SIGNAL-TTT: CLOSED-FORM CROSS-SECTIONAL TEST-TIME TRAINING FOR EVENT-DRIVEN PORTFOLIO CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Event-driven asset ranking aims to predict market impact and construct long-short portfolios following market-moving information, such as Donald Trump's Truth Social posts. While frozen large language models can score individual (post, instrument) pairs, test-time adaptation (TTA) is critical to adjust for post-training market regimes. However, existing TTA methods fit a single parameter per test input, which in this context represents an entire event cross-section. We prove that any additive or strictly monotonic per-event correction inherently preserves within-event relative orderings, leaving rank metrics and portfolio performances unchanged. To address this limitation, we propose SIGNAL-TTT, a closed-form cross-sectional test-time training framework. SIGNAL-TTT computes a per-(event, instrument) offset using ridge-regularized least squares over retrieved historical outcomes, automatically calibrating its regularizer via empirical Bayes. Requiring no gradients, learning rates, or access to model weights, our update is provably a shrunk per-instrument mean residual with a explicit error-reduction condition. Evaluated on a development set of 321 Truth Social policy events across 9 frozen backbones, SIGNAL-TTT improves Rank IC by +0.0156 and boosts dollar-neutral net returns by +2.32 bps per event, whereas collapsing to a single per-event offset completely eliminates these gains.

1 INTRODUCTION

Policy announcements move markets within minutes (Tsang & Yang, 2026; Lindahl, 2025), and they move instruments unevenly: a tariff can favor domestic producers over importers. Acting on such an event means answering two questions at once, *which* instruments the event moves and *in which direction*. Large language models are a natural tool for this (Lopez-Lira & Tang, 2026; Yue et al., 2026). A frozen model is asked, once per (event, instrument) pair, whether the instrument moves up, down or not at all. The probability of UP minus that of DOWN is then a single score whose sign is the direction call and whose size is the model's conviction. Sorting the instruments of one event by this score and trading its two ends gives a long-short book, so the quantity a method improves and the quantity a desk trades are the same. However, frozen models are poorly calibrated on this task, and their errors differ from one instrument to the next. The outcome of every past event matures within the hour and can be stored, which is exactly the setting test-time methods are built for (Liang et al., 2025). Existing methods typically advance along two pathways: **label-free adaptation** updates a model on the unlabeled test stream by entropy minimization or cached confident predictions (Wang et al., 2021; Niu et al., 2023; Karmanov et al., 2024); and **retrieval-based adaptation and test-time training** correct a prediction with the labels of retrieved neighbors (Khandelwal et al., 2020; Zhang et al., 2023; Hardt & Sun, 2024). Calibration methods for language models share the same structure (Zhao et al., 2021; Zhou et al., 2024) prove the work significance.

While these methods have achieved commendable progress, they share a fundamental assumption: **one test input carries one prediction to correct**. On a cross-sectional task this assumption fails, for three reasons. ❶ A test input is an event carrying a whole vector of instruments scored against each other. A parameter indexed by the event adds the same number to every instrument, so it *cannot change their order*, and the order is what a ranking metric and a rank-built book respond to (Figure 1). ❷ The label-free objectives sharpen the model's confidence, and a frozen model's most confident

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.

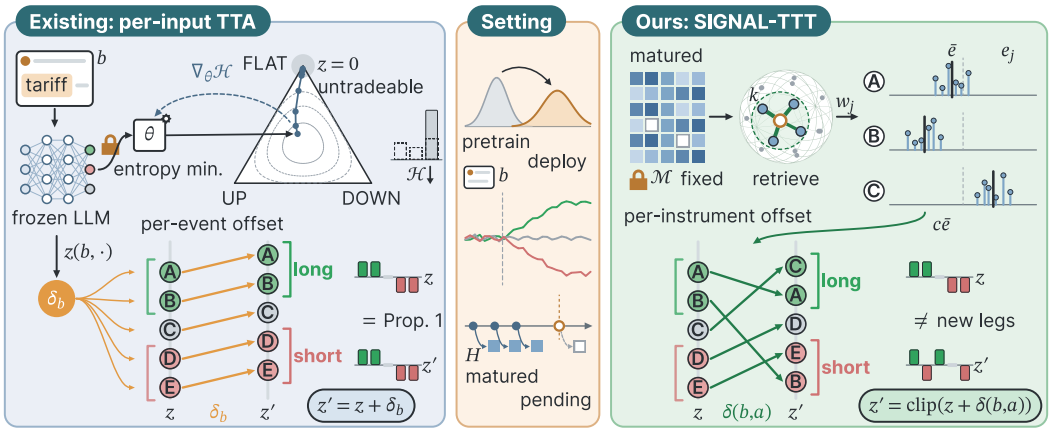

Figure 1: Cross-Sectional Test-Time Adaptation with SIGNAL-TTT. Standard TTA (left) either collapses to zero-yield predictions via entropy minimization or fits a shared offset that leaves relative orderings unchanged (Prop. 1). By leveraging retrieved historical outcomes for uneven policy impacts (middle), SIGNAL-TTT (right) fits closed-form per-(event, instrument) offsets to correct cross-sectional orderings and modify portfolio legs.

answer about a policy event is that nothing moves, which cannot be traded. ❸ The evidence that matters is instrument-specific: how the frozen score has been mis-centered for *this* instrument on similar past events. A correction shared by the event cannot use it.

In this work, we argue that cross-sectional test-time adaptation requires moving the fitted parameter from the test input to the (event, instrument) cell. We introduce **SIGNAL-TTT**, a closed-form cross-sectional test-time training method discovery event driven market prediction. The core innovation of SIGNAL-TTT is that, once the parameter is indexed by (event, instrument), the test-time objective becomes a one-variable ridge problem whose exact solution needs no learning rate, no step count and no backward pass. No parameter of the language model is read or modified. SIGNAL-TTT operates at two levels. **At the instrument level**, it retrieves the past events whose text resembles the query and fits, for each instrument, an offset equal to the mean residual of the frozen score on their matured outcomes. **At the evidence level**, it shrinks each offset by empirical Bayes: an instrument whose neighbors disagree is shrunk hard, and an event whose instruments genuinely differ is shrunk little. We show that the update is exactly a shrunk per-instrument residual, and we derive the condition under which applying it lowers squared error (Proposition 2). On a development window of 321 events with 9 frozen backbones, SIGNAL-TTT improves rank IC by +0.0156 and the net return of a dollar-neutral book by +2.32 bps per event over the same frozen backbone.

Our main contributions are summarized as follows: ❶ **First Quantitative Trump Trade Dataset.** To the best of our knowledge, we present the first quantitative dataset that couples President Donald Trump's Truth Social posts with granular market outcomes. This benchmark captures the unique market-moving dynamics of high-profile political announcements during a presidential tenure, providing a novel testbed for financial text-driven task. ❷ **Closed-Form Cross-Sectional TTT.** We formulate cross-sectional test-time adaptation and prove that standard per-event corrections preserve within-event orderings. To overcome this, we propose SIGNAL-TTT, a gradient-free method that computes per-(event, instrument) offsets with empirical Bayes shrinkage. SIGNAL-TTT consistently achieves state-of-the-art ranking performance and universally boosts the performance of diverse frozen backbones. ❸ **Realistic Evaluation Protocol.** We establish a rigorous cross-sectional TTA evaluation protocol that compares SIGNAL-TTT against 12 baselines under validation-selected configurations. The protocol strictly enforces dollar-neutral constraints, accounts for transaction costs on entry and exit, and evaluates multi-backbone inference across identical event streams.

2 RELATED WORK

Test-Time Adaptation and Training updates a deployed model on post-training data (Liang et al., 2025). Label-free methods adapt normalization statistics (TENT; Wang et al., 2021), filter samples

by entropy (EATA; Niu et al., 2022), add sharpness-aware updates (Niu et al., 2023), or update non-parametric caches without gradients (T3A, TDA; Iwasawa & Matsuo, 2021; Karmanov et al., 2024). TTAB (Zhao et al., 2023a) shows these gains are highly sensitive to protocol choices. Test-time training fine-tunes on nearest neighbors (Sun et al., 2020; Hardt & Sun, 2024; Akyürek et al., 2025). Time-series variants recalibrate on partially observed labels (Kim et al., 2025), retrain output layers (Chen et al., 2024), model label delay (Zhao & Shen, 2025), or meta-learn adapters (Zhao et al., 2023b). As summarized in Table 1, these methods index updates by the test input or series, failing to reorder items. SIGNAL-TTT indexes parameter updates by the (event, instrument) pair as the only granularity capable of modifying cross-sectional relative orderings.

Table 1: Cross-sectional TTA properties of SIGNAL-TTT versus closely related baseline methods.

Method	Frozen	Outcome-Sup.	Closed-Form	Reorders	Order-Invar.
TENT (Wang et al., 2021)	✗	✗	✗	✗	✗
EATA (Niu et al., 2022)	✗	✗	✗	✗	✗
SAR (Niu et al., 2023)	✗	✗	✗	✗	✗
T3A (Iwasawa & Matsuo, 2021)	✓	✗	✓	✗	✗
TDA (Karmanov et al., 2024)	✓	✗	✓	✗	✗
TTT (Hardt & Sun, 2024)	✗	✓	✗	✗	✓
AdaNPC (Zhang et al., 2023)	✓	✓	✓	✗	✗
SIGNAL-TTT (ours)	✓	✓	✓	✓	✓

Frozen: no weight modification/access; **Outcome-Sup.**: uses realized market labels; **Closed-Form**: no backward pass; **Reorders**: fits per-(event, instrument) offsets to reorder items; **Order-Invar.**: robust to test input arrival order.

Retrieval and Calibration of Language Models adjust outputs using retrieved labels or estimated biases. kNN-LM (Khandelwal et al., 2020) interpolates predictions with retrieved context labels via calibrated weights (Xu et al., 2023; Drozdov et al., 2022). AdaNPC (Zhang et al., 2023) maintains a non-parametric memory at test time. Contextual and batch calibrations remove label biases estimated from dummy inputs or test batches (Zhao et al., 2021; Zhou et al., 2024), while label-shift methods reweight class posteriors (Saerens et al., 2002; Lipton et al., 2018; Alexandari et al., 2020). When a batch equals one event, batch calibration collapses to an additive per-event shift, rendered inert by Proposition 1. Prior reweighting is non-additive yet shared, making it insufficient for per-instrument fitting (Remark 2). Unlike prior approaches, SIGNAL-TTT fits a separate correction for each instrument, on matured outcomes rather than on the test stream.

Language Models and the Cross-Section of Returns are increasingly integrated. Headline-prompted models predict short-term returns (Lopez-Lira & Tang, 2026), while cross-sectional ranking is standard in financial ML (Gu et al., 2020; Feng et al., 2019), especially as political shocks are priced rapidly (Tsang & Yang, 2026). Look-ahead contamination from pre-training (Glasserman & Lin, 2023; Sarkar & Vafa, 2024; Lopez-Lira et al., 2025) is controlled via backbone release dates (Appendix E). Our shrinkage draws on empirical Bayes and credibility theory (Efron & Morris, 1973; Bühlmann, 1967). Crucially, rather than training a return predictor, SIGNAL-TTT corrects a frozen backbone's scores at test time to optimize cross-sectional performance.

3 PRELIMINARY

In this work, we tackle event-driven asset ranking via efficient test-time training to learn how complex policy announcements. Donald Trump's Truth Social posts impact diverse financial instruments. Given a frozen language model that scores raw (event, instrument) pairs, our goal is to adaptively reorder the cross-section to reflect realized relative returns using matured historical outcomes. To overcome the rank-invariance limitations of standard TTA without incurring heavy gradient computation, we introduce SIGNAL-TTT leverages closed-form, per-(event, instrument) offset updates to capture heterogeneous event-to-asset market transmission, substantially improving both computational speed and market forecasting efficiency.

Problem Definition. Let $\mathcal{A}$ be an instrument universe with $|\mathcal{A}| = 39$, trading in 15-minute bars. A policy post published at time t_e is aligned to the first bar opening at or after t_e, and the posts aligned to bar b form one *event*. With a horizon of $H = 4$ bars (one hour), the response of instrument a is its open-to-close return $r_{b,a}$. The target is the cross-sectionally demeaned return

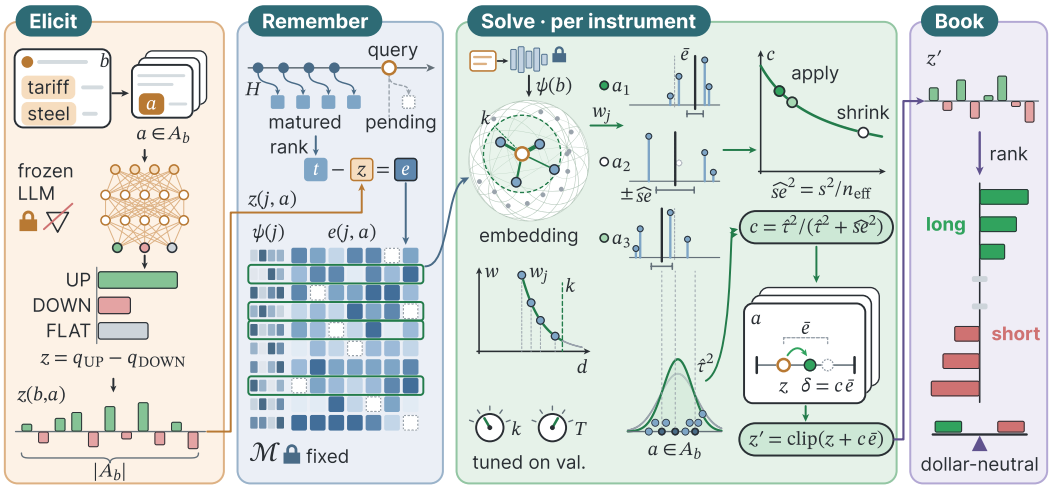

Figure 2: Overview of SIGNAL-TTT. **Elicit**: A frozen LLM scores each (event, instrument) pair. **Remember**: A memory bank stores historical residual errors from matured events. **Solve**: For each instrument, kernel-weighted residuals are computed from nearest past events and shrunk via empirical Bayes to update scores. **Book**: The adapted scores form a dollar-neutral long–short portfolio.

$\tilde{r}_{b,a} = r_{b,a} - \frac{1}{|A_b|}\sum_{c\in A_b} r_{b,c}$, where $A_b \subseteq \mathcal{A}$ is the set of instruments with a complete window at event b. An event enters the panel only if $|A_b| \geq 20$. For each (event, instrument) pair, a frozen language model is prompted with the post and the instrument and asked whether the instrument moves UP, DOWN or FLAT over the next hour relative to the market. Generation stops after one token, and the three answer probabilities are read from the next-token distribution.

$$z(b,a) \;=\; q_{\mathrm{UP}}(b,a) - q_{\mathrm{DOWN}}(b,a) \;\in\; [-1,1], \tag{1}$$

Equation 1 is the frozen score, sign is the direction call and whose magnitude is the model's conviction. *Cross-sectional test-time adaptation* produces an adapted score $z'(b,\cdot)$, within-event ordering agrees better with $\tilde{r}(b,\cdot)$. It may use only information available at the bar open and a memory $\mathcal{M}$ of past events whose outcomes have matured.

Windows. The panel is split chronologically. Events before 2025-07-01 fit every transform, and events before 2025-10-01 form the memory $\mathcal{M}$, which is fixed and never grows. The validation window (2025-10-01 to 2025-12-31) is the only surface on which any hyper-parameter or design choice is made. We report results on the development window (2026-01-02 to 2026-04-11). It was examined repeatedly while the method was developed, so it is a development window, not a held-out test. Labels mature one hour after the bar opens and are purged at every boundary.

Evaluation Metric. The claim concerns ordering *within* an event, so we score each event separately and weight events equally:

$$\mathrm{IC}_b \;=\; \rho_{\mathrm{Spearman}}\big(z'(b,\cdot),\, \tilde{r}(b,\cdot)\big) \text{ over } a \in A_b, \qquad \overline{\mathrm{IC}} \;=\; \tfrac{1}{|B|}\textstyle\sum_{b\in B} \mathrm{IC}_b, \tag{2}$$

where an event contributes only if at least 5 instruments are priced and the score is not constant across them. Since ρ depends on $z'(b,\cdot)$ only through ranks, an adapter can improve $\overline{\mathrm{IC}}$ only by changing which instruments outrank which. All 9 backbones score the same events, so every test in this paper averages over backbones within an event before testing across events.

From Ranking to a Book. The scores are traded directly. Within event b the book goes long the instruments with $z' > 0.20$ and short those with $z' < -0.20$, with half the capital on each side, so it is dollar-neutral with equal dollars per side. An event on which only one side clears the margin is not traded. Positions are entered at the bar open and exited H bars later. A cost of γ bps per leg is charged on both entry and exit, so an event's net return is $\sum_a w_{b,a}\tilde{r}_{b,a} - 2\gamma\sum_a |w_{b,a}|$ with $\sum_a |w_{b,a}| = 1$. This book reads score levels through the margin as well as the ordering. Proposition 1 therefore applies to the rank metric and to books built from the ordering alone. A shared per-event shift can still move instruments across the margin, and so change which of them are traded.

4 METHODOLOGY

The core idea of SIGNAL-TTT is to index the fitted parameter by (event, instrument), which turns the test-time objective into a one-variable problem with an exact solution. As illustrated in Figure 2, SIGNAL-TTT then comprises three main components: ❶ *Matured-outcome retrieval.* The past events nearest the query are retrieved in a frozen text-embedding space, from a memory whose outcomes had all matured before the query bar. ❷ *Per-instrument closed-form fit.* For each instrument of the query event, a scalar offset is fitted by ridge-regularized least squares to the frozen score's residuals on the neighbors. ❸ *Empirical-Bayes shrinkage.* Each offset is shrunk by the noise of its own neighborhood relative to how much the event's instruments differ, both estimated from the retrieved data.

4.1 PER-EVENT CORRECTIONS: AN ORDERING INVARIANCE

Standard adaptation methods fit one parameter per test input, applying a per-event mapping ϕ_b across all instruments: $z'(b,a) = \phi_b(z(b,a))$. Common examples include additive shifts $\phi_b(z) = z + \delta_b$ (e.g., retrieval adaptation, batch calibration) and positive score rescalings.

Proposition 1 (Per-event strictly increasing corrections preserve within-event order)**.** *For any event b and strictly increasing $\phi_b : \mathbb{R} \to \mathbb{R}$, the within-event ordering of $\phi_b(z(b,\cdot))$ equals that of $z(b,\cdot)$. Consequently, rank IC (Eq. (2)) and any portfolio constructed solely from relative rankings (e.g., quantile or rank-weighted legs) remain unchanged.*

Score clipping to $[-1, 1]$ is the sole mechanism by which a shared shift alters metrics—by creating ties that destroy ordering rather than refining it (Remark 1). Empirically, across 14,508 (backbone, event) cells, an unclipped shared update yields 0 rank reversals. (Appendix B)

Label-shift updates $q'_y \propto q_y\, \pi^b_y/\pi_y$ (Saerens et al., 2002) integrate q_{FLAT} with z, enabling reordering based on FLAT mass rather than instrument-specific histories. Though not covered by Proposition 1, these updates remain shared event-level mappings rather than per-instrument fits.

Takeaway: Per-event corrections (whether additive or monotonic) cannot alter within-event rankings or downstream portfolios. Effective cross-sectional TTA must fit parameters that vary *across instruments within* each event.

4.2 MATURED-OUTCOME RETRIEVAL: NEIGHBORS IN TEXT SPACE

For a target event b, SIGNAL-TTT retrieves its k-nearest matured events $\mathcal{N}_k(b)$ from memory $\mathcal{M}$ using frozen text embeddings ψ. Neighbors are weighted by relative distance: $w_{bj} = \exp\{-(d_{bj} - d_b^{\min})/T\}$. Let $m_{j,a} \in \{0,1\}$ indicate whether instrument a has an outcome at event j.

Signed Rank Target. To align the adaptation objective with ranking metrics, SIGNAL-TTT defines the target as the normalized within-event return rank, $t(j,a) = 2\,\mathrm{rank}_a(\tilde{r}_{j,a})/(|A_j|-1) - 1 \in [-1,1]$. The frozen score residual on neighbor j is then $e_j = t(j,a) - z(j,a)$. These matured residuals serve as direct test-time evidence of systematic instrument-specific miscalibration (Figure 3).

4.3 PER-INSTRUMENT CLOSED-FORM FIT

To enable cross-sectional reordering, SIGNAL-TTT fits an instrument-specific offset $\delta(b,a)$ for each target (event, instrument) pair via ridge-regularized weighted least squares:

$$\delta(b,a) \;=\; \arg\min_{\delta\in\mathbb{R}} \sum_{j\in\mathcal{N}_k(b)} w_{bj}\, m_{j,a}\, \big(z(j,a)+\delta-t(j,a)\big)^2 \;+\; \lambda(b,a)\,\delta^2 \;=\; \frac{W}{W+\lambda(b,a)}\,\bar{e}(b,a), \tag{3}$$

where $W = \sum_j w_{bj} m_{j,a}$ is the effective neighbor mass and $\bar{e}(b,a) = \sum_j w_{bj} m_{j,a}\, e_j / W$ is the weighted mean residual. Because Eq. (3) is a strictly convex univariate quadratic, its solution is exact and closed-form, requiring no iterative optimization, gradients, or weight access. Crucially, as $\delta(b,a)$ varies across a, the update directly modifies the within-event instrument rankings.

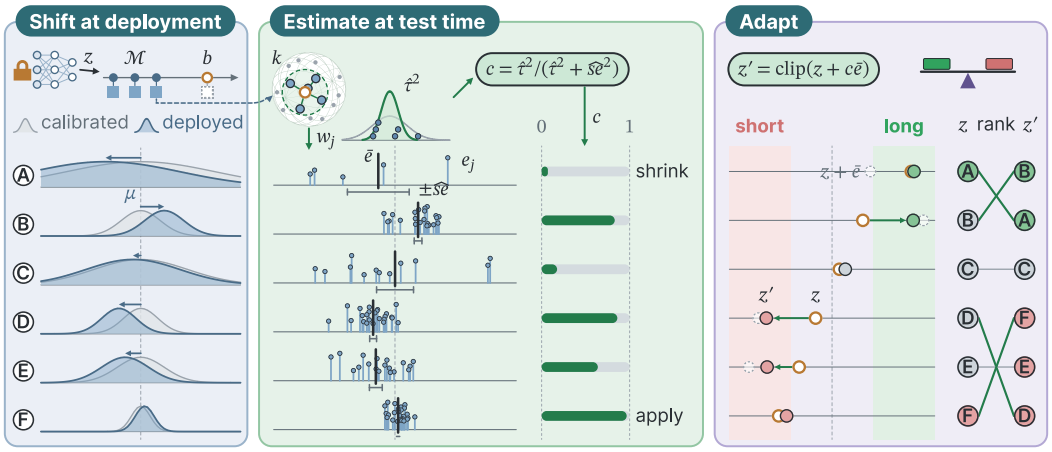

Figure 3: Adaption Mechanism of SIGNAL-TTT. **Shift**: Deployed scores suffer from asset-specific miscalibration. **Estimate**: Historical residuals are weighted by similarity and shrunk via empirical Bayes to filter noise. **Adapt**: Shrunk offsets reorder instruments and update the long-short portfolio.

4.4 EMPIRICAL-BAYES SHRINKAGE

Rather than tuning a static regularizer, SIGNAL-TTT estimates $\lambda(b,a)$ dynamically from retrieved neighbor data using empirical Bayes (Efron & Morris, 1973). Let $\widehat{se}^{\,2}(b,a) = s^2(b,a)/n_{\text{eff}}$ denote the squared standard error of $\bar{e}(b,a)$, where s^2 is the weighted residual variance and $n_{\text{eff}} = W^2/\sum_j w_{bj}^2 m_{j,a}$ is the effective sample size. The cross-sectional variance of the true offsets is estimated via the method of moments:

$$\hat{\tau}^2(b) \;=\; \max\Big\{0,\; \mathrm{Var}_a\,\bar{e}(b,a) - \mathrm{mean}_a\,\widehat{se}^{\,2}(b,a)\Big\}. \tag{4}$$

Setting $\lambda(b,a) = W\,\widehat{se}^{\,2}(b,a)/\hat{\tau}^2(b)$ in Eq. (3) yields the closed-form shrinkage gate $c(b,a)$ and adapted score $z'(b,a)$:

$$\delta(b,a) \;=\; c(b,a)\,\bar{e}(b,a), \qquad c(b,a) \;=\; \frac{\hat{\tau}^2(b)}{\hat{\tau}^2(b) + \widehat{se}^{\,2}(b,a)}, z'(b,a) = \mathrm{clip}\big(z(b,a)+\delta(b,a), -1, 1\big). \tag{5}$$

The shrinkage mechanism naturally balances noise and signal: large standard errors $\widehat{se}$ pull noisy updates toward zero ($c \to 0$), whereas strong cross-sectional variance $\hat{\tau}^2$ preserves asset-specific adjustments ($c \to 1$). When cross-sectional signal is absent ($\hat{\tau}^2 = 0$), SIGNAL-TTT seamlessly defaults to the unadapted frozen score z. Hyperparameters k and T are tuned on validation (Algorithm 1).

4.5 THEORETICAL ANALYSIS: ERROR REDUCTION BOUNDS

Equation (5) shows that SIGNAL-TTT applies a weighted-mean residual $\bar{e}$ shrunk by $c \in [0,1]$, which mirrors the classic credibility weighting of Bühlmann (1967) estimated at the event level. We formalize the precise condition under which this adjustment reduces score estimation error:

Proposition 2 (Error Reduction Condition)**.** *Fix an (event, instrument) pair and $c \in (0,1]$. Assume retrieved residuals are conditionally unbiased for a systematic score miscentering μ, with conditional variance σ^2 independent across neighbors, and effective sample size $n_{\text{eff}} = W^2/\sum_j w_{bj}^2 m_{j,a}$. Applying $c\,\bar{e}$ strictly reduces the expected squared error of the score if and only if*

$$\frac{\mu^2}{\sigma^2} \;>\; \frac{c}{(2-c)\,n_{\text{eff}}}. \tag{6}$$

The bound decreases with n_{eff} as $\mathcal{O}(1/n_{\text{eff}})$, guaranteeing error reduction whenever systematic miscentering μ dominates local variance. Conversely, when neighbor evidence is sparse or noisy (n_{eff} small or σ^2 large), empirical-Bayes shrinkage drives $c \to 0$, safeguarding the score from noisy updates (Appendix B.2).

Takeaway: SIGNAL-TTT lowers score error whenever systematic miscentering outweighs local retrieval noise. When evidence is sparse or discordant, empirical Bayes shrinks the correction, recovering the unadapted frozen baseline.

Computational Complexity. At inference, SIGNAL-TTT executes one k-NN query per event ($\mathcal{O}(|\mathcal{M}|d)$ for d-dimensional embeddings) and a closed-form $\mathcal{O}(k)$ update per instrument. Requiring zero backward passes or parameter updates, it integrates seamlessly with black-box API-hosted models. Because memory $\mathcal{M}$ is fixed prior to evaluation, predictions are inherently order-invariant.

5 EXPERIMENT

5.1 EXPERIMENTAL SETTINGS

Datasets & Protocols. The dataset pairs policy events with 15-minute returns across 39 instruments (1,612 events total; 321 in evaluation window). Appendix D details protocol specifics. Based on Trump post and SPY500 stock.

Baselines. We benchmark SIGNAL-TTT against 12 test-time baselines across key paradigms: (1) label-free test-time adapters (TENT, EATA, SAR, T3A, TDA); (2) matured-outcome retrieval baselines (kNN-LM, AdaNPC, kNN vote, Nadaraya–Watson regression); (3) the shared per-event offset; (4) gradient-based TTT on our objective; and (5) a per-instrument static affine calibration fitted on memory. All baselines select hyperparameters on validation (Appendix D).

Implementation & Evaluation. We evaluate 9 frozen backbones (8B–36B parameters, 7 instruction-tuned, 2 base). Text embeddings are reduced to 64 principal components fitted on train. SIGNAL-TTT tunes only k and T on validation, with no other hyperparameters. We report event-weighted rank IC (Eq. (2)) and net book return, evaluating statistical significance via one-sided 5-day block sign-flip permutation tests and moving-block bootstrap intervals.

5.2 EXPERIMENT RESULTS

Performance Comparison.

We compare SIGNAL-TTT with the frozen backbone and 12 test-time operators in Table 2 and Figure 4, and draw four observations. ❶ **SIGNAL-TTT is the best operator on both metrics.** It has the highest net return per traded event and the highest rank IC of all operators in Table 2. Against the same backbone frozen, it raises net return by +2.32 bps per event ($[+1.30, +3.09]$; $p = 0.0001$) and rank IC by +0.0156 ($p = 0.008$). ❷ **Label-free adaptation does not move the score.** The 5 label-free adapters stay at the frozen level (best change in rank IC +0.0001), and SIGNAL-TTT is ahead of every one of them after Holm correction. Entropy- and prototype-based updates have no access to realised outcomes, so they cannot correct a score whose error is in its direction. ❸ **SIGNAL-TTT leads the retrieval operators on every comparison.** Its paired difference from each retrieval operator is positive (Figure 4c), and its curve is the highest at every cost in Figure 4b, so the ranking does not depend on the cost assumption. These differences do not survive Holm correction. The retrieval operators come close on rank IC, which scores the whole cross-section, while the book trades only its tails; SIGNAL-TTT gains more on the tails, which is where net return is decided. ❹ **The gain holds across backbones.** In Figure 4a, SIGNAL-TTT is above the frozen backbone on all 9 backbones, while each rival leads on at most a few. At 2 bps per

Table 2: Gain of SIGNAL-TTT over baseline operators on the development window. p-values are Holm-corrected over rivals. Backbone: Seed-OSS-36B.

operator	Δ net bps ↑	p-val	Rank IC ↑
frozen (no adaptation)	+2.32	<0.001	0.0030
TENT	+2.41	0.001	0.0021
EATA	+2.37	0.001	0.0030
SAR	+2.46	0.001	0.0021
T3A	+2.49	0.004	0.0026
TDA	+2.20	0.001	0.0031
shared per-event offset	+1.10	0.254	0.0028
AdaNPC	+1.28	0.169	0.0072
kNN-LM	+0.57	0.921	0.0178
Nadaraya–Watson	+2.40	0.404	0.0162
kNN vote	+0.94	0.921	0.0173
gradient TTT (retrieved)	+0.40	0.841	0.0148
SIGNAL-TTT	–	–	**0.0186**

Table 3: Per-backbone comparison of SIGNAL-TTT with the same backbone frozen on the development window, ordered by size ([B]: base model). Net return is per traded event at 2 bps per leg, reported relative to the worst frozen backbone ([†], set to zero). The paired difference counts untraded events as zero and carries a 95% five-day block-bootstrap interval. Bold marks the better.

		rank IC ↑			net bps vs. worst frozen ↑		
backbone	params	frozen	SIGNAL-TTT	Δ	frozen	SIGNAL-TTT	paired Δ [95% CI]
`Qwen3-8B`	8B	-0.0018	**0.0002**	+0.0020	+3.25	**+5.62**	+1.70 [+0.05, +3.63]
`Qwen3.5-9B`	9B	0.0018	**0.0223**	+0.0205	+8.06	**+9.40**	+1.11 [-1.49, +3.16]
`Qwen3.5-9B-Base`[B]	9B	-0.0186	**0.0150**	+0.0336	+4.97	**+11.33**	+4.84 [+2.39, +7.05]
`Qwen3-14B`	14B	-0.0120	**0.0220**	+0.0340	0.00[†]	**+6.88**	+3.66 [+0.87, +6.59]
`Qwen3.6-27B`	27B	0.0058	**0.0149**	+0.0091	+7.62	**+10.57**	+1.95 [-1.85, +5.40]
`Qwen3.8-27B`	27B	0.0221	**0.0299**	+0.0078	+15.02	**+15.70**	+0.17 [-0.57, +1.13]
`Qwen3.5-35B-A3B-Base`[B]	35B	0.0103	**0.0222**	+0.0119	+9.60	**+13.50**	+1.94 [-1.15, +4.61]
`Qwen3.6-MoE-35B`	35B	0.0042	**0.0241**	+0.0199	+4.10	**+11.37**	+3.90 [+1.09, +6.13]
`Seed-OSS-36B`	36B	0.0149	**0.0267**	+0.0118	+10.41	**+14.33**	+1.61 [-1.12, +3.61]
mean		0.0030	**0.0186**	+0.0156	+7.00	**+10.97**	+2.32

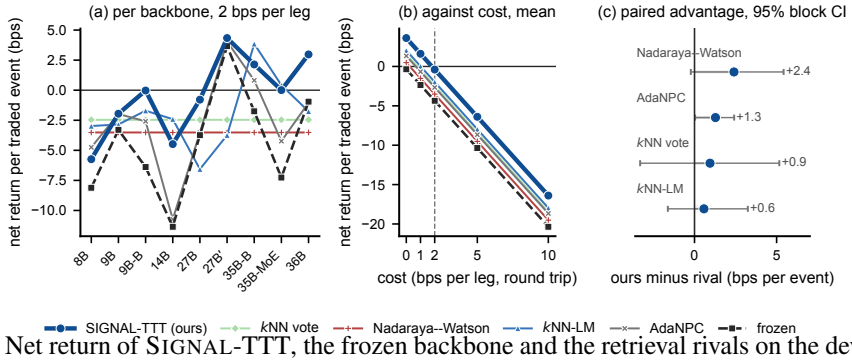

Figure 4: Net return of SIGNAL-TTT, the frozen backbone and the retrieval rivals on the development window. **(a)** Net return per traded event for each backbone, ordered by size, at 2 bps per leg. kNN vote and Nadaraya–Watson do not use the frozen score, so their value is the same for every backbone. **(b)** Net return against the cost per leg, charged on entry and exit, mean over backbones; the dashed line marks the cost used in the rest of the paper. **(c)** Paired difference between SIGNAL-TTT and each rival in bps per event, with 95% block-bootstrap intervals.

leg the book is still below zero in absolute terms (-0.39 against -4.36 bps per traded event for frozen), and the per-instrument affine calibration is the one operator above zero.

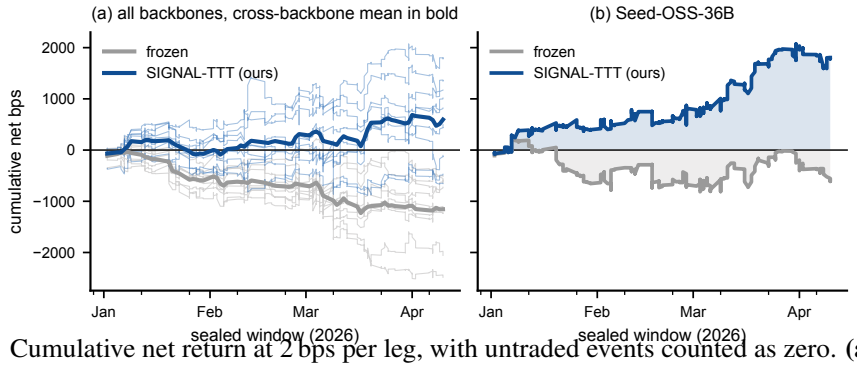

Figure 5: Cumulative net return at 2 bps per leg, with untraded events counted as zero. **(a)** Frozen (grey) and SIGNAL-TTT (blue) for every backbone, with the cross-backbone mean in bold. **(b)** `Seed-OSS-36B`, with the area between each curve and zero shaded.

Per-Backbone Analysis. Table 3 and Figure 5 break the gain down by backbone. SIGNAL-TTT raises net return per traded event on all 9 backbones and rank IC, with a mean paired gain of +2.32 bps per event. The paired interval excludes zero on 4 backbones; on the remaining ones the point estimate is positive but the interval includes zero. Notably, the performance scales positively with stronger base capability: backbones with higher frozen rank IC generally achieve stronger absolute performance under SIGNAL-TTT, reaching up to 0.0299 rank IC and +15.70 bps net return on `Qwen3.8-27B`.

The gains hold consistently across base models, dense architectures, and mixture-of-experts (MoE) backbones. Over time, the frozen backbones lose steadily at this cost, while SIGNAL-TTT loses much less (Figure 5a). The advantage over frozen builds up across the whole window rather than coming from a few events (Figure 5b).

Ablation Analysis. Table 4 isolates each design component: **Cross-Sectional vs. Shared Updates.** Replacing per-instrument updates with a shared per-event offset collapses rank IC to frozen levels (0.0027 vs. 0.0030), validating Proposition 1. Target binarization yields marginal difference (0.0192). **Empirical-Bayes Shrinkage.** On the development window, disabling shrinkage gives a higher rank IC (0.0245 vs. 0.0186), but it lowers validation rank IC (0.0409 vs. 0.0461), the only window used for design choices, so shrinkage is kept. Empirical-Bayes shrinkage matches optimal tuned fixed shrinkage (0.0186 vs. 0.0181) without adding hyperparameters. **Retrieval Mechanisms.** Random matured events capture only 17% of the gain (0.0057), while market-state retrieval performs below frozen (0.0015), confirming that text similarity is essential. Event-permutation controls verify that gains stem from semantic alignment rather than static recalibration. **Entropy Minimization Baseline Failure.** Label-free adapters collapse because entropy minimization drives predictions toward the non-tradable FLAT outcome (Appendix C).

Table 4: Ablation of SIGNAL-TTT.

variant (development window)	rank IC
frozen	0.0030
one shared offset per event	0.0027
binary target	0.0192
no shrinkage	0.0245
fixed shrinkage tuned on validation	0.0181
uniform random memory	0.0057
market-state neighbors	0.0015
SIGNAL-TTT	**0.0186**

Sensitivity Analysis. Table 5 varies one setting at a time; rank IC stays above frozen in every row. From 16 to 128 PCA dimensions rank IC lies between 0.0186 and 0.0244, and it falls only at 8 dimensions. Rank IC is flat in the number of neighbors, so the result does not depend on a tuned k. The gain is largest at the elicited horizon and at twice it, and it persists, smaller, at one trading day. With the memory's outcomes permuted within each instrument, rank IC is still 0.0149, because the permutation keeps each instrument's outcome distribution. The gain is present in both halves of the development window (+0.0110 and +0.0203), adapted scores are identical under every stream order, and mean rank IC varies by 0.0006 across BLAS thread counts. Full results, including the neighbor space and the operator form, are in Appendix C.2.

Table 5: Sensitivity of SIGNAL-TTT.

	setting	rank IC	Δ
embedding dimension	16	0.0244	+0.0215
	64*	**0.0186**	**+0.0156**
	128	0.0188	+0.0158
neighbors k	32	0.0197	+0.0167
	48	0.0184	+0.0154
	64	0.0187	+0.0157
horizon	elicited*	0.0186	+0.0156
	2×	0.0235	+0.0165
	one day	−0.0173	+0.0046
memory	fully permuted	0.0149	+0.0119

Δ is the change in rank IC from frozen at the same setting. *: registered setting.

5.3 LIMITATIONS

Macro-level political shocks (Trump post) are inherently sparse in historical logs, leaving space to test adaptation under rare regime shifts. Second, our current setup targets SPY constituent dynamics; extending the framework to cross-asset and global markets offers a promising extension. Evaluating adaptation across multi-year news cycles will further characterize long-term robustness (Appendix E).

6 CONCLUSION

This paper introduces SIGNAL-TTT for cross-sectional test-time adaptation. We prove that per-event corrections strictly preserve within-event rank orderings and cannot realign cross-sectional predictions. SIGNAL-TTT resolves this by indexing updates by (event, instrument), solving a closed-form objective on retrieved matured outcomes, and regulating adjustment via empirical Bayes shrinkage. Theoretically, we establish exact error-reduction bounds. Empirically, SIGNAL-TTT consistently boosts rank IC and net returns across diverse LLM backbones over unadapted baselines, whereas per-event baseline offsets eliminate all gains. SIGNAL-TTT provides a lightweight, gradient-free method to transform past matured outcomes into actionable, instrument-specific test-time evidence, without reading or modifying any backbone weights.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to provide feedback on research methodology and experiments. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, or interpret results. Translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, such as experiment and evaluation scripts, and to edit the text of the paper for readability. We have reviewed all AI-assisted work. All AI-assisted code was read, executed, and tested by the authors, and all reported results were checked against the saved experiment logs. Feedback from AI tools on methodology and experiments was treated as suggestions; the research questions, method design, and conclusions were determined by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of AI.

ETHICS STATEMENT

This research adheres to the ICLR Code of Ethics. The panel is built from public market data and public policy communications. The method is a research artifact, not investment advice.

REPRODUCIBILITY STATEMENT

The algorithm is in Appendix A and the proofs in Appendix B. The source code in https://anonymous.4open.science/r/SignalTTT

REFERENCES

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. In *International Conference on Machine Learning (ICML)*, 2025.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/alexandari20a.html.

Hans Bühlmann. Experience rating and credibility. *ASTIN Bulletin*, 4(3):199–207, July 1967. ISSN 1783-1350. doi: 10.1017/s0515036100008989. URL http://dx.doi.org/10.1017/s0515036100008989.

Mouxiang Chen, Lefei Shen, Han Fu, Zhuo Li, Jianling Sun, and Chenghao Liu. Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 341–352. ACM, August 2024. doi: 10.1145/3637528.3671926. URL http://dx.doi.org/10.1145/3637528.3671926.

Andrew Drozdov, Shufan Wang, Razieh Rahimi, Andrew McCallum, Hamed Zamani, and Mohit Iyyer. You can't pick your neighbors, or can you? when and how to rely on retrieval in the kNN-LM. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2997–3007, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.218. URL https://aclanthology.org/2022.findings-emnlp.218/.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, March 1973. ISSN 1537-274X. doi: 10.1080/01621459.1973.10481350. URL http://dx.doi.org/10.1080/01621459.1973.10481350.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2):1–30, March 2019. ISSN 1558-2868. doi: 10.1145/3309547. URL http://dx.doi.org/10.1145/3309547.

Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *arXiv preprint arXiv:2309.17322*, 2023.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, February 2020. ISSN 1465-7368. doi: 10.1093/rfs/hhaa009. URL http://dx.doi.org/10.1093/rfs/hhaa009.

Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Training-free Dynamic Adapter; arXiv:2403.18293.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*, 2020.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965. arXiv:2501.04970.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133:31–64, 2025. doi: 10.1007/s11263-024-02181-w.

Robert Lindahl. Tariff communication and the repricing of U.S. sovereign risk. SSRN Working Paper 5601612, Social Science Research Network, 2025. URL https://ssrn.com/abstract=5601612. Available at SSRN.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? return predictability and large language models. *Journal of Financial Economics*, 2026. doi: 10.1016/j.jfineco.2026.104335. Article 104335.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust llms' economic forecasts? *SSRN Electronic Journal*, 2025. doi: 10.2139/ssrn.5217505. URL http://dx.doi.org/10.2139/ssrn.5217505.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a *priori* probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, January 2002. ISSN 1530-888X. doi: 10.1162/089976602753284446. URL http://dx.doi.org/10.1162/089976602753284446.

Suproteem Sarkar and Keyon Vafa. Lookahead bias in pretrained language models. *SSRN Electronic Journal*, 2024. doi: 10.2139/ssrn.4754678. URL http://dx.doi.org/10.2139/ssrn.4754678.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9229–9248, 2020.

Kwok Ping Tsang and Zichao Yang. Political shocks and price discovery in prediction markets: Evidence from the 2024 U.S. presidential election. *arXiv preprint arXiv:2603.03152*, 2026.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Frank F. Xu, Uri Alon, and Graham Neubig. Why do nearest neighbor language models work? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38325–38341. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23a.html.

Tian Yue, Zhixian Huang, and Jiexiang Huang. Large Language Model, trade threats and stock markets. *Economics Letters*, 268:113130, 2026. ISSN 0165-1765. doi: 10.1016/j.econlet.2026.113130. URL https://doi.org/10.1016/j.econlet.2026.113130.

Yi-Fan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. AdaNPC: Exploring non-parametric classifier for test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41647–41676. PMLR, 2023.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023a. URL https://proceedings.mlr.press/v202/zhao23d.html.

Lifan Zhao and Yanyan Shen. Proactive model adaptation against concept drift for online time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 2020–2031. ACM, July 2025. doi: 10.1145/3690624.3709210. URL http://dx.doi.org/10.1145/3690624.3709210.

Lifan Zhao, Shuming Kong, and Yanyan Shen. Doubleadapt: A meta-learning approach to incremental learning for stock trend forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3492–3503. ACM, August 2023b. doi: 10.1145/3580305.3599315. URL http://dx.doi.org/10.1145/3580305.3599315.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.17249.

APPENDIX CONTENTS

A Algorithm **14**

B Proofs **14**

B.1 Proof of Proposition 1 14

B.2 Proof of Proposition 2 14

C Additional Experimental Results **15**

C.1 Baselines Against the Frozen Score 15

C.2 Robustness Analysis 15

D Additional Implementation Details **18**

D.1 Splits 18

D.2 Event Filter 18

D.3 Retrieval and Selection 18

D.4 Book and Costs 18

D.5 Inference 18

D.6 Porting the Label-Free Adapters 18

E Scope and Limitations **18**

E.1 Backbone Release Dates 18

E.2 Cross-Sections Without Differential Exposure 19

A ALGORITHM

This section gives the full procedure of SIGNAL-TTT and its cost. The instrument a enters only

Algorithm 1 SIGNAL-TTT: closed-form, empirical-Bayes-shrunk offset per (event, instrument)

Require: frozen scores z; memory $\mathcal{M}$ of events with matured outcomes; query event b with instruments A_b; k, T
Ensure: adapted scores $z'(b, \cdot)$
$\mathcal{N}_k(b) \leftarrow$ the k events of $\mathcal{M}$ nearest b in the frozen embedding space; $w_{bj} \leftarrow \exp\{-(d_{bj} - d_b^{\min})/T\}$
for $a \in A_b$ **do** ▷ one independent scalar problem per instrument
  $e_j \leftarrow t(j,a) - z(j,a)$ on neighbors where a matured; W, $\bar{e}(b,a)$, $s^2(b,a)$, n_{eff} as in Sections 4.3 and 4.4
  $\widehat{se}^{\,2}(b,a) \leftarrow s^2(b,a)/n_{\text{eff}}$ ▷ noise of this instrument's neighborhood
end for
$\hat{\tau}^2(b) \leftarrow \max\{0, \mathrm{Var}_a\, \bar{e}(b,a) - \mathrm{mean}_a\, \widehat{se}^{\,2}(b,a)\}$ ▷ spread of the offsets across the event
for $a \in A_b$ **do**
  $z'(b,a) \leftarrow \mathrm{clip}\big(z(b,a) + \bar{e}(b,a)\,\hat{\tau}^2(b)/(\hat{\tau}^2(b) + \widehat{se}^{\,2}(b,a)),\, -1,\, 1\big)$ ▷ Eq. (5)
end for
return $z'(b, \cdot)$; an instrument with no matured neighbor keeps $z(b,a)$

inside the loop. Collapsing that loop to one shared δ_b recovers exactly the operator of Proposition 1, which is the "shared per-event offset" arm of Table 4. **Complexity.** In addition to the cost given in Section 4.5, computing $\hat{\tau}^2(b)$ takes one $O(|A_b|)$ pass per event.

B PROOFS

This section proves Propositions 1 and 2.

B.1 PROOF OF PROPOSITION 1

Fix an event b, any two instruments $a, a' \in A_b$ and a strictly increasing ϕ_b. Then $z(b,a) > z(b,a')$ if and only if $\phi_b(z(b,a)) > \phi_b(z(b,a'))$, and equality is preserved as well, so every pairwise comparison is unchanged. A total preorder on a finite set is determined by its pairwise comparisons, so the within-event ordering of $\phi_b(z(b,\cdot))$ equals that of $z(b,\cdot)$. The Spearman correlation of Eq. (2) depends on the scores only through their ranks, so it is unchanged, as is any book whose legs are a function of the ordering alone. The shift $z \mapsto z + \delta_b$ is the special case $\phi_b(z) = z + \delta_b$. □

Remark 1 (What the range clip does). An implementation writes $z' = \mathrm{clip}(z + \delta_b, -1, 1)$. The clip is weakly but not strictly increasing: once it saturates, two instruments on the same side of the boundary map to the same value, and their strict order is lost. The clip binds on 66% of (backbone, event) cells. This is why the implemented per-event operator measures -0.0002 rather than exactly zero.

Remark 2 (Shared but non-monotone maps). A per-event label-shift correction $q'_y \propto q_y\, \pi_y^b/\pi_y$ (Saerens et al., 2002) is not a function of z alone, because it also reads q_{FLAT}. It can therefore reorder instruments that share z but differ in FLAT mass, which is why Proposition 1 is stated for strictly increasing maps of z. Such a map is still shared by the whole event: any reordering it induces comes from each instrument's FLAT mass, not from that instrument's outcome history.

B.2 PROOF OF PROPOSITION 2

Write $\bar{e} = \sum_j w_{bj} m_{j,a} e_j / W$. Under the assumptions, $\mathbb{E}[\bar{e}] = \mu$ and $\mathrm{Var}(\bar{e}) = \sigma^2 \sum_j w_{bj}^2 m_{j,a}/W^2 = \sigma^2/n_{\text{eff}}$. The error after the update is $(z + c\bar{e}) - t = c\bar{e} - \mu + \varepsilon$ with ε independent of $\bar{e}$. The update therefore changes expected squared error by $\mathbb{E}[(c\bar{e} - \mu)^2] - \mu^2 = c^2\sigma^2/n_{\text{eff}} + (1-c)^2\mu^2 - \mu^2$. This is negative exactly when $c^2\sigma^2/n_{\text{eff}} < c(2-c)\mu^2$, and dividing by $c > 0$ gives the stated condition. Because $c \le 1$, the threshold never exceeds $1/n_{\text{eff}}$. Conditional unbiasedness is an idealization that we do not test. □

C Additional Experimental Results

This section reports additional development-window results: the baselines against the frozen score and the robustness analysis.

C.1 Baselines Against the Frozen Score

Table 6: Change in rank IC against the same backbone frozen for the baselines, development window, with 5-day block-bootstrap intervals. Upper block: label-free adapters. Lower block: operators that use matured outcomes. SIGNAL-TTT is compared with them in Table 2.

method	venue	Δ rank IC ↑	95% block CI	improved
Label-free self-supervised adaptation (entropy or cache, no outcomes)				
TDA	CVPR 2024	+0.0001	$[-0.0003, +0.0006]$	4/9
EATA	ICML 2022	+0.0001	$[-0.0003, +0.0005]$	4/9
T3A	NeurIPS 2021	-0.0004	$[-0.0015, +0.0007]$	4/9
SAR	ICLR 2023	-0.0008	$[-0.0016, +0.0000]$	2/9
TENT	ICLR 2021	-0.0009	$[-0.0036, +0.0020]$	1/9
Adaptation from matured outcomes (retrieval)				
kNN-LM	ICLR 2020	+0.0148	$[-0.0031, +0.0306]$	9/9
kNN vote	–	+0.0143	$[-0.0141, +0.0396]$	8/9
Nadaraya–Watson	–	+0.0132	$[-0.0159, +0.0392]$	8/9
AdaNPC	ICML 2023	+0.0042	$[-0.0053, +0.0142]$	8/9

C.2 Robustness Analysis

All studies below use SIGNAL-TTT at its registered per-backbone k and T, with the memory fixed to the training and selection splits. They use validation and development data only, and none of them was used to change the method.

Neighbor Space. Adding market-state descriptors to the text lowers rank IC to 0.0127. Smaller text spaces score higher on the development window (0.0244 and 0.0224). They were not selected, because the embedding dimension was fixed in advance and every design choice is made on validation.

Table 7: Rank IC and gain over frozen on the development window when the neighbor space changes; frozen rank IC is 0.0030.

neighbor space (development window)	rank IC ↑	Δ vs frozen	improved
text embedding, 64 PCA dims	**0.0186**	**+0.0156**	**8/9**
text + market state	0.0127	+0.0097	8/9
market state only	0.0015	-0.0015	3/9
text embedding, 8 PCA dims	0.0083	+0.0054	6/9
text embedding, 16 PCA dims	0.0244	+0.0215	9/9
text embedding, 32 PCA dims	0.0224	+0.0194	9/9
text embedding, 128 PCA dims	0.0188	+0.0158	9/9

Operator Form. Validation selects the residual form of SIGNAL-TTT (0.0461, against 0.0409 for interpolation and 0.0180 for direct replacement). On the development window the order reverses for ranking, where direct replacement reaches 0.0425 and interpolation 0.0203 against 0.0186. We keep the validation choice.

Permutation Nulls. Under the within-event outcome permutation, the observed gain of +0.0156 gives $p = 0.008$. Permuting which matured event each memory outcome belongs to breaks the link between an event's text and its outcomes; this null has mean +0.0065 ($p = 0.069$).

Table 8: Three ways to combine the frozen score with the neighbors' signed targets, on validation and on the development window.

operator on the signed target	validation IC	development IC
frozen	0.0375	0.0030
direct (Nadaraya–Watson form)	0.0180	0.0425
interpolation (kNN-LM form)	0.0409	0.0203
residual (SIGNAL-TTT)	**0.0461**	**0.0186**

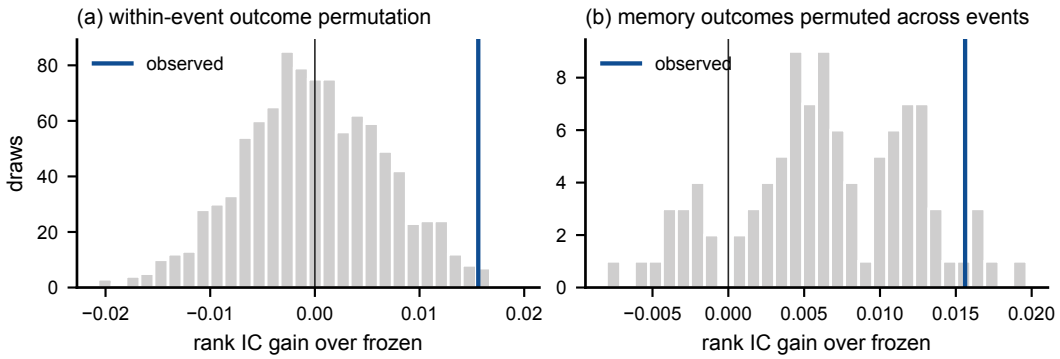

Figure 6: Permutation nulls of the development-window rank IC gain of SIGNAL-TTT. **(a)** Within-event outcome permutation (1000 draws). **(b)** Memory outcomes permuted across events (100 draws). The vertical line is the observed gain.

Memory Corruption, Horizon and Time. Rank IC degrades under corruption but not monotonically: it falls to 0.0083 at 75% and returns to 0.0149 at full permutation, for the reason given in Section 5.2. Figure 7 and Table 11 give the full horizon results summarized in Table 5. On validation the gain is concentrated in the first half: +0.0161 against +0.0008, with a net-return change of -1.38 bps in the second.

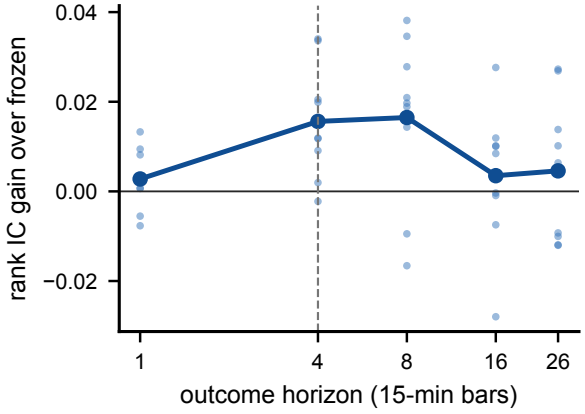

Figure 7: Rank IC gain of SIGNAL-TTT over frozen against the outcome horizon, development window: per backbone (dots) and mean (line). The dashed line marks the elicited horizon.

Order Invariance and Numerical Reproducibility. Tables 13 and 14 give the stream-order and thread-count results summarized in Section 5.2. The spread across thread counts arises because top-k selection is a discrete choice over near-tied distances.

Table 9: Permutation nulls for the development-window rank IC gain of SIGNAL-TTT over frozen.

null (development Δ rank IC)	observed	null mean	null sd	draws	p
memory outcomes permuted across events	+0.0156	+0.0065	0.0058	100	0.069
within-event outcome permutation (registered)	+0.0156	-0.0001	0.0068	1000	0.008

Table 10: Development-window rank IC when a fraction of the memory's matured outcomes is permuted within each instrument.

memory outcomes permuted	rank IC ↑
0% (clean)	**0.0186**
10%	0.0182
25%	0.0179
50%	0.0127
75%	0.0083
100%	0.0149

Table 11: The same elicited answers scored against several outcome horizons, development window.

outcome horizon	frozen	SIGNAL-TTT	Δ	improved
1 bar of 15 min	-0.0145	-0.0118	+0.0028	7/9
4 bars of 15 min (elicited)	**0.0030**	**0.0186**	**+0.0156**	**8/9**
8 bars of 15 min	0.0070	0.0235	+0.0165	7/9
16 bars of 15 min	-0.0004	0.0031	+0.0035	5/9
26 bars of 15 min	-0.0219	-0.0173	+0.0046	5/9

Table 12: Validation and development windows split at their median event.

window	frozen	SIGNAL-TTT	Δ IC	improved	Δ net bps
validation, first half	0.0467	0.0628	+0.0161	8/9	+1.08
validation, second half	0.0276	0.0284	+0.0008	4/9	-1.38
development, first half	0.0100	0.0210	+0.0110	7/9	+2.16
development, second half	-0.0041	0.0162	+0.0203	8/9	+2.48

Table 13: Adapted scores under four stream orders on the development window.

stream order	max $\lvert\Delta z'\rvert$	events checked	rank IC
chronological	0	321	0.0186
shuffled	0	321	0.0186
reversed	0	321	0.0186
one event at a time	0	60	–

Table 14: Mean development-window rank IC under several BLAS thread counts.

BLAS threads	frozen	SIGNAL-TTT
1 thread	0.0030	0.0182
2 threads	0.0030	0.0188
4 threads	0.0030	0.0188
8 threads	**0.0030**	**0.0186**
16 threads	0.0030	0.0186

D Additional Implementation Details

This section details the data splits, the event filter, retrieval and selection, the book, the inference and the porting of the baselines.

D.1 Splits

The panel is split chronologically:

- `train`: before 2025-07-01 (737 events);
- `select`: to 2025-10-01 (337 events);
- `validation`: 2025-10-01 to 2025-12-31 (217 events);
- `development`: 2026-01-02 to 2026-04-11 (321 events).

The memory is `train`$\cup$`select` and is never extended with evaluated events. All hyper-parameters were chosen on `validation`, for SIGNAL-TTT and for every rival.

D.2 Event Filter

A bar is an event if at least one post aligns to it and its 4-bar window is contiguous within one trading day. It enters the cross-section if at least 20 instruments have a complete window, and it contributes to rank IC if at least 5 instruments are priced and the score is not constant. An instrument without a complete window at an event is absent from A_b, and nothing is imputed.

D.3 Retrieval and Selection

Neighbors are the k nearest matured events in the frozen text-embedding space, reduced to 64 principal components fitted on `train`. The standardization is fitted once and applied frozen to every later event. Only k and T are selected, per backbone, on `validation`; the shrinkage has no hyper-parameter. Every design and hyper-parameter choice in this paper was made on `validation`. The search covered 614 configurations, or 5,526 (configuration, backbone) cells.

D.4 Book and Costs

The book is the dollar-neutral margin book of Section 3. Cost is charged on entry and exit of every leg, and it is swept over {0,1,2,3,5,10} bps per leg (Figure 4b).

D.5 Inference

Backbones are averaged within each event before any test, because all backbones score the same events. Intervals are 5-day moving-block bootstraps over trading days. Tests are one-sided 5-day block sign-flip permutations, and rival comparisons are Holm-corrected over the 12 rivals.

D.6 Porting the Label-Free Adapters

TENT, EATA, SAR, T3A and TDA cannot backpropagate into a frozen model queried only for log-probabilities. Each one therefore adapts a per-instrument affine recalibration of the three answer logits, initialized at identity, and applies its selection rule unchanged.

E Scope and Limitations

This section collects the scope conditions of the method and the caveats of our data.

E.1 Backbone Release Dates

No model card states a knowledge cutoff, so the public release date of the weights bounds it from above (Table 15). 3 backbones predate the development window, 3 were released during it and 3 after it. Development-window scores of the latter may carry look-ahead information.

Table 15: Public release date of each backbone's weights (Hugging Face model API), relative to the development window.

backbone	weights released	before development
`Qwen3-8B`	2025-04-27	✓
`Qwen3-14B`	2025-04-27	✓
`Seed-OSS-36B`	2025-08-20	✓
`Qwen3.5-35B-A3B-Base`	2026-02-24	during
`Qwen3.5-9B-Base`	2026-02-26	during
`Qwen3.5-9B`	2026-02-27	during
`Qwen3.6-MoE-35B`	2026-04-15	✗
`Qwen3.6-27B`	2026-04-21	✗
`Qwen3.8-27B`	2026-08-05	✗

E.2 Cross-Sections Without Differential Exposure

On a panel of 40 crypto perpetual futures, the frozen model answers FLAT on 96.1% of pairs, against 84.4% on equities, and only 4.2% of pairs are decided, against 16.6%. Members that load on one common factor leave the method little cross-sectional signal to correct, so we treat such panels as out of scope.