | 4.131 | [2.371, 5.928] |
| Uniform weights | TF-IDF | Event – random mean | 4.044 | [2.366, 5.779] |
| Decay | FinBERT | Constant: decay – uniform | 1.604 | [0.681, 2.579] |
| Decay | FinBERT | Event – constant, both decayed | 0.858 | [0.133, 1.553] |
| Identity/history | FinBERT | Event – constant | 0.054 | [-0.708, 0.806] |
| Identity/history | FinBERT | Event – random mean | 0.162 | [-0.560, 0.902] |

Table 44: FinBERT absolute AP (percent), averaged over five seeds then six periods. No content retains history and common metadata. Differences are point estimates.

| Context | No content | Constant | Event | Event – no content |
|---|---|---|---|---|
| Base | 14.003 | 13.757 | 16.273 | 2.270 |
| Identity + past reactions | 18.105 | 18.128 | 18.182 | 0.077 |

Table 45: Direct changes in event-minus-constant AP. Negative means attenuation. These descriptive 95% linked-firm intervals are outside the two primary comparison families. Decay and company controls are separate interventions.

| Change in event – constant margin | AP pp | Paired 95% interval |
|---|---|---|
| After decay | -1.658 | [-2.528, -0.847] |
| After identity + past reactions | -2.462 | [-3.215, -1.617] |

**Inference and interpretation.** All AP contrasts average scores over seeds before periods, not probabilities before AP. The four uniform-weight routing comparisons retain positive adjusted intervals under both linked-firm and circular 5-/20-date-block analyses. These intervals resample fixed fitted predictions; training and selection are not repeated. With FinBERT, event-minus-constant is positive in five of six starts and has five positive seed-averaged effects; after decay it is positive in four starts and all five seed averages. With joint identity/history, three of six period contrasts are positive. The linked-firm change in the routing margin is $-2.462\,\mathrm{pp}$ $[-3.215, -1.617]$, descriptive 95%. All three models improve in absolute AP, with the Constant and no-content models gaining more. This joint-addition experiment leaves the separate roles of identity and past reactions unresolved. The Background study in Appendix A.4 separates the two additions under its own shared-selection protocol. A normalized-recency supplement is descriptive: with only one eligible record it preserves that record regardless of age, whereas unnormalized decay changes its magnitude.

# F Learned-Age Comparison

This study supplies the learned-age comparison in Table 2. It uses six 2022–2024 starts, frozen full-text FinBERT/SVD24, and the $20 \times 18$ history tensor. Training-fitted transforms, vocabulary, and availability features remain fixed during the final pre-test refit. All previous-call reactions have completed before use. Linked-firm inference uses 20,000 draws with each firm's weight synchronized across periods. The four age contrasts use Bonferroni percentile limits $0.025/4$ and $1 - 0.025/4$, conditional on fitted predictions. Five-/20-date blocks assess alternative dependence structures.

## F.1 Learned Age Weights

A learned gate uses training-standardized $\log(1 + \mathrm{age})$ with a $1 \to 4 \to 1$ tanh/sigmoid network, adding 13 parameters to the same two-expert predictor. Shared expert initialization and dropout random streams are retained across variants; an absent record has zero text contribution. Three allocation rules (constant, event-day, learned age) are crossed with no decay and common exponential decay. Each cell selects from the same six configurations using two tuning seeds; five final seeds give 432 tuning and 180 final fits.

The common half-life comes from the shared-routing decay study in Appendix E, with period choices 1, 30, or infinity. It stays fixed during learned-gate selection, giving the gate the same network search budget. Within this per-cell-selection study, event-minus-constant is $2.161\,\mathrm{pp}$ without decay and $0.761\,\mathrm{pp}$ with decay; the shared-routing comparison in Appendix E uses a different selection protocol. After common decay, event-minus-learned is $0.604\,\mathrm{pp}$. Its four-comparison-adjusted linked-firm interval is $[-0.097, 1.323]$, compared with [0.026, 1.210] for 20-date blocks and $[-0.076, 1.323]$ for five-date blocks. We use the primary linked-firm result in the main text; evidence for the residual advantage is sensitive to the resampling scheme. Learned-age-minus-constant with common decay is $0.157\,\mathrm{pp}$ $[-0.167, 0.454]$, leaving the additional learned-gate gain uncertain.

Table 46: All six learned-age-study cell means, AP percent.

| Condition | AP (%) |
|---|---|
| Decay, Constant | 15.4116 |
| Decay, Event | 16.1723 |
| Decay, Learned age | 15.5687 |
| No decay, Constant | 14.1179 |
| No decay, Event | 16.2790 |
| No decay, Learned age | 13.7631 |

Table 47: All four primary age comparisons, AP percentage points; four-comparison-adjusted linked-firm intervals. "Age" denotes the learned gate.

| Contrast | Estimate | Adjusted 95% interval |
|---|---|---|
| Event – learned age, no decay | +2.5158 | [+1.1919, +3.9851] |
| Learned age – constant, no decay | -0.3548 | [-0.8411, -0.0264] |
| Event – learned age, decay | +0.6035 | [-0.0969, +1.3230] |
| Learned age – constant, decay | +0.1572 | [-0.1666, +0.4536] |

**Training and period sensitivity.** For Event minus learned-age routing under common decay, four of six period means are positive; they range from $-0.521$ to $1.568\,\mathrm{pp}$. Leave-one-period-out means range from $0.411$ to $0.828\,\mathrm{pp}$, and five seed means from $0.568$ to $0.651\,\mathrm{pp}$. These are descriptive ranges. A broader 24-contrast historical-family check gives a linked-firm interval of $[-0.272, 1.526]\,\mathrm{pp}$, retaining the uncertainty about this comparison.

## G Additional History and Text Conditions

**Metric interpretation.** EventGain centers AP by the positive-label prevalence, $\mathrm{EventGain}_s = P_s - \pi_{+,s}$. On identical rows, subtracting the same prevalence leaves paired EventGain and AP differences equal. The separate history study additionally measures negative log loss in bits per prediction; a gain is the reduction in out-of-sample log loss for the fitted learner. This follows the computationally usable-information framing of Xu et al. (2020); unrestricted Shannon information is not estimated here. The history study compares a text predictor with its metadata-retaining reference at each history level and tests the change in that margin. Useful history is assessed through reference-model improvement.

**History-study configurations.** The three numerical-history conditions have 18 baseline features, 26 with prior-call summaries, and 41 with lagged market-state variables (Table 48). Dimensionality is descriptive only; predictive value is assessed by no-content improvement and the change in text margin. LR candidates are $C \in \{0.1, 1, 10\}$; histogram gradient boosting (HGB) uses $\{3, 7, 15\}$ maximum leaves, 80 iterations, learning rate 0.05, L2 regularization 1, minimum leaf size 30, and no early stopping. Both use seed 90531296. The final 126 calendar days of the training prefix provide validation. A shared hyperparameter is chosen using mean log loss over the nine history–record configurations. The grid comprises 108 final fits and 324 validation fits. Inner target thresholds, vocabulary, SVD, imputation, and scaling are trained independently before validation; final fits refit preprocessing on the outer training prefix. Outcome endpoints are purged at boundaries.

Table 48: Separate LR/HGB history-study inputs and nine history-record cells. This study uses daily feature vectors rather than the matched-architecture $20 \times 18$ sequence.

| Pack | Width | Fields / record conditions |
|---|---|---|
| X0 | 18 | Base 18 daily lag/rolling fields; not a 20-step tensor in this separate LR/HGB study. |
| X1 | 26 | X0 plus the eight completed prior-reaction fields. |
| X2 | 41 | X1 plus five prior daily OHLCV quantities, each lag1 / mean5 / mean20: overnight gap, intraday return, log high-low range, close location, log Amihud. |
| Nine cells | 3 x 3 | Each X0/X1/X2 crossed with no record, local presence+log length, local content24+presence+log length. Detailed formulas in field_dictionary.csv. |

Synchronized-ticker resampling uses the same firm multiplicities across all six starts. Date-block resampling selects ordered dates and their company rows jointly within a start. We use 2,000 ticker or date-block draws and 10,000 rolling-fold draws. The AP and log-score tables retain all four learner–source interactions and their separate uncertainty definitions.

Table 49: Separate LR/HGB history study: representation comparison in AP pp, with rolling-start intervals. Here $H$ is history gain given metadata (95% CI); $J$ is the change in the additional text margin (98.75% CI, adjusted for four learner–history interactions per representation). All four expansions are retained. Separate representation intervals do not test a between-representation difference.

| Text representation | $H$ [95% CI] | $J$ [98.75% CI] |
|---|---|---|
| **LR + prior-call history** | | |
| TF-IDF | -0.11 [-0.36, +0.15] | +0.40 [-0.02, +0.78] |
| FinBERT (full text) | -0.11 [-0.36, +0.15] | +0.07 [-0.15, +0.25] |
| **LR + market state** | | |
| TF-IDF | +0.82 [+0.49, +1.11] | +0.45 [+0.24, +0.72] |
| FinBERT (full text) | +0.82 [+0.49, +1.11] | +0.26 [-0.03, +0.57] |
| **HGB + prior-call history** | | |
| TF-IDF | +2.48 [+0.82, +4.15] | -1.67 [-3.09, -0.58] |
| FinBERT (full text) | +2.48 [+0.82, +4.15] | -0.62 [-1.20, +0.11] |
| **HGB + market state** | | |
| TF-IDF | +0.09 [-1.37, +1.61] | +0.30 [-0.45, +1.24] |
| FinBERT (full text) | +0.09 [-1.37, +1.61] | +0.17 [-0.88, +1.02] |

Table 50: Separate LR/HGB history study: representation comparison in millibits per prediction, with synchronized-ticker intervals. $H$ uses 95% intervals; $J$ uses four-comparison-adjusted 98.75% intervals per representation. Useful-history and text-interaction evidence must be checked separately.

| Text representation | $H$ [95% CI] | $J$ [98.75% CI] |
|---|---|---|
| **LR + prior-call history** | | |
| TF-IDF | -1.05 [-2.60, +0.43] | +0.36 [-0.07, +0.80] |
| FinBERT (full text) | -1.05 [-2.60, +0.43] | +0.16 [-0.22, +0.49] |
| **LR + market state** | | |
| TF-IDF | +4.00 [+2.12, +6.15] | +0.17 [-0.15, +0.49] |
| FinBERT (full text) | +4.00 [+2.12, +6.15] | +0.24 [-0.10, +0.58] |
| **HGB + prior-call history** | | |
| TF-IDF | +2.40 [-0.53, +5.23] | -2.26 [-3.42, -1.27] |
| FinBERT (full text) | +2.40 [-0.53, +5.23] | -0.82 [-1.81, +0.02] |
| **HGB + market state** | | |
| TF-IDF | +4.00 [+2.07, +6.00] | +0.37 [-0.31, +1.12] |
| FinBERT (full text) | +4.00 [+2.07, +6.00] | -0.04 [-0.65, +0.56] |

## H Controlled Persistence and State Conditions

Fifty independent simulated datasets each contain 48 entities and 960 time steps. A stationary Gaussian two-coordinate latent signal follows an AR(1) process with $\rho \in \{0.2, 0.95\}$. Records arrive every 40 steps with entity-specific phase and contain the signal at release plus Gaussian measurement noise and independent distractors. An observed iid binary state $s$ either leaves the predictive coefficient unchanged (equal condition) or selects which latent coordinate predicts the outcome (different condition). Conditional event probability is logistic in lagged numerical features and the relevant current latent coordinate. The intercept targets a marginal positive rate near 10%. Common random streams pair the four conditions within each dataset.

All predictors receive the same 20-step numerical history, observed state, record age, and availability metadata. Dense exposure retains a single mapped record through age 29, whereas Local exposure changes the available content and therefore represents a different comparison from the equal-information controls. Shared content weights are compared with state interactions, four age bins, their joint interaction, continuous exponential decay, and state plus continuous decay. The state is independent of record age, unlike the deterministic event-day routing in the financial experiment. The reference integrates unobserved current signal conditional on the record assigned under the stated availability rule; it does not observe the hidden signal or extra historical records.

Training ends at step 639, validation covers 640–799, and testing covers 800–959. Each fitted logistic model selects regularization, and where applicable decay, by validation log loss; final fits use training plus validation with training-fitted transforms. Main scores are negative unweighted log loss in *nats per row*, not the bits used by the real-history study. All conditions, models, and worlds are retained. The six core and two continuous-age contrasts use separate Bonferroni percentile families over 10,000 paired world resamples. Their 99.1667% and 97.5% intervals appear in Table 51. The predeclared 0.002-nat threshold is a generator-specific working scale for practical effect size. Function classes and parameter counts differ across interaction bases.

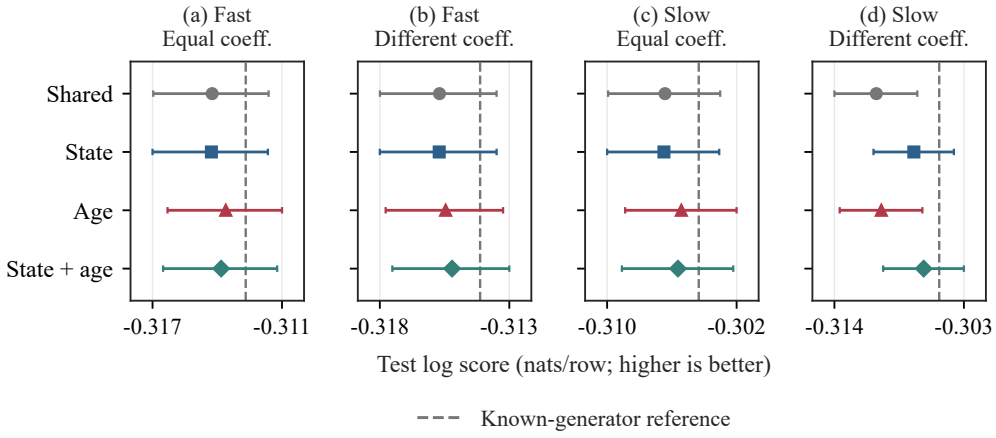



Figure 14: All four simulated conditions and the four core fitted models. Dashed references use the known generator with the same record assigned under the stated availability rule. Marginal world intervals describe absolute scores; paired comparisons are in Table 51.

Two adjusted lower bounds exceed the 0.002-nat working scale: the slow-decay coefficient interaction and the slow/different state benefit beyond continuous age. The other six intervals lie within $[-0.002, 0.002]$, though several exclude zero. Within this generator, persistence also changes learning difficulty and temporal dependence.

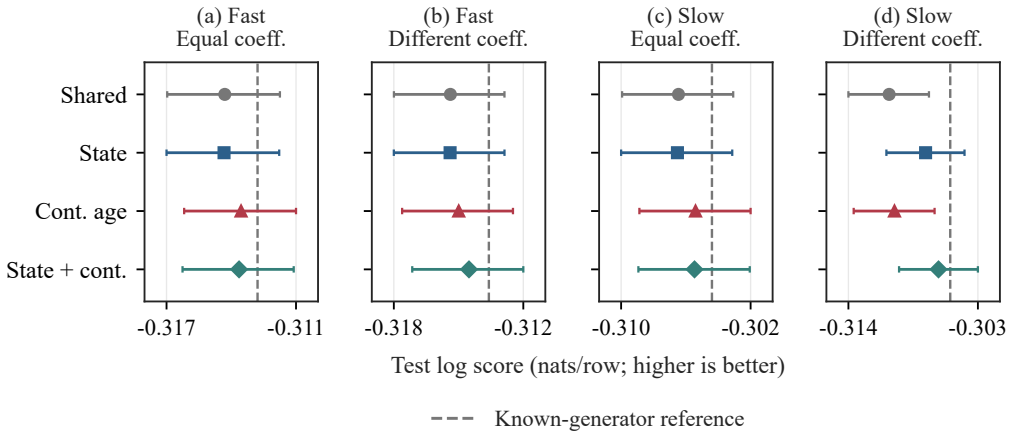



Figure 15: Continuous decay evaluated on the same set of 50 independent datasets. Left to right: fast/equal, fast/different, slow/equal, and slow/different coefficients. Benefits of state interaction depend on signal persistence and state-specific coefficients; these controlled conditions describe the specified generators.

Table 51: All eight registered log-score contrasts, nats per row. B1/B2 subtract the state-minus-shared benefit with equal coefficients from that with different coefficients at the stated persistence. The remaining rows are direct differences as labeled. Families are separate.

| Test | Paired log-score contrast | Estimate | Adjusted interval |
|---|---|---|---|
| B1 | State-coefficient interaction, fast decay | 0.000024 | [-0.000035, 0.000080] |
| B2 | State-coefficient interaction, slow decay | 0.003236 | [0.002680, 0.003799] |
| B3 | Age benefit: slow – fast, equal coefficients | 0.000370 | [0.000079, 0.000631] |
| B4 | Age benefit: slow – fast, different coefficients | 0.000165 | [-0.000029, 0.000350] |
| B5 | State + age – state, fast/different | 0.000536 | [0.000325, 0.000745] |
| B6 | State + age – age, fast/different | 0.000269 | [0.000086, 0.000455] |
| BD1 | Continuous age – shared, fast/equal | 0.000811 | [0.000656, 0.000971] |
| BD2 | State + continuous age – continuous age, slow/different | 0.003829 | [0.003355, 0.004314] |

# I External Scope and Uncertainty

The external studies intervene at different released interfaces. The VoT cache audit fixes forecasting weights and changes cached post-reasoner forecasts; the Aurora grid fixes a pretrained checkpoint and changes record selection. These comparisons retain each system's metric and evaluation dates. VoT normalizes across batch elements, so forecasts at different origins can influence one another. Target alignment and fixed batch composition preserve a paired retrospective comparison, while the normalization contribution remains unresolved.

## I.1 Target-Aligned Cache Sensitivity

The cache audit fixes VoT model weights and varies the age of released post-reasoner forecasts. SocialGood uses a six-month target and cache ages 0–6 months; Energy uses a twelve-week target and ages 0–12 weeks. At each forecast origin, an older donor's vector is sliced to the same target dates. Results average five training seeds. Figure 16 shows all age–lead cells; Table 52 averages the six or twelve output steps on common origins within each dataset.

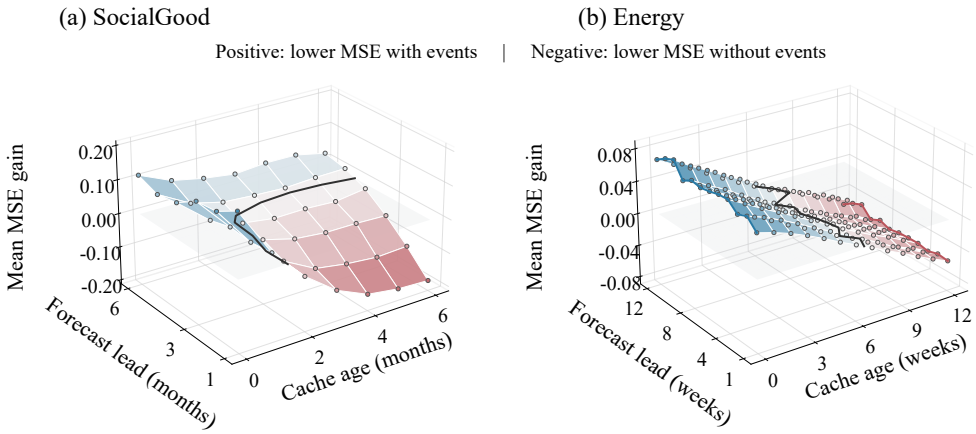



Figure 16: **Module gains across cache ages and forecast leads.** Frozen VoT models receive cached event forecasts aligned with the same target dates. Height and color show mean MSE gain over the separately trained no-event reference; the grey plane and black contour mark zero mean gain. Surfaces connect evaluated grid points, shown as dots. The panels use different vertical scales. Table 52 reports simultaneous intervals for horizon-averaged gains and their changes.

Define $G(a) = \mathrm{MSE}_{\text{no event}} - \mathrm{MSE}_{\mathrm{Full},a}$ at cache age $a$. On Energy, $G(12) - G(0) = -0.1195$, with a 38-comparison simultaneous interval $[-0.1882, -0.0508]$. The change remains negative with half and double the primary block length, and across all five seeds and both three-period and five-period partitions. The age-12 gain itself varies across three periods: $-0.1088$, $+0.0333$, and $-0.0792$. The evidence supports a decline relative to the latest cache; the whole-task sign reversal remains uncertain. SocialGood's horizon-averaged gains likewise have simultaneous intervals crossing zero at every tested age.

Table 52: **All horizon-averaged VoT cache gains and direct changes.** $G$ is no-event minus Full MSE; $\Delta G = G(a) - G(0)$. Brackets are 38-comparison simultaneous 95% intervals at the primary block length.

| Age | $G$ | Interval | $\Delta G$ | Interval |
|---|---|---|---|---|
| *SocialGood; age in months* | | | | |
| 0 | $+0.1336$ | $[-0.0569, +0.3242]$ | — | — |
| 1 | $+0.0448$ | $[-0.0593, +0.1490]$ | $-0.0888$ | $[-0.2814, +0.1038]$ |
| 2 | $-0.0093$ | $[-0.1655, +0.1469]$ | $-0.1430$ | $[-0.4335, +0.1476]$ |
| 3 | $-0.0395$ | $[-0.2197, +0.1406]$ | $-0.1732$ | $[-0.5000, +0.1537]$ |
| 4 | $-0.0500$ | $[-0.2148, +0.1148]$ | $-0.1836$ | $[-0.5039, +0.1367]$ |
| 5 | $-0.0541$ | $[-0.1935, +0.0854]$ | $-0.1877$ | $[-0.4881, +0.1128]$ |
| 6 | $-0.0557$ | $[-0.1808, +0.0695]$ | $-0.1893$ | $[-0.4725, +0.0939]$ |
| *Energy; age in weeks* | | | | |
| 0 | $+0.0680$ | $[+0.0029, +0.1331]$ | — | — |
| 1 | $+0.0575$ | $[-0.0042, +0.1192]$ | $-0.0105$ | $[-0.0182, -0.0027]$ |
| 2 | $+0.0474$ | $[-0.0112, +0.1060]$ | $-0.0206$ | $[-0.0356, -0.0056]$ |
| 3 | $+0.0365$ | $[-0.0195, +0.0926]$ | $-0.0315$ | $[-0.0531, -0.0099]$ |
| 4 | $+0.0257$ | $[-0.0282, +0.0796]$ | $-0.0423$ | $[-0.0706, -0.0141]$ |
| 5 | $+0.0151$ | $[-0.0369, +0.0671]$ | $-0.0529$ | $[-0.0876, -0.0182]$ |
| 6 | $+0.0047$ | $[-0.0458, +0.0551]$ | $-0.0633$ | $[-0.1040, -0.0226]$ |
| 7 | $-0.0057$ | $[-0.0554, +0.0441]$ | $-0.0737$ | $[-0.1201, -0.0272]$ |
| 8 | $-0.0144$ | $[-0.0635, +0.0346]$ | $-0.0824$ | $[-0.1346, -0.0303]$ |
| 9 | $-0.0239$ | $[-0.0732, +0.0254]$ | $-0.0919$ | $[-0.1492, -0.0346]$ |
| 10 | $-0.0327$ | $[-0.0830, +0.0176]$ | $-0.1007$ | $[-0.1625, -0.0388]$ |
| 11 | $-0.0418$ | $[-0.0940, +0.0105]$ | $-0.1098$ | $[-0.1753, -0.0442]$ |
| 12 | $-0.0515$ | $[-0.1060, +0.0031]$ | $-0.1195$ | $[-0.1882, -0.0508]$ |

**Paired inference.** The 38-comparison family contains 20 age-specific gains and 18 changes from age 0 across both datasets. Paired circular date-block and seed resampling uses 10,000 draws, with primary blocks of six months for SocialGood and twelve weeks for Energy; half and double lengths check sensitivity. The 198 output-step cells form a separate simultaneous family. Intervals condition on the fitted models and selected rules, excluding full retraining and reselection.

**Evaluation conditions.** This exploratory audit reuses previously inspected periods. Cache age changes content and the donor's original prediction lead together. Target dates are aligned within the released library, whose historical online availability is unverified. Batch composition stays fixed across conditions, but normalization can couple an origin's prediction to later-origin inputs; its contribution to the measured loss differences remains unresolved.

### I.2 Record Selection with a Fixed Checkpoint

The Aurora grid uses the released checkpoint, a 192-month numerical lookback, five generation seeds, and 100 generated predictions per forecast. It evaluates 18 record positions at 6-, 8-, 10-, and 12-month horizons on 88 common test origins (144 training and 39 validation origins). All 72 mean gains over the same checkpoint without text are positive, ranging from 0.003181 to 0.006333 MSE (Table 53). Record position changes content, age, coverage, and length together. This retrospective grid uses common origins across horizons on previously studied periods.

In a separate exploratory validation-selected record-choice experiment, mean gain over the default was $-0.001085$ MSE (two-comparison-adjusted interval $[-0.002896, +0.000448]$), with negative means in all three periods. Grid intervals and full selector results are provided in the aggregate supplement.

Table 53: **Aurora: all 72 mean text gains.** Values are no-text minus selected-record MSE on 88 common test origins, averaged over five generation seeds. Position 0 is the oldest record in the 192-month lookback; 191 is the latest. Positive point estimates describe the grid, not simultaneous evidence of positive gains in every cell.

| Record position | 6 months | 8 months | 10 months | 12 months |
|---|---|---|---|---|
| 0 | 0.003900 | 0.004685 | 0.005324 | 0.005979 |
| 6 | 0.004233 | 0.005064 | 0.005698 | 0.006333 |
| 12 | 0.003181 | 0.004111 | 0.004798 | 0.005548 |
| 24 | 0.003535 | 0.004347 | 0.004911 | 0.005558 |
| 36 | 0.003600 | 0.004484 | 0.005150 | 0.005831 |
| 48 | 0.003534 | 0.004420 | 0.005235 | 0.005892 |
| 60 | 0.003703 | 0.004554 | 0.005223 | 0.005900 |
| 72 | 0.004046 | 0.004858 | 0.005441 | 0.006047 |
| 84 | 0.003918 | 0.004752 | 0.005344 | 0.005968 |
| 96 | 0.003826 | 0.004616 | 0.005293 | 0.005938 |
| 108 | 0.003283 | 0.004094 | 0.004842 | 0.005589 |
| 120 | 0.003203 | 0.004133 | 0.004869 | 0.005644 |
| 132 | 0.003699 | 0.004474 | 0.005076 | 0.005714 |
| 144 | 0.004112 | 0.004912 | 0.005441 | 0.006077 |
| 156 | 0.003827 | 0.004569 | 0.005109 | 0.005771 |
| 168 | 0.004042 | 0.004777 | 0.005305 | 0.005880 |
| 180 | 0.003603 | 0.004405 | 0.005052 | 0.005730 |
| 191 | 0.003614 | 0.004362 | 0.005045 | 0.005752 |

### I.3 Screening and Unresolved Cases

Table 54: External evidence is screened for accessible inputs and an interpretable intervention. Null or unavailable cases remain documented; no frequency claim is inferred from the positive subset.

| System | Status | Consequence for evidence |
|---|---|---|
| VoT | Released post-reasoner inputs; matched retraining | Positive module-by-mapping case at that stage |
| Aurora | Fixed released checkpoint; available-record selection | Small selection sensitivity; text-gain date uncertainty |
| CAMEF (Zhang et al., 2025) | Checkpoint reproduced; five-seed retraining text gain uncertain | No stable positive module evidence in the tested retraining. |
| StockNet (Xu & Cohen, 2018) | Data recovered; faithful original execution unresolved | Only a protocol-level surrogate was available. |
| Time-MMD Public Health | Five-seed text and mapping gains uncertain | No stable sensitivity at this audited interface. |
| Time-MMD Climate | Released future facts repaired | The mean text gain remains after the validity repair. |
| CFA / UniCA (Lee et al., 2026; Han et al., 2026) | Baseline gain absent or unstable in tested setup | No stable positive module-gain case |
| IGSTGNN / MEANT (Fan et al., 2026; Irving & Schoene, 2024) | Required timelines or data unavailable | Complete external audit unavailable. |
| VolTAGE (Sawhney et al., 2020) | Released representation pool crosses data partitions | Excluded from clean predictive evidence |

CAMEF's official checkpoint gives scaled MSE 0.0004790 versus the reported 0.0004886. In five matched retraining seeds, its text gain averages $-5{,}584$ raw-MSE units with a 95% seed interval [$-30{,}214$, 15,732], positive in 2/5 seeds. Checkpoint reproduction and a stable retrained module advantage are different requirements. Public Health's five-seed horizon-12 reported text gain is $-0.0060$ MSE with an interval crossing zero. Climate's repaired gain is $+0.0675$ [0.0181, 0.1416], versus $+0.0579$ [0.0165, 0.1134] before repair; the interaction is uncertain. A fact-order transformation preserving the truncated token multiset in the Public Health average-pooling interface changes

MSE by 0.0079, inside the specified 0.01 band. These cases document uncertain, positive, and small sensitivities across the audited interfaces.

## J TRAFFIC AS A SCOPE TEST

The released XTraffic archive (Gou et al., 2025) supplies numerical sensors and categorical incident reports, not transcript text. We retain 128 sensors selected using January data. An earlier low-flow target mainly captured recurring low demand and its category encoder collapsed several raw categories. We preserve that analysis as an exploratory predecessor, then evaluate a separately defined negative-speed-anomaly task with eight raw category coordinates. This redesign followed inspection of the archive and reuses the same data.

Predictions occur every 30 minutes from twelve preceding 5-minute observations. Histories retain flow, occupancy, speed, and training-normalized speed residuals, with calendar inputs. Normal speed is the training median by sensor, weekday and half-hour, with training-only fallbacks for sparse groups. The target is the future mean speed residual below that sensor's training tenth percentile. Horizons 15/30/60 minutes share eligible rows. Incident availability is first-report time plus 5 or 30 minutes, rounded up to the next 5-minute slot. Recent exposure uses one hour; dense uses 24 hours. No resolution time, duration, or description is supplied. Category revision histories, exact timezone/DST and historical data vintages are unavailable, limiting real-time interpretations. For fourteen sensors in the earliest training fold, the 30-minute anomaly threshold is smaller than 0.1 released speed unit in magnitude, so the statistical event can correspond to a very small speed change.

Three expanding training periods end in April, June and August 2023; two-month validation periods precede July–August, September–October and November–December tests. Target endpoints are purged at boundaries. Train-only transforms are frozen, then model fits use training plus validation. LR and GRU compare history, history plus event clock, category-free metadata, and category-content models under pooled, timed, and split encodings. During training/validation development, we strengthened the common history baseline after a simple residual predictor outperformed raw-history LR.

Splitting gains vary across settings, despite some nonzero exposure and learner interactions. The 0.5-pp value is a descriptive practical reference; no equivalence margin was registered. The financial mitigation pattern is therefore not stable across the tested traffic settings. This grid holds recent numerical-state visibility fixed; Appendix J.1 tests that factor directly. Whether recent history accounts for the weak event margins remains unresolved by this grid.

Table 55: Complete primary traffic families at 30-minute horizon, 5-minute report delay, overall population. AP percentage points; 2,000 paired date-cluster draws, preserving same-date sensors and seeds. Bonferroni family sizes are 4 for event increments, 6 for exposure gaps, 3 for learner interactions, and 2 for dense splitting. These intervals condition on fitted predictions under date-cluster sampling and do not include full retraining uncertainty.

| Learner | Contrast | AP pp | Adjusted interval |
|---|---|---|---|
| LR | Content beyond history/clock, recent | 0.0079 | [0.0048, 0.0108] |
| LR | Content beyond history/clock, dense | 0.0003 | [-0.0094, 0.0098] |
| LR | Recent – dense, pooled | 0.0008 | [-0.0091, 0.0103] |
| LR | Recent – dense, timing | 0.0076 | [-0.0017, 0.0167] |
| LR | Recent – dense, split | 0.0080 | [-0.0011, 0.0166] |
| LR | Split – timing, dense | -0.0003 | [-0.0026, 0.0020] |
| GRU | Content beyond history/clock, recent | -0.0366 | [-0.0841, 0.0079] |
| GRU | Content beyond history/clock, dense | -0.1858 | [-0.2389, -0.1259] |
| GRU | Recent – dense, pooled | -0.0710 | [-0.1266, -0.0132] |
| GRU | Recent – dense, timing | 0.1493 | [0.0858, 0.2119] |
| GRU | Recent – dense, split | 0.1224 | [0.0682, 0.1746] |
| GRU | Split – timing, dense | 0.0269 | [-0.0229, 0.0769] |
| GRU–LR | Exposure interaction, pooled | -0.0718 | [-0.1228, -0.0187] |
| GRU–LR | Exposure interaction, timing | 0.1416 | [0.0868, 0.1968] |
| GRU–LR | Exposure interaction, split | 0.1144 | [0.0633, 0.1635] |

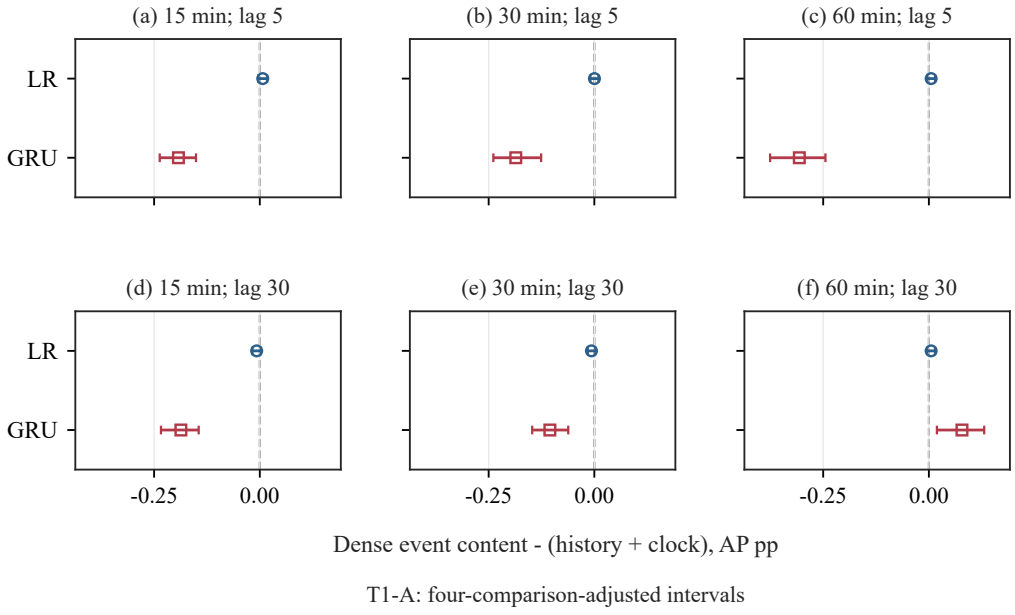



Figure 17: Complete traffic horizon and report-delay sensitivity on the common overall sample. Event increments are measured beyond history plus the report clock. Bars use 98.75% marginal limits from each four-comparison event-increment family (2,000 paired date-cluster draws, conditional on fitted predictions), not joint coverage of the entire display. Direct contrasts appear in Table 55.

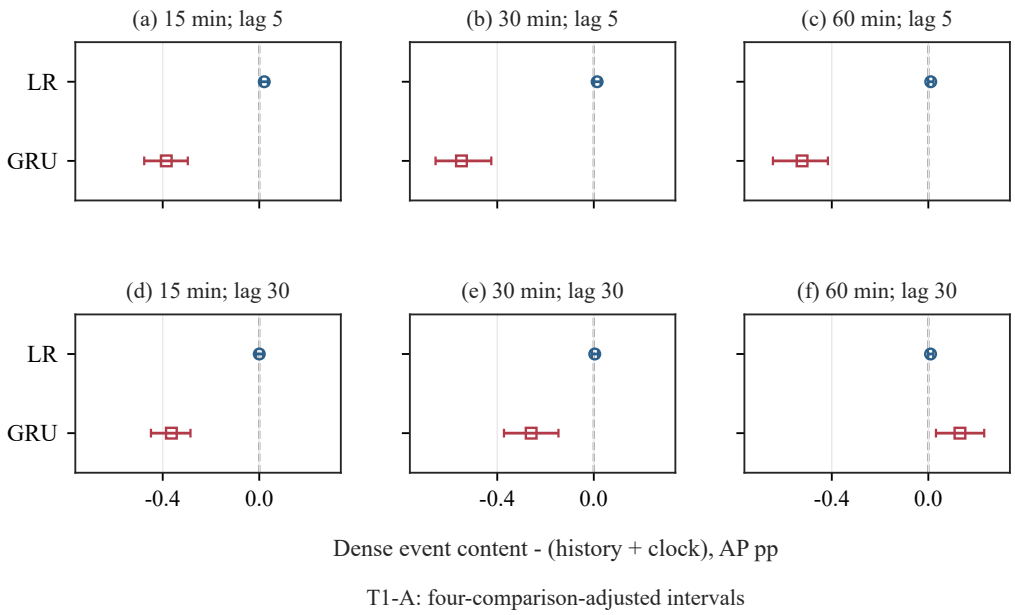



Figure 18: Predefined onset subset: the recent observed residual is not yet below its training threshold. Eligibility uses only past observations and is common across methods. Bars use the same per-setting four-comparison 98.75% limits as Figure 17. The overall sample and primary horizon remain the main evaluation.

### J.1 Traffic History and Exploratory Contrasts

This follow-up uses 128 sensors and three 2023 test periods, July–August, September–October, and November–December. It retains the corrected negative-speed-anomaly task, a 30-minute horizon, and a five-minute report delay. Event inputs contain the eight original incident categories and availability metadata, not description text. Training-only normal-speed estimates define the target.

Earlier history provides observations at lags 60 through 20 minutes (nine slots). Full history adds the 15-, 10-, and 5-minute lags. Both use a $12 \times 5$ tensor with masking and visibility indicators, identical prediction rows, incidents, and future labels. Each history is crossed with no event, event metadata, pooled categories, and split categories for LR and GRU. The primary event-content margin is *pooled minus metadata*, not split minus no-event. LR searches $C \in \{0.001, 0.01, 0.1, 1\}$ with class balancing and 600 iterations; GRU searches three configurations with two tuning and five final seeds. There are 240 tuning fits and 144 final fits.

Recent state improves metadata AP by $9.858\,\mathrm{pp}$ for LR and $5.671\,\mathrm{pp}$ for GRU. Because the earlier-history pooled-content margin is already nonpositive, its decline with recent history does not test absorption of a previously positive pooled signal. In contrast, the exploratory split-minus-metadata contrast reveals a positive earlier-history margin for GRU. Here split AP (75.7911%) exceeds metadata (75.5560%) but not no-event (75.9719%). Descriptively, the category input partly offsets the metadata pipeline's lower mean. The conditional margin's seed-averaged decline occurs in all three periods, while the full-history margin changes sign across periods. Across seeds, the period-averaged changes are $+0.365$, $-0.286$, $-0.253$, $-0.593$, and $-0.346\,\mathrm{pp}$; their directions vary. LR's corresponding margin is nonpositive, limiting this pattern to the GRU split-minus-metadata comparison.

**Eight explicitly exploratory comparisons.** The first two contrasts use a prior 16-cell factorial comparison: base/joint company context, Local/Dense exposure, pooled + timing/fresh-stale encoding, and LR/GRU. It used the same six financial periods, full-text FinBERT/SVD24, and matched histories, with five LR and six GRU candidates selected separately per cell. Neural tuning used two

Table 56: All 16 traffic cell means, AP percent. "No event" uses history without event features; "Metadata" is the event-metadata reference for content increments.

| Condition | AP (%) |
|---|---|
| Earlier history, No event, GRU | 75.9719 |
| Earlier history, No event, LR | 54.8758 |
| Earlier history, Metadata, GRU | 75.5560 |
| Earlier history, Metadata, LR | 54.8739 |
| Earlier history, Pooled, GRU | 75.2569 |
| Earlier history, Pooled, LR | 54.8732 |
| Earlier history, Split, GRU | 75.7911 |
| Earlier history, Split, LR | 54.8738 |
| + Recent state, No event, GRU | 81.2121 |
| + Recent state, No event, LR | 64.7330 |
| + Recent state, Metadata, GRU | 81.2267 |
| + Recent state, Metadata, LR | 64.7322 |
| + Recent state, Pooled, GRU | 80.7810 |
| + Recent state, Pooled, LR | 64.7247 |
| + Recent state, Split, GRU | 81.2393 |
| + Recent state, Split, LR | 64.7329 |

Table 57: All six primary traffic contrasts, AP percentage points, with six-comparison-adjusted seven-day calendar-block intervals. $H$ improves the metadata baseline, $J$ changes pooled-minus-metadata, and $K$ changes split-minus-pooled when recent state is added.

| Contrast | Estimate | Adjusted 95% interval |
|---|---|---|
| $H$: LR | +9.8582 | [+9.5014, +10.2239] |
| $J$: LR | -0.0067 | [-0.0112, -0.0027] |
| $K$: LR | +0.0075 | [+0.0050, +0.0097] |
| $H$: GRU | +5.6706 | [+5.4113, +5.9572] |
| $J$: GRU | -0.1466 | [-0.2732, -0.0596] |
| $K$: GRU | -0.0759 | [-0.1717, +0.0511] |

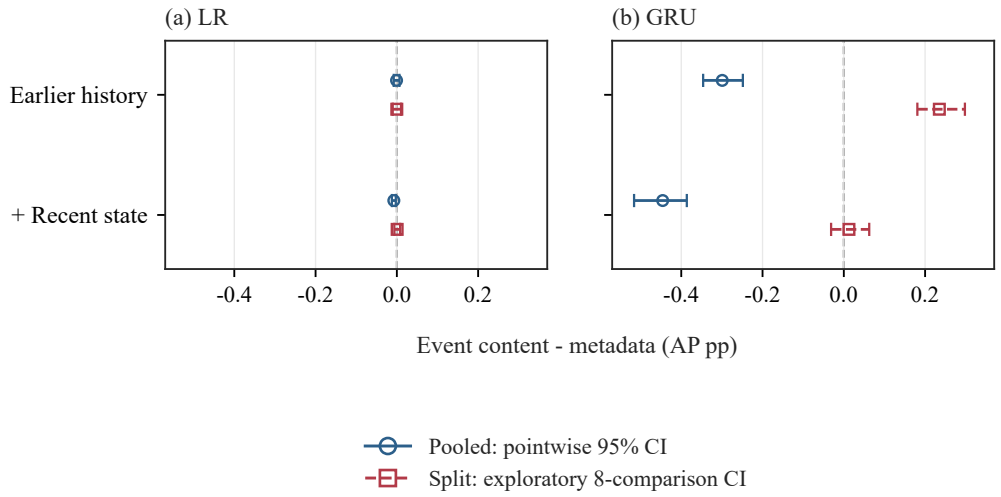



Figure 19: **The traffic history result depends on the learner and encoding.** Pooled-minus-metadata uses descriptive pointwise 95% calendar-block intervals; the derived split-minus-metadata intervals use the separate eight-comparison exploratory adjustment. The pooled and split intervals belong to different comparison families; no direct pooled-versus-split interval is shown here. Both learners, encodings, and histories are shown.

seeds and final fits used five. Full historical scores remain in the research archive; all eight derived contrasts are retained here to preserve their joint family. After reading the completed primary analyses, we form two split-minus-timing model interactions and six split-traffic contrasts. Within each source experiment, draws are combined with their original paired weights, retaining covariance. Their intervals use a separate eight-comparison Bonferroni family over the 20,000 saved draws. These contrasts remain exploratory; their eight-contrast family is separate from the 24 primary contrasts.

Table 58: All eight exploratory derived contrasts, AP pp, preserving their joint adjustment family. The first two rows use the earlier per-cell-selected model-comparison protocol, specified below, rather than the main Model-gap study. The six LR/GRU rows use the traffic history intervention: split-minus-metadata or full-minus-earlier-history changes.

| Derived contrast | Estimate | Adjusted 95% interval |
|---|---|---|
| Interaction change, base history | +1.7805 | [-0.3819, +3.8824] |
| Interaction change, both controls | +3.7273 | [+1.7282, +5.7780] |
| LR: split gain, earlier | -0.0001 | [-0.0101, +0.0103] |
| LR: split gain, recent state | +0.0007 | [-0.0062, +0.0080] |
| LR: change in split gain | +0.0009 | [-0.0045, +0.0052] |
| GRU: split gain, earlier | +0.2351 | [+0.1809, +0.2979] |
| GRU: split gain, recent state | +0.0126 | [-0.0313, +0.0625] |
| GRU: change in split gain | -0.2224 | [-0.2819, -0.1593] |

## K Nearest-Work Comparison

Table 59 connects existing work on exposure, company information, and recency to our paired learner-gap comparison. The Model-gap study changes assignment with company information fixed; the separate Background study changes the text-model reference margin. These extend prior work through two distinct comparisons.

Table 59: Nearest-work overlap and the narrower EventAlign tests. The table specifies the comparison added by each study.

| Prior work | Already studied | Narrower EventAlign test |
|---|---|---|
| Yu et al. (2025) | Release-aware volatility targets, same-firm reaction baselines, and identity-preserving random embeddings (Secs. 4.2, 5, 6.2). | Daily extreme-return classification: test learner-by-exposure interaction with company information fixed; separately vary the background reference for the text-model margin. |
| Zhang et al. (2026) | Numerical-backbone strength, numerical/dynamic-text context lengths, and unique versus redundant text (Secs. 4.3, 4.5–4.6; App. D.4). | Hold numerical history and the record corpus fixed; vary record reuse across prediction rows and directly compare learner gaps. |
| Time-MMD (Liu et al., 2024) | Text start/end timestamps and a text lookback independent of the numerical lookback (Secs. 2.3, 3.1). | Pair exposure rules on unchanged rows and labels to test comparison stability. |
| Time-IMM (Chang et al., 2025) | Normalized Gaussian recency weighting, timestamp-aware attention, and fusion ablations (Sec. 4; App. I.1, Eqs. 6–7). | Use common unnormalized decay to test changes in allocation margins; its weights retain age-dependent magnitude. |
| TimesX (Liu et al., 2026) | Benchmark-dependent rankings and text replacement with numerical series, model, and prompt fixed (Secs. 3.1.1, 3.1.3; Table 3). | Retain the record corpus, change its exposure across the same prediction rows, and estimate paired learner-gap interactions. |

Time-IMM's normalized Gaussian recency average and our common unnormalized exponential decay are different controls: with a single eligible record, normalization preserves the embedding's

magnitude, whereas unnormalized decay attenuates it. Likewise, TimesX's cross-benchmark ranking comparison and its single-model controlled text replacement answer different questions; our learner-by-exposure interaction retains the record collection and prediction rows. Company-history substitution motivates our controls: Yu et al. (2025) study post-earnings volatility, while our task classifies daily extreme returns.

Table 60: Additional neighboring questions. Recent preprints are cited for their specific reported findings.

| Work | Verified contribution | Relation to this study |
|---|---|---|
| Context is Key (Williams et al., 2025) | Forecasting tasks constructed to require textual context (abstract and benchmark design). | Establishes a context-use evaluation question; our reuse comparison permits either direction of the text effect. |
| MUSE-Bench (Chen et al., 2026) | Wrong Context mixes semantic-replacement and timing-error perturbations 1:1 (Section 4.2.2), reporting their combined effect. | We instead compare learner gaps under one-day and prolonged same-firm reuse, alongside negative controls. |
| Fidel-TS (Xu et al., 2025) | Source integrity, temporal/description leakage, and explicitly scheduled inputs (Sections 1 and 3.1). | Motivates separate availability checks. Our daily availability convention provides a weaker source-time guarantee than Fidel-TS. |
| MILM (Chung et al., 2026) | Value-redaction tests separate sampling-pattern information from values in healthcare (abstract and evaluations). | Provides cross-domain motivation for retaining metadata in no-content references. |

Koval et al. (2023) study long-transcript earnings-surprise prediction, distinct from text–numerical financial forecasting studied by Koval et al. (2024). Yuan et al. (2023) develop sparse attention and multi-source counterfactual augmentation for earnings-call analysis. These approaches model long documents and multi-source representations; our experiments use frozen-vector interfaces. Siano (2025) relates textual earnings disclosures to short-window market responses; that evidence motivates the event-response task; our record-availability assumptions are specified separately in Appendix B.1.

**Fusion and audited system identities.** Zhou et al. (2026) vary last-value shortcuts and context quality to probe contextual-fusion benefits. Our primary contrast instead changes record assignment while retaining the record collection and prediction rows, and measures the resulting change in the between-learner AP gap.

Constrained fusion addresses cases where naive fusion underperforms a unimodal model (Lee et al., 2026); UniCA adapts heterogeneous covariates to a time-series foundation model (Han et al., 2026). These architectural questions complement our assignment comparisons. CAMEF combines macroeconomic announcements and numerical series (Zhang et al., 2025). StockNet uses tweets and historical prices (Xu & Cohen, 2018), whereas MEANT organizes temporally distributed price, tweet, and graphical information (Irving & Schoene, 2024). Our primary task instead uses earnings-call transcripts.

IGSTGNN evaluates an incident-guided forecasting architecture with temporal impact decay (Fan et al., 2026); our traffic audit varies the XTraffic incident-input interface. VolTAGE combines text/audio with a company graph (Sawhney et al., 2020). Our audit identifies the released repository at commit `597d69d`: company representations pool call features across partitions. This artifact inspection concerns the released pooling implementation; we did not retrain the released VolTAGE system. Table 54 reports our audit results and execution limits for these released systems.