WHAT DOES TEXT ADD TO FORECASTING? A CONTROLLED AUDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

In our controlled studies, text-induced forecast changes do not consistently translate into lower error, and embedding-source rankings depend on the evaluation metric. We distinguish forecast response, control-relative benefit, semantic selectivity, and embedding-source effects using paired event worlds and matched training. Across 240 synthetic processes, a fixed Qwen instruction model favors correct over incorrect event descriptions on average and moves forecasts in the event's direction, but the benefit is concentrated in level changes and reverses for capacity constraints. Follow-up comparisons on the same inspected cohort have negative mean scores against no auxiliary context (-0.1283) and metadata only (-0.1579): resampling intervals are negative, while conservative bounded intervals include zero. Across 16 matched GPT-2 training units, original and permuted pretrained embeddings both reduce an integrated unsigned loss gap relative to random embeddings, yet final error and signed-benefit readouts rank the sources differently. These findings identify concrete dependence on the comparator, event family, and readout, without establishing general semantic benefit or a shared semantic mechanism.

1 INTRODUCTION

A promotion, weather event, or disruption may shift the future without a precursor in numerical history. An event description can supply that missing information. Context-required benchmarks and scenario-guided forecasters address this setting (Williams et al., 2025; Jang et al., 2026; Liu et al., 2024; Jin et al., 2024; Zhou et al., 2023; Xue & Salim, 2024). We ask how to interpret the resulting forecast changes, which may reflect event information, surface wording, token correspondence, or optimization shaped by pretraining.

A performance gain from a language-model backbone or multimodal fusion architecture (Gruver et al., 2023; Chang et al., 2023; Cao et al., 2024; Zhou et al., 2025; Mou et al., 2025; Qiu et al., 2026a) establishes a comparison, not every interpretation of that comparison. Beating incorrect text need not imply benefit over neutral context. Moving in an event's direction need not imply an appropriate response magnitude. A parameter-source effect need not favor original token correspondence. Recent language-model and multimodal forecasting audits motivate examining these possibilities directly (Tan et al., 2024; Zhang et al., 2025; Sun et al., 2025; Qiu et al., 2026b; Shchur et al., 2025; Liu et al., 2026; Sridhar et al., 2026; Andrews & Mengaldo, 2026).

We distinguish forecast response, control-relative benefit, semantic selectivity, and parameter attribution (Figure 1). Each requires its own comparison. Known-behavior predictors first check the loss score's sign and scale. A frozen Qwen2.5-7B-Instruct forecaster then receives six text conditions on paired event worlds to examine response, benefit, and expression controls. A separate GPT-2 Medium study matches non-target initial states while varying original pretrained, permuted pretrained, and random token embeddings. This intervention tests source dependence of a loss diagnostic, not the same mechanism as direct prompting. Earlier Qwen3 and Pythia comparisons provide context without pooling their distinct outcomes.

The contribution is an audit in which the comparator, target world, readout, and assumptions are explicit. Qwen tests how the benefit judgment changes with the baseline and event family. Matched GPT-2 training tests whether a source ordering on a loss-separation diagnostic also orders forecasting utility. Together, the studies expose different ways of misreading a text-related effect.

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.

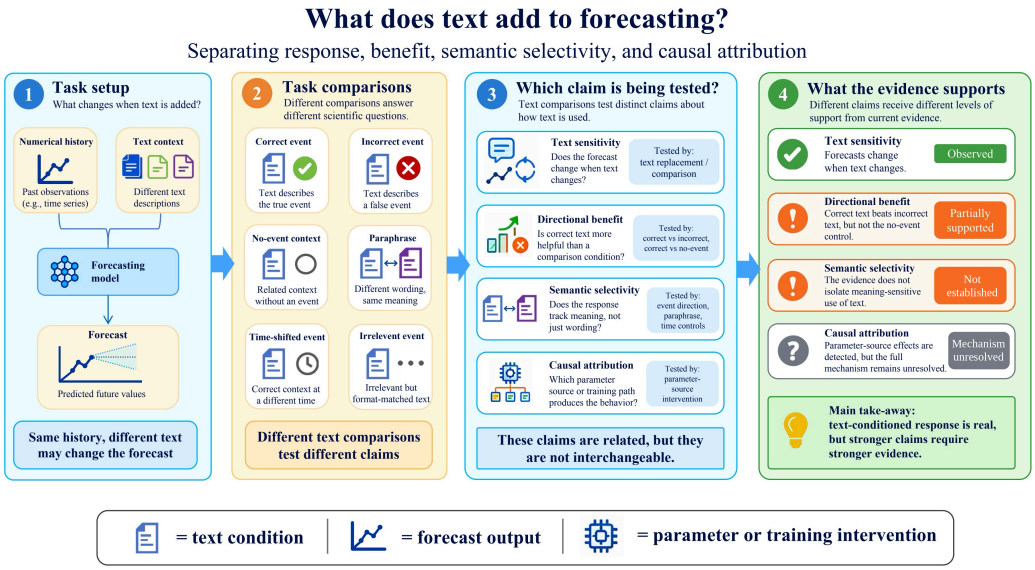


Figure 1: **Four questions about text-conditioned forecasting.** Text comparisons test forecast changes, control-relative benefit, and semantic selectivity; matched training and embedding replacement address attribution. Qwen's aggregate correct–incorrect contrast is favorable, unlike the neutral no-event comparison. GPT-2 source effects concern an integrated unsigned loss gap. Panel order guides reading, not logical implication.

Anonymous reproducibility materials are available at https://anonymous.4open.science/r/What-Does-Text-Add-to-Forecasting-Repository-4C00.

2 Related Work

Text-conditioned forecasting. Prompting, reprogramming, and multimodal fusion connect language models to temporal prediction (Jin et al., 2024; Zhou et al., 2023; Chang et al., 2023; Gruver et al., 2023; Xue & Salim, 2024; Cao et al., 2024; Zhou et al., 2025; Mou et al., 2025; Qiu et al., 2026a). Context is Key assembles context-required tasks, while What If TSF (WIT) evaluates plausible and counterfactual scenarios primarily through direction (Williams et al., 2025; Jang et al., 2026). Our paired numerical futures permit separate loss, displacement, and direction comparisons. Time-MMD and TimesX instead emphasize multimodal breadth and real-world evaluation (Liu et al., 2024; 2026); Aurora is a trained multimodal foundation model (Wu et al., 2026). These provide complementary settings, not evidence of transfer from our controlled studies.

Text-replacement audits examine content versus pathway presence (Sridhar et al., 2026), and concurrent information-theoretic diagnostics compare annotations before training (Andrews & Mengaldo, 2026). We instead compare fixed-model forecasts and ask whether matched embedding-source effects retain their interpretation across readouts. Work on pretraining, fusion, distribution shifts, and temporal reasoning motivates these checks (Tan et al., 2024; Zhang et al., 2025; Sun et al., 2025; Lee et al., 2026; Shchur et al., 2025; Qiu et al., 2026b; Merrill et al., 2024; Cakiroglu et al., 2026).

Forecasting and transfer. Accuracy measures and benchmarks (Hyndman & Koehler, 2006; Makridakis et al., 2020; 2022), neural architectures (Oreshkin et al., 2020; Wu et al., 2021; Nie et al., 2023; Zeng et al., 2023; Zhou et al., 2021), and temporal foundation models (Ansari et al., 2024; Das et al., 2023; Woo et al., 2024; Rasul et al., 2023; Garza et al., 2023; Ekambaram et al., 2024) establish the forecasting context. Language pretraining can reshape optimization and representations (Erhan et al., 2010; Neyshabur et al., 2020; Gururangan et al., 2020). Transformer and language-model research supplies our architectural background (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Bommasani et al., 2021); restricted adaptation

(Hu et al., 2022; Aghajanyan et al., 2021) and in-context learning (Min et al., 2022; Akyürek et al., 2022; Xie et al., 2021) motivate distinguishing sources of an observed effect.

Behavioral and causal evaluation. Causal inference distinguishes interventions from associations (Pearl, 2009; Peters et al., 2017); neural abstraction, mediation, editing, and probing make particular interventions testable (Geiger et al., 2021; Vig et al., 2020; Meng et al., 2022; Elazar et al., 2021). CheckList and contrast sets motivate controlled behavioral comparisons (Ribeiro et al., 2020; Gardner et al., 2020), while robust evaluation and multiplicity methods address heterogeneous effects and repeated testing (Sagawa et al., 2020; Holm, 1979; Benjamini & Hochberg, 1995). Our contribution is their application to comparator- and readout-dependent forecasting judgments, not a new forecasting architecture.

3 Evaluation Framework and Methods

3.1 Questions and controlled event worlds

For numerical history x, text t, and future y, let $f(x,t)$ be the 24-step forecast and $L(f(x,t),y)$ its loss. *Response* asks whether the forecast changes. *Control-relative benefit* asks whether loss falls relative to a specified text control. *Semantic selectivity* combines event-direction response with meaning-preserving and meaning-changing controls, whereas *causal attribution* requires parameter-source or training interventions. Evidence for one question does not mathematically answer the others.

Paired futures expose the event difference while holding history and noise fixed. The evaluation contains 240 synthetic base processes: 80 additive level changes, 80 timed interruptions, and 80 capacity constraints. Each process has 96 historical points and a 24-point horizon. Its two event worlds share the history, latent base future, and noise. The futures differ through opposite horizon-wide shifts, opposite shifts on a predetermined interior interval, or upper-capacity versus lower-floor constraints derived from history. Let $z_i \in \mathbb{R}^{24}$ be the shared base future including its noise. The paired structures can be written as

$$\begin{aligned} \text{Level change:} &\quad z_i + \delta_i \mathbf{1}, \quad z_i - \delta_i \mathbf{1}, \\ \text{Interruption:} &\quad z_i + \delta_i \mathbf{1}_{I_i}, \quad z_i - \delta_i \mathbf{1}_{I_i}, \\ \text{Capacity:} &\quad \min(z_i, u_i), \quad \max(z_i, \ell_i). \end{aligned} \tag{1}$$

Here I_i is an interior time interval, and the capacity and floor are history-derived; min/max act elementwise. These equations state the event operations, not additional assumptions about parameter distributions. Let y_i^c and y_i^w denote the designated reference and alternative futures. The alternative description is incorrect relative to y_i^c. Reference-role counts and two-role scoring are detailed in Appendix A.

We specify six text conditions: correct event, matched incorrect event, neutral no-event, paraphrase, time-shifted event, and format-matched irrelevant text. Time shifts apply only to 80 timed interruptions, yielding $240 \times 6 - 160 = 1{,}280$ evaluations. In the original evaluation, losses use y_i^c; direction alignment uses $y_i^c - y_i^w$. The base process is the inference unit. Appendix A details the construction and dependence structure.

3.2 Direct-prompt outcomes

Forecast changes measure response; loss contrasts assess benefit against a named control. Let L_i^c, L_i^w, L_i^0 be 24-point mean squared errors under correct, incorrect, and no-event text. With $\varepsilon = 10^{-15}$, define

$$S_i^{c:w} = \frac{2(L_i^w - L_i^c)}{L_i^w + L_i^c + \varepsilon}, \qquad S_i^{c:0} = \frac{2(L_i^0 - L_i^c)}{L_i^0 + L_i^c + \varepsilon}. \tag{2}$$

These dimensionless scores lie in $[-2, 2]$; positive values favor correct text against the named control, rather than expressing percentage error reductions. For forecasts $f_i^c, f_i^w \in \mathbb{R}^{24}$, define

$$C_i = \frac{\langle f_i^c - f_i^w,\, y_i^c - y_i^w \rangle}{\|f_i^c - f_i^w\|_2 \, \|y_i^c - y_i^w\|_2 + \varepsilon}, \tag{3}$$

$$P_i = \frac{24^{-1} \sum_h |f_{i,h}^c - f_{i,h}^w|}{\frac{1}{2}(24^{-1} \sum_h |f_{i,h}^c|) + \frac{1}{2}(24^{-1} \sum_h |f_{i,h}^w|) + \varepsilon}. \tag{4}$$

C_i is event-direction alignment, P_i is forecast displacement, and $D_i = |S_i^{c:w}|$ is the unsigned loss gap. Paraphrase selectivity subtracts the analogous correct-versus-paraphrase signed contrast from $S_i^{c:w}$; time-shift selectivity applies only to interruptions. Both selectivity summaries are exploratory.

Across the original conditions, the fixed Qwen2.5-7B-Instruct model receives the same history and instruction; only the event text changes. Greedy constrained generation produces 24 numerical predictions, with no additional training. The input serializes the 96 history values to six significant digits, supplies the requested future labels, and asks for ordered timestamp–value pairs inside forecast tags without commentary. The model and decoder remain fixed across conditions. The no-event control states "No additional event is supplied for the forecast period." This is not an empty prompt and may itself signal event absence, so the comparison measures a specific instruction contrast rather than text-off incremental information.

After pre-submission review, we specified two additional controls before generating outputs for the same 240 processes. The *no-auxiliary-context* prompt removes the full context block but keeps the forecasting instruction, numeric history, future labels, and output format; the model remains language based. The *metadata-only* prompt states only that the history has 96 observations, the horizon has 24 observations, and timestamps are ordinal labels. For each process i, let $g_{i,b}$ be its forecast under baseline b and let $k \in \{c, w\}$ index the two event worlds. With event-conditioned forecast f_i^k and target y_i^k, we compute

$$S_i^k(b) = \frac{2\{L(g_{i,b}, y_i^k) - L(f_i^k, y_i^k)\}}{L(g_{i,b}, y_i^k) + L(f_i^k, y_i^k) + 10^{-15}}, \qquad \bar{S}_i(b) = \tfrac{1}{2}\{S_i^c(b) + S_i^w(b)\}. \tag{5}$$

We average worlds within process, then aggregate across 240 units. Both controls, all families, and raw losses are retained. The follow-up's timing and evidential role are summarized in Table 7.

3.3 Statistical analysis

Median intervals and sum tests address different summaries. Table 7 separates the original analyses, exploratory replacements, saved-output reanalyses, and new-condition follow-ups. Here, "review-driven" and "post-review" refer to pre-submission scrutiny of the initial results, not to official ICLR 2027 reviews. For medians over the 240 processes, we use bias-corrected and accelerated (BCa) 95% bootstrap intervals (10,000 resamples; seed 42) (Efron, 1979; Wilcox, 2012). Three pre-specified one-sided tests use $T = \sum_i z_i$ for $z_i \in \{S_i^{c:w}, C_i, S_i^{c:0}\}$. Each test draws $B = 1{,}048{,}576$ random sign assignments to $|z_i|$ with seed 42. If b simulated sums are at least as large as the observed sum, $p_{\mathrm{MC}} = (b+1)/(B+1)$; Holm correction controls the three-comparison family at $\alpha = .05$. These tests target sums, not the displayed median intervals.

Sign flipping assumes that the joint null distribution is invariant to the applied sign changes. Independent effects symmetric about zero are sufficient; independence or zero mean alone is not. Jointly exchanging both event worlds and forecasts leaves C_i unchanged, so the event construction provides no design-based justification for sign flipping. We therefore interpret the tests under a sign-symmetry model, not as randomization guaranteed by the design (Winkler et al., 2014). Appendix A states the corresponding GPT-2 assumptions.

Post-review saved-output analyses average the three family means equally. Within-family resampling yields approximate simultaneous 95% mean intervals; bounded-variable intervals require independence but not sign symmetry (Hoeffding, 1963). All outcomes, families, and leave-one-family-out summaries are retained (Appendix K). These sensitivity analyses neither replace the original tests nor constitute prospective confirmation.

For the two added baselines, the supplementary plan resamples 80 processes per family for 20,000 paired replicates using the 64-bit permuted congruential generator (PCG64), seed 20260917. The

1.25th and 98.75th percentiles give approximate Bonferroni simultaneous 95% intervals. Simultaneous Hoeffding intervals for scores in $[-2, 2]$ require independent processes, not sign symmetry or physical randomization. No new p-value, practical threshold, or subgroup decision is introduced. A descriptive capacity check makes only the substitutions "a upper" to "an upper" and "an lower" to "a lower".

3.4 Matched training across token-embedding sources

This study asks how embedding source changes correct–incorrect loss separation during adaptation, rather than assuming that a source effect measures utility. The GPT-2 Medium study uses two fixed sets of 133 training examples, with history 96 and horizon 24. Their 266 occurrences comprise 148 from Time-MMD and 118 from Context is Key; repeated windows are not new source series. Each split and initialization repetition defines a matched unit, with eight repetitions per split. Within a unit, conditions share non-target and forecasting-component initialization, data order, and optimization settings. Their input embeddings are original pretrained vectors, the same row-vector multiset with permuted token correspondence, or random vectors drawn from $\mathcal{N}(0, 0.02^2)$. Across the 16 units, 16 distinct pre-specified vocabulary derangements are each fixed within one unit. Derangement identity and training repetition are not independently crossed.

The selected token embeddings and pretrained backbone weights remain frozen, with backbone stochastic layers disabled. The additional forecasting and text-conditioning parameters, including cross-attention and low-rank adaptation, are trainable; "frozen backbone" does not mean that these added parameters are frozen. They receive 200 AdamW updates: learning rate 10^{-3}, weight decay 0.01, no schedule, and gradient clipping at 1.0. The retained architectural detail and its limits are given in Appendix E.

At $t \in \{0, 10, 25, 50, 100, 200\}$, both text conditions are evaluated on all 240 processes against y_i^c, using the recorded past-normalized squared losses (Appendix E). For source condition a within a matched unit,

$$d_{i,t}^{(a)} = \frac{2|L_{i,t}^{c,(a)} - L_{i,t}^{w,(a)}|}{L_{i,t}^{c,(a)} + L_{i,t}^{w,(a)} + 10^{-15}}, \qquad m_t^{(a)} = \operatorname{median}_{i=1}^{240} d_{i,t}^{(a)}, \tag{6}$$

$$W^{(a)} = -\sum_{j=0}^{4} \frac{m_{t_j}^{(a)} + m_{t_{j+1}}^{(a)}}{2} \frac{t_{j+1} - t_j}{200}, \qquad \Delta W^{a:b} = W^{(a)} - W^{(b)}. \tag{7}$$

Here $(t_0, \ldots, t_5) = (0, 10, 25, 50, 100, 200)$. Larger W means a smaller integrated absolute loss gap, not lower error or greater displacement. It remains the pre-specified diagnostic, not an optimization objective; post-review error and signed-benefit summaries assess its relation to utility.

The primary contrasts are original minus random, permuted minus random, and original minus permuted. We enumerate all 2^{16} sign assignments to the summed matched-unit contrasts and apply Holm correction at $\alpha = .025$. Mean intervals are descriptive 95% Student t intervals. Exhaustive enumeration does not remove the symmetry assumption because source labels were not randomized across units.

At update 200, original and permuted embeddings are exchanged between their trained models, holding other parameters fixed within each replacement. Before adaptation, matched conditions share the same non-target initialization; later divergence may result from learning. Unlike the combined training-and-installed-source W comparison, these replacements examine a two-factor layout outside the primary Holm family. Appendix F defines its conditional contrasts and distinguishes them from the incompletely specified archived summary labels.

4 Results

4.1 Reference behaviors anchor the score

The reference predictors test how the signed score represents known relative-loss behavior (Figure 2). The mean is 2.000 for the event rule, 0.000 for history-only, -0.00576 for surface form, and -2.000 for the reversed event. The over-amplified predictor scores 0.8376. These values anchor sign and

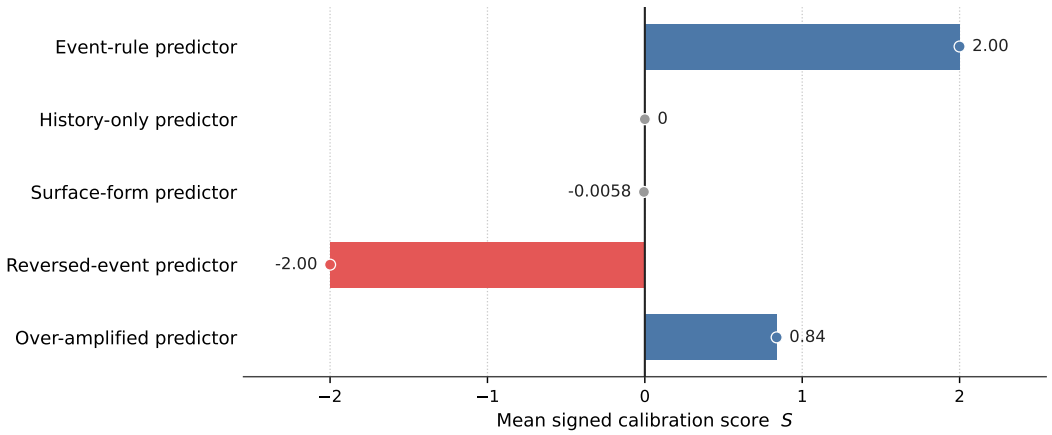


Figure 2: **Signed-score calibration on five reference behaviors.** Bar lengths and markers show the mean $S = 2(L_w - L_c)/(L_w + L_c + \varepsilon)$ over 240 processes; they do not encode uncertainty. Positive values favor correct text. Event-rule and reversed-event predictors approach ± 2; history-only and surface-form scores are near zero. The over-amplified score of 0.8376 shows that positive relative benefit need not imply correctly scaled forecasts.

scale. The positive over-amplified score also shows why beating incorrect text does not guarantee an appropriately scaled response. This check concerns the specified predictors, not general semantic recognition.

4.2 Qwen's benefit depends on the comparison

Table 1 brings the family and baseline summaries together; full process summaries and median intervals are in Table 3.

Table 1: **Qwen's mean effects depend on the event family and baseline.** The upper block uses the original reference world, with 80 processes per family; S is a normalized loss contrast and C is direction cosine. The lower block uses $\bar{S}_i(b)$, averaging two world-specific scores within each of 240 processes. Positive scores favor the matching event description. These are descriptive means, not subgroup tests; Table 3 and Appendix L give intervals and their assumptions.

Original event family	**Correct vs. incorrect**	**Direction cosine**	**Correct vs. no-event**
Level change	0.7110	0.9020	−0.2311
Timed interruption	0.1490	0.3574	−0.3685
Capacity constraint	−0.1351	−0.1924	0.0963
Follow-up baseline	**Mean role-averaged score $\bar{S}_i(b)$**		
No auxiliary context	−0.1283		
Metadata only	−0.1579		

Correct versus incorrect text gives mean signed benefit 0.241621, median 0.071474, and BCa 95% median interval $[0, 0.415214]$ (Figure 3); the 240 effects comprise 126 positive, 96 negative, and 18 zero values. The pre-specified sum test meets its positive criterion under sign symmetry (Holm $p = 0.001129$).

Event-direction cosine has mean 0.355653, median 0.497459, and median interval $[0.397324, 0.607594]$, with 163 positive, 57 negative, and 20 zero effects. Its sum test also meets the criterion under sign symmetry (Holm $p = 2.86102 \times 10^{-6}$; simulation counts in Table 4).

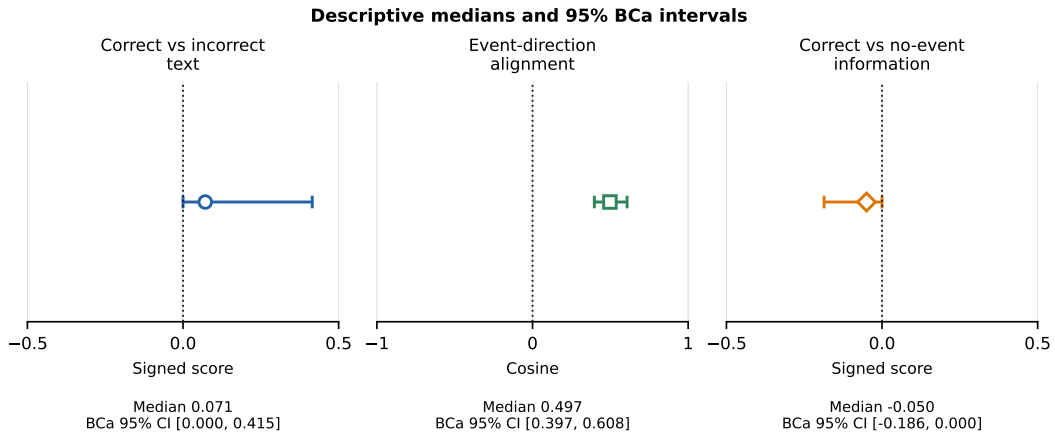


Figure 3: **Descriptive Qwen medians, separate from the sum tests.** Markers and bars show medians and 95% bias-corrected and accelerated (BCa) intervals over 240 processes. Panels (a) and (c) share the signed-score axis; panel (b) uses the cosine range. Positive scores favor correct text against the labeled control; positive cosine denotes event-direction alignment. The plotted intervals concern medians, not the summed-effect tests and Holm probabilities reported in Table 4.

Against neutral no-event information, the mean is -0.167744, the median -0.049799, the median interval $[-0.186300, 0]$, and Holm $p = 0.995060$, with 104 positive, 128 negative, and 8 zero effects. The positive-benefit criterion is unmet; this one-sided test does not establish harm.

Event forecasts versus no auxiliary context and metadata. In the review-driven fixed-cohort follow-up, event-conditioned forecasts have world-averaged mean scores -0.128311 against no auxiliary context and -0.157928 against metadata only. Approximate family-stratified simultaneous intervals are $[-0.238007, -0.015417]$ and $[-0.276363, -0.041840]$; conservative Hoeffding simultaneous intervals are $[-0.510499, 0.253877]$ and $[-0.540115, 0.224260]$. Approximate intervals are negative; conservative intervals include zero. The original tests remain unchanged.

Event families change the aggregate interpretation. Mean correct-versus-incorrect scores are 0.7110 for level changes, 0.1490 for timed interruptions, and -0.1351 for capacity constraints; the corresponding mean direction cosines are 0.9020, 0.3574, and -0.1924. Removing level changes leaves a signed mean of 0.00695 and a median of zero, although the primary aggregate retains all families. The post-review family-stratified mean interval for signed benefit is $[0.0751, 0.4024]$; the conservative bounded-mean interval includes zero. The direction-cosine bounded-mean interval is $[0.1559, 0.5554]$. Appendix K reports both simultaneous interval sets for all three outcomes. The smallest correct-plus-incorrect loss denominator is 0.151525, which rules out machine-scale denominators as the source of the aggregate, though normalization still changes process weighting. The post-review bound supports positive mean alignment under process independence without requiring sign symmetry. Uniform benefit across event types remains unresolved (Table 3).

Expression and wording controls. Exploratory summaries give a forecast-displacement median of 1.3398, an unsigned-loss-gap median of 0.9969, a paraphrase-selectivity median of 0 (mean -0.0143), and a time-shift-selectivity median of 0.2913 on the 80 applicable interruption processes. A post-review descriptive contrast between correct and format-matched irrelevant text has a mean of -0.178524, a median of -0.053953, and 108/127/5 positive/negative/zero effects. Taken together, these controls do not establish semantic selectivity (Appendix D).

For capacity constraints, correcting the two identified article errors changes 70 of 160 prompt strings and at least one forecast in 28 of 80 processes. The role-averaged correct-versus-incorrect score R_i in Eq. (9), over these 80 capacity processes, shifts from -0.219522 to -0.190806 (mean change 0.028716). This is not the event-versus-baseline score $\bar{S}_i(b)$; other surface cues remain uncontrolled.

4.3 EMBEDDING SOURCE CHANGES THE UNSIGNED LOSS-GAP TRAJECTORY

Original pretrained embeddings have larger W than random embeddings in all 16 matched units: mean $\Delta W = 0.265906$, descriptive 95% mean interval $[0.223408, 0.308404]$. Permuted embeddings also have larger W than random embeddings in all 16 units: mean 0.282395, interval $[0.237172, 0.327618]$. Under the matched-unit sign-symmetry model, both Holm-adjusted probabilities are 4.58×10^{-5} (Figure 4; exact values in Appendix E). Thus, relative to random embeddings, the smaller integrated absolute correct/incorrect loss gap appears for both original and specified permuted pretrained sources.

For original minus permuted embeddings, the mean is -0.016489, the median is -0.011418, the interval is $[-0.035559, 0.002580]$, and 3/16 units are positive. The one-sided $p = 0.959778$ does not establish the pre-specified original-correspondence advantage. This comparison leaves the contribution uncertain; it does not establish zero. Permutation identity varies with training repetition, and the final-state embedding replacements form a separate exploratory analysis (Appendix F).

Source ordering depends on the readout and evaluation time. At update 200, the mean signed correct-versus-incorrect scores are 0.0171, 0.0232, and 0.0523 for original, permuted, and random sources; mean correct-text normalized errors are 2.5739, 4.1793, and 3.0044. These post-review summaries use all 16 units but describe final outcomes, not the primary curve estimand (Table 5). Source ordering in W differs from final loss and signed-benefit orderings; larger W alone does not identify a better forecaster.

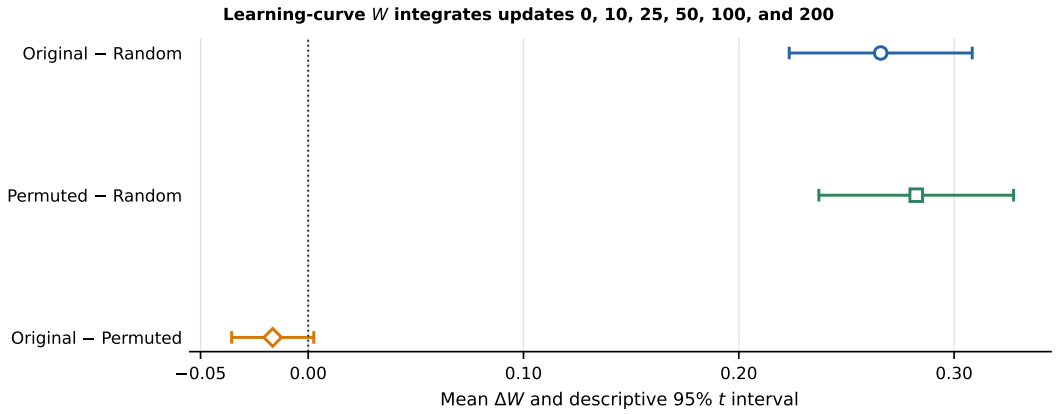


Figure 4: **Embedding-source contrasts on the learning-curve diagnostic.** Original uses pretrained token vectors; Permuted uses the same vector multiset with a fixed derangement within each unit; Random uses randomly initialized vectors. Points are mean ΔW over 16 matched units, with descriptive 95% Student t intervals. W integrates median absolute correct/incorrect loss gaps across six training times; positive contrasts denote smaller integrated gaps, not lower error. Original–Permuted does not establish an advantage or equivalence.

4.4 CROSS-MODEL REPLICATION DEPENDS ON THE MEASURED BEHAVIOR

The retained records of two earlier Pythia-410M studies report a narrow text-replacement effect in their prescribed negative direction (Table 6, Appendix G). The later replication also meets split and seed consistency requirements. Stronger alignment, stability across pretraining runs, and trajectory-wide criteria remain unmet. A historical Qwen3 comparison, meanwhile, has an opposite-direction point estimate. Because these records use different endpoints, we do not pool them with the current experiments. The records do not fully define historical endpoint formulas or unit membership, so these results provide historical context rather than an independently reconstructed replication claim.

Table 2 in Appendix A separates the evidence supported by each comparison from interpretations that remain unresolved.

5 DISCUSSION

Why an event response need not reduce error. The Qwen comparisons distinguish direction from utility. Level changes drive the favorable aggregate contrast with incorrect text, whereas capacity constraints reverse both signed benefit and direction. The more neutral baselines change the question: on the observed cohort, event-conditioned forecasts have negative mean contrasts against neutral no-event, no-auxiliary-context, and metadata-only inputs. The latter two are follow-up comparisons, as classified in Table 7.

The squared-loss identity in Appendix L makes this distinction concrete. Its positive alignment term is outweighed, on average, by the response-magnitude and center-error terms. Direction alone omits both penalties. These are terms in the same loss difference, not independent confirmations or an explanation of the model's internal mechanism. Article correction attenuates the capacity result without reversing its mean.

Why a source effect need not select the better forecaster. Original and permuted pretrained embeddings both reduce the integrated unsigned loss gap, but their final correct-text losses lie on opposite sides of the random reference. Random embeddings have the largest final signed-benefit score. The diagnostic thus detects a source-dependent training trajectory without ranking forecasting utility. Permutation preserves the vector multiset while changing correspondence, so the intervention does not isolate geometry, scale, frequency, or semantic content.

Equal squared losses can arise from different forecasts at the same distance from the target. A reduced correct–incorrect gap therefore need not indicate identical predictions or lower correct-text error. Reporting displacement alongside both losses exposes this distinction.

Before adaptation, matched conditions share the same non-target initialization; later divergence can be an effect of training. The primary W contrast combines training and installed sources. Exploratory replacements have imprecise summaries, with the available conditional-baseline documentation limited as specified in Appendix F. No component-level mechanism follows from these records.

6 LIMITATIONS

Scope and controls. The findings are limited to three synthetic event families, one Qwen direct-prompt system, and one GPT-2 adaptation design. The no-auxiliary-context and metadata-only conditions retain different amounts of prompt structure. Both follow-ups use the same inspected cohort and therefore serve as sensitivity analyses rather than independent replications.

Inference. The sign-flip analyses are conditional on sign symmetry, whereas the bounded-mean analyses require independent process effects. Conservative signed-benefit intervals include zero; neither analysis provides design-based randomization.

Attribution and reproducibility. The GPT-2 study estimates a total source-condition contrast under the matched-unit design. Because derangement identity and training repetition are not independently crossed, the present design does not decompose that contrast into finer embedding properties. The released supplement supports statistical reconstruction from saved outputs rather than full end-to-end retraining.

7 CONCLUSION

A forecasting audit should report the comparator, target world, and readout together. In the studied Qwen system, event-direction alignment coexists with baseline-dependent benefit and event-family reversals. In matched GPT-2 training, a source ordering on the integrated unsigned loss gap differs from the final error and signed-benefit orderings. These complementary results show where an observed text-related change can be mistaken for forecasting utility. The conclusions remain conditional on the studied processes, interventions, and stated inferential assumptions.

AI USE STATEMENT

Generative AI tools supported experimental design and data analysis, including research planning, method implementation, synthetic-data generation, and result interpretation. AI-assisted analyses were checked against experiment records and separate numerical calculations. The authors are responsible for the final methodology, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

We provide an anonymous reproduction package containing the controlled-process definitions, text-condition metadata, saved Qwen forecasts and targets, GPT-2 per-process loss records, and code for reconstructing the headline reported statistics and selected post-review saved-output analyses. Additional diagnostic summaries are released as CSV/JSON records. The anonymous repository is available at `https://anonymous.4open.science/r/What-Does-Text-Add-to-Forecasting-Repository-4C00`. Experimental definitions, statistical procedures, and additional implementation details are documented in the appendix. The released materials support statistical reconstruction from saved model outputs; they do not provide full end-to-end reproduction of GPT-2 training or model re-evaluation.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.568. URL https://doi.org/10.18653/v1/2021.acl-long.568.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, et al. What learning algorithm is in-context learning? investigations with linear models, 2022. URL https://arxiv.org/abs/2211.15661.

Emma Andrews and Gianmarco Mengaldo. When does text inform? benchmarking information-theoretic metrics for multimodal time-series forecasting, 2026. URL https://arxiv.org/abs/2609.11282.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Mert Onur Cakiroglu, Mehmet Dalkilic, and Hasan Kurban. The spectrum is not enough: When context helps time-series forecasting, 2026. URL https://arxiv.org/abs/2607.13006.

Defu Cao, Furong Jia, Sercan O. Arik, et al. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YH5w12OUuU.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, et al. LLM4TS: Aligning pre-trained LLMs as data-efficient time-series forecasters, 2023. URL https://arxiv.org/abs/2308.08469.

Abhimanyu Das, Weihao Kong, Rajat Sen, et al. A decoder-only foundation model for time-series forecasting, 2023. URL https://arxiv.org/abs/2310.10688.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://doi.org/10.18653/v1/N19-1423.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, et al. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series, 2024. URL https://arxiv.org/abs/2401.03955.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, et al. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9: 160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://doi.org/10.1162/tacl_a_00359.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, et al. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 2010. URL https://www.jmlr.org/papers/v11/erhan10a.html.

Matt Gardner, Yoav Artzi, Victoria Basmov, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1, 2023. URL https://arxiv.org/abs/2310.03589.

Atticus Geiger, Hanson Lu, Thomas Icard, et al. Causal abstractions of neural networks, 2021. URL https://arxiv.org/abs/2106.02997.

Nate Gruver, Marc Finzi, Shikai Qiu, et al. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2310.07820.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, et al. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://doi.org/10.2307/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 2006. doi: 10.1016/j.ijforecast.2006.03.001. URL https://doi.org/10.1016/j.ijforecast.2006.03.001.

Jinkwan Jang, Hyunbin Jin, Hyungjin Park, et al. What if TSF: A benchmark for reframing forecasting as scenario-guided multimodal forecasting, 2026. URL https://arxiv.org/abs/2601.08509.

Ming Jin, Shiyu Wang, Lintao Ma, et al. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Seunghan Lee, Jun Seo, Jaehoon Lee, et al. Rethinking multimodal fusion for time series: Text modalities need constrained fusion, 2026. URL https://arxiv.org/abs/2603.22372.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, et al. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.08627.

Haoxin Liu, Yichen Zhou, Rajat Sen, et al. Rethinking multimodal time-series forecasting evaluation, 2026. URL https://arxiv.org/abs/2607.06973.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 2020. doi: 10.1016/j.ijforecast.2019.04.014. URL https://doi.org/10.1016/j.ijforecast.2019.04.014.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 2022. doi: 10.1016/j.ijforecast.2021.11.013. URL https://doi.org/10.1016/j.ijforecast.2021.11.013.

Kevin Meng, David Bau, Alex Andonian, et al. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, et al. Language models still struggle to zero-shot reason about time series, 2024. URL https://arxiv.org/abs/2404.11757.

Sewon Min, Xinxi Lyu, Ari Holtzman, et al. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Shangyang Mou, Qiang Xue, Jinhui Chen, et al. MM-iTransformer: A multimodal approach to economic time series forecasting with textual data. *Applied Sciences*, 15(3):1241, 2025. doi: 10.3390/app15031241. URL https://doi.org/10.3390/app15031241.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2008.11687.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, et al. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161. URL https://www.cambridge.org/core/books/causality/.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Wei Qiu, Ying Zhao, and Mengxia Ren. Large language models for unstructured economic data processing: Enhancing GDP forecasting accuracy with heterogeneous textual information. *Journal of King Saud University Computer and Information Sciences*, 38:416, 2026a. doi: 10.1007/s44443-026-00788-2. URL https://doi.org/10.1007/s44443-026-00788-2.

Xin Qiu, Junlong Tong, Yirong Sun, et al. Rethinking the role of LLMs in time series forecasting, 2026b. URL https://arxiv.org/abs/2602.14744.

Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, et al. Lag-Llama: Towards foundation models for probabilistic time series forecasting, 2023. URL https://arxiv.org/abs/2310.08278.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, et al. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://doi.org/10.18653/v1/2020.acl-main.442.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-group accuracy. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, et al. fev-bench: A realistic benchmark for time series forecasting, 2025. URL https://arxiv.org/abs/2509.26468.

Karthik Sridhar, Atharva Gupta, Nishant Pradhan, et al. Semantics or structure? auditing text sensitivity in multimodal time-series forecasting. In *ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://arxiv.org/abs/2608.22321.

Zhengke Sun, Hangwei Qian, and Ivor Tsang. Exploring the effectiveness and interpretability of texts in LLM-based time series models, 2025. URL https://arxiv.org/abs/2504.08808.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, et al. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.16964.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, et al. Causal mediation analysis for interpreting neural NLP: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.12265.

Rand R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 2012. doi: 10.1016/C2010-0-67044-1. URL https://doi.org/10.1016/C2010-0-67044-1.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, et al. Context is key: A benchmark for forecasting with essential textual information. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66887–66944. PMLR, 2025. URL https://proceedings.mlr.press/v267/williams25a.html.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, et al. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060. URL https://doi.org/10.1016/j.neuroimage.2014.01.060.

Gerald Woo, Chenghao Liu, Akshat Kumar, et al. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Haixu Wu, Jiehui Xu, Jianmin Wang, et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.13008.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, et al. Aurora: Towards universal generative multimodal time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a8e18adf6489fd3a27417d128660eaa9-Abstract-Conference.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, et al. An explanation of in-context learning as implicit bayesian inference, 2021. URL https://arxiv.org/abs/2111.02080.

Hao Xue and Flora D. Salim. PromptCast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864, 2024. doi: 10.1109/TKDE.2023.3342137. URL https://doi.org/10.1109/TKDE.2023.3342137.

Ailing Zeng, Muxi Chen, Lei Zhang, et al. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Xiyuan Zhang, Boran Han, Haoyang Fang, et al. When does multimodality lead to better time series forecasting?, 2025. URL https://arxiv.org/abs/2506.21611.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, et al. 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. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Tian Zhou, PeiSong Niu, Xue Wang, et al. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.11939.

Xin Zhou, Weiqing Wang, Shilin Qu, et al. Unveiling the potential of text in high-dimensional time series forecasting, 2025. URL https://arxiv.org/abs/2501.07048.

A Complete Protocol and Evidence Scope

Table 2: **What each comparison establishes.** Calibration uses reference predictors; Qwen uses 240 processes; GPT-2 uses 16 matched training units. Signed scores favor correct text against a named control; larger W denotes a smaller integrated loss gap. Forecast displacement is a separate outcome. Main tests retain their symmetry assumptions; subgroup and joint-readout descriptions added after review are diagnostic. Unmet positive criteria and intervals spanning zero do not establish equivalence or absence.

Question	Evidence	Interpretation
Score calibration	Event rule 2.000; history only 0.000; surface form -0.006.	Known reference behaviors anchor the signed score.
Forecast response	Forecast displacement median 1.33984.	Text changes forecasts; displacement is distinct from a loss gap.
Event direction	Cosine mean 0.3557; negative mean for capacity constraints.	Pre-specified sum tests meet their criteria under sign symmetry.
No-event benefit	Signed mean -0.167744; Holm $p = 0.995060$.	Positive benefit criterion not met; harm not established.
Meaning selectivity	Paraphrase median 0; time-shift results on 80 processes.	Exploratory controls do not establish selectivity; no global conjunction was pre-specified.
Embedding source	Both original and permuted sources yield larger W than random embeddings in 16/16 units.	Pretrained-source contribution to the integrated unsigned loss gap.
Token correspondence	Original minus permuted mean $\Delta W = -0.016489$; Holm $p = 0.959778$.	Original-correspondence advantage not established; not an equivalence result.
Installed embeddings	Within-model replacements; archived summary interval spans zero.	Conditional baseline is not fully specified; no isolated installed effect is claimed.
Total source effect	Initial state and data order matched in 16/16 units; W changes.	Combines training and evaluation effects, not history alone.
Training interaction	Exploratory final-state summary interval spans zero.	The recorded subtraction convention is incomplete; no interaction mechanism is established.

A.1 Study records and pre-specification

The primary Qwen and GPT-2 comparisons were specified before their corresponding model results were examined. Review-driven analyses are identified separately in Table 7. The preserved study records document these specifications and their subsequent numerical verification. Here, "pre-specified" describes that timing; it is not a claim of registration in an external public registry. Model versions and evaluation membership are documented in the accompanying reproducibility records.

A.2 Controlled event processes and text conditions

The direct-prompt study used 240 synthetic base processes: 80 level-shift processes, 80 timed-interruption processes, and 80 capacity-limited processes. Paired construction holds the numerical history and background noise fixed while changing the event. Each example contains a 96-point history and a 24-point forecast horizon. Six text conditions were defined: correct event text, matched

incorrect event text, no-event-information text, a meaning-preserving paraphrase, a time-shifted event description, and format-matched irrelevant text. The design contains 1,440 potential process–condition combinations. Time-shifted descriptions apply only to the 80 timed-interruption processes; they are inapplicable to the 80 level-shift and 80 capacity-limited processes, for 160 inapplicable combinations and 1,280 evaluations. In the original 1,280 evaluations, text-condition losses share the future designated as the correct event world; the direction contrast uses both worlds. The follow-up computes a normalized loss contrast separately in each event world and then averages the two scores within each process. Raw two-world MSE averages are a separate readout (Appendix L).

Reference-world assignment and role averaging. The original reference-world assignment was not exactly counterbalanced. Level changes contain 36 negative-sign and 44 positive-sign processes, timed interruptions 44 and 36, and capacity constraints 35 and 45, for 115 and 125 overall. The deterministic process-generation rule and both saved futures are included in the supplement; these counts alone do not establish a randomized assignment design.

A post-review calculation evaluates both stored event-conditioned forecasts against each saved future. Write

$$A_i(f, g; y) = \frac{2\{L(g, y) - L(f, y)\}}{L(g, y) + L(f, y) + 10^{-15}}, \tag{8}$$

$$R_i = \tfrac{1}{2}\{A_i(f_i^c, f_i^w; y_i^c) + A_i(f_i^w, f_i^c; y_i^w)\}. \tag{9}$$

The mean of R_i over 240 processes is 0.191434, with family means 0.665269, 0.128554, and -0.219522 for level changes, timed interruptions, and capacity constraints. Omitting level changes gives -0.045484. These saved-output descriptions do not replace the original single-reference-world comparison or supply new hypothesis tests. Averaging the two roles leaves 240 process units, not 480 independent observations, and does not establish sign-flip exchangeability or broader event-family generalization.

A.3 Dependence and inference units

Text conditions within a base process are paired measurements. The 240 generated base processes are the direct-prompt inference units; neither 1,280 prompt evaluations nor the 24 points in each forecast constitute additional independent observations. The sign-flip procedure uses the sum of process-level effects and assumes that their joint null distribution is invariant under the sampled sign changes. Independent, zero-centered symmetric effects would justify that reference distribution; independence or a zero mean alone would not. For the direction cosine, simultaneously exchanging both futures and forecasts leaves the statistic unchanged, so this exchange is not a design-based justification for sign flips.

For GPT-2, the 16 paired units are the two training splits crossed with eight initialization repetitions. The same evaluation processes and six training times are reused within each trajectory. The procedure enumerates all 2^{16} signs of the matched-unit contrasts. Its probabilities are computationally exhaustive but retain a sign-symmetry model assumption because the source-condition labels were not randomly assigned across units. The reported intervals and tests are conditional on this study's matched-unit structure; they do not establish generalization to arbitrary training samples or model families.

B Metric Definitions

Let $L_{c,i}$, $L_{w,i}$, and $L_{0,i}$ be the mean-squared forecast losses for correct event text, matched incorrect event text, and no-event-information text on base process i. All have squared target units. With $\epsilon = 10^{-15}$,

$$S_i = \frac{2(L_{w,i} - L_{c,i})}{L_{c,i} + L_{w,i} + \epsilon}, \qquad D_i = \frac{2|L_{c,i} - L_{w,i}|}{L_{c,i} + L_{w,i} + \epsilon}, \tag{10}$$

$$S_{0,i} = \frac{2(L_{0,i} - L_{c,i})}{L_{c,i} + L_{0,i} + \epsilon}. \tag{11}$$

$S_i > 0$ favors the correct description; D_i is the unsigned normalized correct–incorrect loss gap, not a direct forecast-change measure. For the direction score, let $\hat{\mathbf{y}}_{c,i}$ and $\hat{\mathbf{y}}_{w,i}$ denote forecasts under

correct and incorrect descriptions, and let $\mathbf{y}_{c,i}$ and $\mathbf{y}_{w,i}$ be futures under the two defined event worlds. Then

$$C_i = \frac{(\hat{\mathbf{y}}_{c,i} - \hat{\mathbf{y}}_{w,i})^\top (\mathbf{y}_{c,i} - \mathbf{y}_{w,i})}{\|\hat{\mathbf{y}}_{c,i} - \hat{\mathbf{y}}_{w,i}\|_2 \, \|\mathbf{y}_{c,i} - \mathbf{y}_{w,i}\|_2 + \epsilon}. \tag{12}$$

The normalized forecast displacement is

$$P_i = \frac{24^{-1} \sum_{h=1}^{24} |\hat{y}_{c,i,h} - \hat{y}_{w,i,h}|}{\frac{1}{2}(24^{-1} \sum_{h=1}^{24} |\hat{y}_{c,i,h}|) + \frac{1}{2}(24^{-1} \sum_{h=1}^{24} |\hat{y}_{w,i,h}|) + \epsilon}. \tag{13}$$

S, D, S_0, C, and P are dimensionless. Medians and BCa intervals are descriptive summaries; the pre-specified one-sided Monte Carlo tests use the sum of the 240 paired unit effects.

Let $L_{p,i}$ and $L_{\text{shift},i}$ denote losses under the meaning-preserving paraphrase and time-shifted text. The exploratory expression controls are

$$S_i^{c:p} = \frac{2(L_{p,i} - L_{c,i})}{L_{p,i} + L_{c,i} + \epsilon}, \qquad Q_i = S_i - S_i^{c:p}, \tag{14}$$

$$T_i = \frac{2(L_{\text{shift},i} - L_{c,i})}{L_{\text{shift},i} + L_{c,i} + \epsilon}. \tag{15}$$

Q_i is a difference between two signed loss contrasts, not a direct forecast-invariance measure. T_i is defined only for the 80 timed-interruption processes; the time-shift condition is not applicable to the other 160 processes. Neither quantity was part of a pre-specified global semantic conjunction.

For GPT-2, let u index one matched training unit, r one of the 240 evaluation processes, a an embedding-source condition, and t a training update. Define

$$d_{u,r,a}(t) = \frac{2|L_{u,r,c,a}(t) - L_{u,r,w,a}(t)|}{L_{u,r,c,a}(t) + L_{u,r,w,a}(t) + \epsilon}, \tag{16}$$

$$m_{u,a}(t) = \text{median}_{r=1}^{240} \, d_{u,r,a}(t). \tag{17}$$

On $(t_0, \ldots, t_5) = (0, 10, 25, 50, 100, 200)$,

$$W_{u,a} = -\sum_{j=0}^{4} \frac{t_{j+1} - t_j}{200} \frac{m_{u,a}(t_j) + m_{u,a}(t_{j+1})}{2}, \qquad \Delta W_u(a,b) = W_{u,a} - W_{u,b}. \tag{18}$$

The median is taken over evaluation processes at each update, then integrated within a training unit, and only then differenced between source conditions and summarized across units. This ordering agrees with the main-text definition and avoids conflating evaluation processes with training repetitions. Larger W means a smaller integrated absolute loss gap, not higher forecasting accuracy or larger forecast displacement.

C Calibration Details

The five deterministic calibration predictors provide known anchors. Mean signed scores are 2.000000 for the event-rule predictor, 0 for history-only, -0.005756 for surface form, -2.000000 for reversed event, and 0.837618 for the over-amplified predictor. The calibration archive contains 6,400 prediction-condition records and 1,200 derived paired summaries from 240 base processes; records sharing a process are dependent. Two separate calculations agree on these reference-behavior checks. The values characterize the specified predictors, not trained-model performance. The supplement reproduces these recorded pair-score means; reconstructing them from saved losses with the direct-prompt epsilon differs at about 1.8×10^{-12} for the event-rule anchors, with no effect on displayed values. The exact historical calibration epsilon is not inferred solely from numerical agreement.

D Qwen Conditions and Statistics

Table 3 gives the three pre-specified comparisons. BCa intervals summarize medians. The one-sided Monte Carlo test uses the sum statistic and $B = 1{,}048{,}576$ draws from the 64-bit permuted

Table 3: **Qwen outcomes by comparison and event family.** Original comparisons use 240 designated-reference worlds and show 95% BCa median intervals. Follow-ups average two world-specific normalized scores within each of the same 240 processes; their intervals target means. The separate capacity correction row uses R_i, correct versus incorrect text in both roles, on only 80 processes. Positive values favor event-conditioned forecasts; Signs gives positive/negative/zero counts. Family rows are descriptive.

Comparison	**Mean**	**Median**	**Interval and target**	**Signs** $+/-/0$
Correct vs. incorrect, $S^{c:w}$	0.2416	0.0715	$[0, 0.4152]$; median	$126/96/18$
Event-direction cosine, C	0.3557	0.4975	$[0.3973, 0.6076]$; median	$163/57/20$
Correct vs. no-event, $S^{c:0}$	-0.1677	-0.0498	$[-0.1863, 0]$; median	$104/128/8$
Follow-up baselines: $\bar{S}_i(b)$, 240 processes				
Event vs. no auxiliary context	-0.1283	-0.0897	$[-0.2380, -0.0154]$; mean	$106/132/2$
Event vs. metadata only	-0.1579	-0.2618	$[-0.2764, -0.0418]$; mean	$95/144/1$
Capacity: role-averaged correct vs. incorrect after article correction, R_i ($n = 80$)	-0.1908	-0.0762	descriptive only	$20/46/14$
Family means ($n = 80$ each)	$S^{c:w}$	C	$S^{c:0}$	
Level change	0.7110	0.9020	-0.2311	
Timed interruption	0.1490	0.3574	-0.3685	
Capacity constraint	-0.1351	-0.1924	0.0963	

Table 4: **Qwen sum tests, separate from the median intervals in Figure 3.** For $B = 1{,}048{,}576$ sign assignments, b counts sums at least as large as observed. The initial b/B is retained for traceability; inference uses $(b + 1)/(B + 1)$ in every row and Holm adjustment across three comparisons at $\alpha = .05$. Under sign symmetry, the first two meet the criterion and the no-event comparison does not.

Comparison	b	**Initial** b/B	**Corrected** p	**Holm** p
Correct vs. incorrect	591	5.63622×10^{-4}	5.64575×10^{-4}	1.12915×10^{-3}
Event-direction cosine	0	0	9.53673×10^{-7}	2.86102×10^{-6}
Correct vs. no-event	1,043,396	0.995060	0.995060	0.995060

congruential generator (PCG64), with seed 42. Table 4 retains the initially reported b/B values beside the uniformly corrected $(b + 1)/(B + 1)$ estimates. Holm correction is recomputed across these three comparisons.

The initial b/B estimates, rounded for display in table order, are

$$(0.000564, \quad 0, \quad 0.995060).$$

The first two pre-specified positive-direction sum tests satisfy the Holm rule; that statement is distinct from a claim that each median is significantly above zero. The no-event comparison does not satisfy its positive-direction condition, which does not establish harm.

Normalized L_1 forecast displacement has median 1.339841 with BCa interval $[1.169412, 1.457796]$. The unsigned correct–incorrect gap has median 0.996852 with interval $[0.805681, 1.104487]$. Paraphrase selectivity has mean -0.014290 and median 0, with interval $[-0.179697, 0]$. Time-shift selectivity, defined for 80 base processes, has median 0.291271 and interval $[0, 0.588924]$. The post-review correct-versus-format-matched-irrelevant contrast has mean -0.178524, median -0.053953, and 108/127/5 positive/negative/zero effects. These exploratory controls do not replace the pre-specified comparison family. No overall semantic-selectivity conjunction was pre-specified for the Qwen experiment.

All fixed event families remain in the primary aggregate. Correct-versus-incorrect mean/median/sign counts are $0.710965/1.352526/58{:}22{:}0$ for level changes, $0.149014/0.097420/44{:}31{:}5$ for timed interruptions, and $-0.135115/-0.069665/24{:}43{:}13$ for capacity constraints. Direction-cosine values are $0.901975/0.951354/79{:}1{:}0$, $0.357424/0.438253/64{:}9{:}7$, and $-0.192440/-0.283617/20{:}47{:}13$, respectively. Correct-versus-no-event values are $-0.231106/-0.178595/35{:}45{:}0$, $-0.368453/-0.293336/29{:}51{:}0$, and $0.096327/0.005530/40{:}32{:}8$. This is descriptive heterogeneity; no post-review family interaction test or family exclusion is introduced.

Raw-loss diagnostics show that the correct-plus-incorrect denominator has minimum 0.151525, first percentile 0.353431, median 6.354268, and no values below 10^{-6}. The unnormalized incorrect-minus-correct loss difference has mean 7.667477 and median 0.077644, while its log loss ratio has mean 0.424650 and median 0.071505. Correct-text loss has median 1.778764 and mean 17.950531; matched-incorrect loss has median 2.959011 and mean 25.618008. Neutral no-event and format-matched irrelevant losses have medians 1.500219 and 1.603782 and means 2.598976 and 2.626908. Small denominators do not drive the aggregate, but normalization still weights processes differently from an absolute loss difference, and the heavy right tails separate mean from median estimands.

E GPT-2 Matched-Training Design

The two training splits contain 133 training examples each. Their 266 training-example occurrences comprise 148 Time-MMD occurrences and 118 Context is Key occurrences; the Time-MMD portion represents 80 unique physical windows, some present in both splits. Three source conditions are evaluated across the two splits and eight initialization repetitions, producing 48 training trajectories. Each trajectory records updates 0, 10, 25, 50, 100, and 200, totaling 288 states. The three arms differ only in the target token-embedding initialization: original pretrained token embeddings, within-unit permuted pretrained token embeddings, or randomly initialized token embeddings drawn with standard deviation 0.02 (variance 0.02^2). Each of the 16 matched units uses a distinct pre-specified vocabulary-row derangement: a bijective permutation with no fixed positions, held constant within the unit. Since derangement identity changes with the matched training unit, these observations do not isolate permutation variation from fold and initialization repetition. In each condition, the selected token-embedding matrix and original GPT-2 backbone weights are frozen; backbone stochastic layers are disabled. This does not freeze the added forecasting and text-conditioning parameters. The trainable text-conditioning path contains 6,438,913 parameters, including 6,433,792 cross-attention/low-rank-adaptation parameters. Within a split-by-repetition unit, non-target initialization, data order, optimization settings, and evaluation conditions are matched. Training uses 200 AdamW updates, learning rate 10^{-3}, weight decay 0.01, no learning-rate scheduler, and gradient-norm clipping at 1.

Architecture detail retained in this report. Cross-attention and low-rank-adaptation parameters belong to the trainable text-conditioning path, rather than to the frozen pretrained weights. The available specification gives their aggregate parameter count but not a layer-by-layer placement, adaptation rank, or a complete numerical encoder and prediction-head definition. Those details cannot be recovered uniquely from parameter counts. We therefore distinguish reconstruction from saved losses from a fully specified model reimplementation (Appendix M).

Normalized loss and the remaining scaling convention. Let $s_i = \mathrm{sd}(x_i)$ use history only, and write

$$\mathrm{MSE}_i(f, y) = \frac{1}{24}\sum_{h=1}^{24}(f_h - y_h)^2, \qquad L_i(f, y) = \frac{\mathrm{MSE}_i(f, y)}{a_i}. \tag{19}$$

Two distinct standard-deviation conventions are possible: $a_i = s_i$ divides the already squared error by the standard deviation, whereas $a_i = s_i^2$ equals $24^{-1}\sum_h((f_h - y_h)/s_i)^2$. The retained wording does not uniquely specify which was implemented, the variance divisor, or zero-scale handling. We do not substitute an unverified convention: the reported GPT-2 scores use the preserved normalized losses. Exact normalized-error re-evaluation requires that implementation detail in addition to this description.

Table 5: **GPT-2 source contrasts and final-state readouts.** Original and Permuted use pretrained vectors with original and within-unit permuted token correspondence; Random uses random vectors. The upper block reports learning-curve ΔW over 16 matched units, with descriptive 95% Student t mean intervals and Holm probabilities from 2^{16} sign assignments at $\alpha = .025$ under sign symmetry. Positive ΔW denotes a smaller integrated loss gap, not lower error. The lower block averages final-state readouts within each evaluation set and then across units; it is a separate post-review description after 200 updates.

Source contrast	**Mean ΔW [95% interval]**	**Median ΔW**	**Positive units**	**Holm p**
Original − Random	0.265906 $[0.223408, 0.308404]$	0.267397	16/16	4.58×10^{-5}
Permuted − Random	0.282395 $[0.237172, 0.327618]$	0.272592	16/16	4.58×10^{-5}
Original − Permuted	-0.016489 $[-0.035559, 0.002580]$	-0.011418	3/16	0.959778

Final source	**Correct-text normalized error**	**Mean signed score**	**Mean displacement**
Original	2.573928	0.017080	0.093866
Permuted	4.179299	0.023194	0.071804
Random	3.004392	0.052314	0.507451

For both pretrained-versus-random contrasts, the exact probabilities are

$$p_{\text{raw}} = \frac{1}{65{,}536} \approx 1.52588 \times 10^{-5}, \qquad p_{\text{Holm}} = \frac{3}{65{,}536} \approx 4.57764 \times 10^{-5}.$$

For original versus permuted, the one-sided probability is $62{,}900/65{,}536 \approx 0.959778$.

Both pretrained-source conditions yield larger W than random embeddings under the specified within-unit derangements. Original-versus-permuted is a positive-direction comparison, not an equivalence test; its result leaves correspondence effects uncertain. Since permutation and training repetition vary together, it also does not isolate robustness to permutation alone.

F Exploratory Embedding Replacement After 200 Updates

At update 200, original and within-unit permuted token-embedding matrices were exchanged between their corresponding trained models without rescaling or reinitialization. Each replacement leaves all other parameters of its recipient model unchanged. Before adaptation, matched conditions share the same non-target initialization and data order; learned parameter differences can therefore be consequences of the source intervention.

Four conditions and conditional contrasts. Let $q_u(h, e)$ be the negative median unsigned correct/incorrect loss gap at update 200 in unit u, where $h \in \{O, P\}$ denotes the training source and $e \in \{O, P\}$ the currently installed embeddings. O means original pretrained and P means the unit's

fixed permutation. The layout is

	Install O	Install P
Train O	$q_u(O,O)$	$q_u(O,P)$
Train P	$q_u(P,O)$	$q_u(P,P)$

. (20)

Explicit conditional history and installation contrasts, with original minus permuted orientation, are

$$H_u(e) = q_u(O,e) - q_u(P,e), \quad e \in \{O,P\}, \tag{21}$$

$$E_u(h) = q_u(h,O) - q_u(h,P), \quad h \in \{O,P\}, \tag{22}$$

$$\begin{aligned} I_u &= H_u(O) - H_u(P) = E_u(O) - E_u(P) \\ &= q_u(O,O) - q_u(P,O) - q_u(O,P) + q_u(P,P). \end{aligned} \tag{23}$$

Averaged main effects, $\{H_u(O) + H_u(P)\}/2$ and $\{E_u(O) + E_u(P)\}/2$, are different quantities from either conditional contrast. None is the learning-curve integral W.

Reported summary labels and identification limit. The retained summary labeled training history has mean -0.016908, median -0.010691, and 95% mean interval $[-0.049306, 0.015490]$. The installed-embedding summary has mean -0.020179, median -0.005076, and interval $[-0.053295, 0.012938]$. The difference-in-differences summary has mean -0.047927, median -0.021447, and interval $[-0.110165, 0.014312]$.

These aggregate summaries do not specify their conditioning baseline, whether the main effects were averaged, or the interaction's subtraction order. Accordingly, the definitions above distinguish the possible contrasts but do not retrospectively assign these three reported numbers to an unverified formula. We retain them as exploratory summaries outside the primary Holm family; all reported intervals span zero, and no conditional component or semantic mechanism is established from them.

G Historical Cross-Model Evidence

Historical studies used different endpoints and expected directions and are not pooled with the current analyses. An earlier Pythia narrow-replacement replication observed median -0.139881, with 46 of 48 units in its prescribed negative direction and $p = 4.181544 \times 10^{-12}$. A later fresh-evaluation replication observed median -0.192827, with 41 of 49 units in the prescribed negative direction, both folds and all five seeds negative, and $p = 9.823269 \times 10^{-7}$. These replicate a narrow replacement behavior. Stronger pre-specified conditions concerning alignment, stability across pretraining runs, and trajectory-wide behavior were not met. A historical Qwen3 cross-family comparison instead had median 0.040274, 20 of 48 units negative, and one-sided $p = 0.903294$, opposite to its prescribed negative direction. The historical evidence therefore supports the reproducibility of a narrow Pythia behavior, not a model-general mechanism.

Limits of historical reconstruction. The retained summaries identify the models, expected directions, counts, and reported probabilities, but do not give the historical endpoint formulas, unit-membership definitions, or exact replacement operations. We therefore report them as historical context rather than defining their units by analogy with the current experiments or treating them as a common-effect replication. "Negative expected" denotes the historical orientation, not a favorable value of the present signed score.

Table 6: **Cross-model evidence under the original comparisons.** Historical comparisons expect negative effects; this orientation is not that of the present signed score. Both narrow Pythia replications meet their criteria, unlike stronger Pythia conditions and the Qwen3 comparison. Historical endpoint formulas and unit memberships are not fully specified by the retained summaries; these values are context, not pooled evidence.

Study	Result and expected direction	Interpretation
Pythia, earlier	Negative expected; median -0.139881; 46/48 negative; $p = 4.18154 \times 10^{-12}$.	Narrow prescribed behavior reproduced.
Pythia, later	Negative expected; median -0.192827; 41/49 negative; both splits and all five seeds negative; $p = 9.82327 \times 10^{-7}$.	A second narrow effect was reported.
Pythia, broader tests	Alignment, stability across pretraining runs, and trajectory-wide criteria not met.	Narrow replication does not establish the broader claim.
Historical Qwen3	Negative expected; median 0.040274; 20/48 negative; one-sided $p = 0.903294$.	Prescribed direction not reproduced.

H Prospective and Diagnostic Status

The present Qwen signed comparisons and GPT-2 learning-curve comparisons were specified before their model results were examined. An earlier ΔS reanalysis of saved losses from three evaluation datasets was retrospective; it is not the present Qwen experiment. The embedding-replacement analysis after 200 updates is pre-specified exploratory secondary evidence.

Table 7: **Timing and inferential role of the analyses.** "Review" denotes pre-submission scrutiny, not ICLR 2027 official review. Planning new conditions before their outputs does not make an already inspected cohort an independent replication. This table groups the status qualifications used throughout the paper.

Analysis	When specified	Statistical role	Location
Original Qwen comparisons	Before the corresponding model outputs were inspected	Median intervals; separate Holm-adjusted sum tests under sign symmetry	Sec. 3; App. D
GPT-2 learning curves	Before their model outputs were inspected	Matched-unit contrasts, exhaustive sign-flip reference distribution	App. E
Final embedding replacement	Pre-specified exploratory secondary analysis	Separate from the primary Holm family	App. F
Role averaging, family sensitivity, joint readouts	After original results were inspected	Descriptive and sensitivity analyses; not new primary decisions	App. A, K
No auxiliary context, metadata, article correction	Plan fixed before new-condition outputs; cohort already inspected	Supplementary baseline means and intervals; descriptive wording check	App. L
Earlier saved-loss ΔS diagnostic	Retrospective	Not the present original Qwen experiment	App. H

I Independent Recomputation and Unexecuted Studies

An independent calculation reconstructed Qwen loss-derived metrics from 1,280 evaluations and 240 base processes and recomputed GPT-2 sign-flip results from the saved matched-unit effects. The independent results agree with the preserved numerical records within reported precision; numerical agreement does not establish the inferential assumptions. Uniformly replacing b/B by $(b+1)/(B+1)$ and recomputing Holm probabilities leaves the pre-specified decisions unchanged.

Studies that were not conducted provide no model-performance evidence and are not counted as negative findings.

J Reproduction Scope

Statistical recomputation uses the separately retained forecasts, targets, unit-level results, and analysis definitions. Repeating model evaluation additionally requires the specified trained model states; repeating training requires the training examples, splits, initial states, and complete optimization settings. The rendered summary tables alone do not recover these experimental inputs.

K Post-Review Sensitivity of Aggregate and Family Conclusions

This section analyzes existing predictions after the original results and review questions were known. No model, prompt, task, or primary decision changes. The estimand is the equally weighted mean over the three studied event families, treated as fixed strata. It is not a mean over a sampled population of event families.

Within-family bootstrap. We resample 80 base processes with replacement within each family, retaining all three outcomes for a selected process together. The three family means are averaged with equal weights. We use 20,000 replicates, PCG64 seed 20260916, and percentile endpoints at $0.05/(2 \cdot 3)$ and $1 - 0.05/(2 \cdot 3)$. Bonferroni allocation gives approximate simultaneous 95% mean intervals for the three outcomes, conditional on the process-sampling assumptions. These are not the pre-specified BCa median intervals and are not exact randomization intervals.

For $(S^{c:w}, C, S^{c:0})$, the simultaneous bootstrap intervals are

$$[0.0751, 0.4024], \qquad [0.2958, 0.4146], \qquad [-0.3162, -0.0176].$$

The negative last interval is a retrospective average contrast with this specific neutral instruction, not a replacement for the original one-sided benefit test or a universal harm claim.

A bound not requiring sign symmetry. For independent $X_i \in [a, b]$, let $\mu = n^{-1} \sum_i \mathbb{E}X_i$. Hoeffding's inequality gives

$$\Pr\{|\bar{X} - \mu| \geq r\} \leq 2 \exp\left(-\frac{2nr^2}{(b-a)^2}\right).$$

Applying a union bound to $m = 3$ outcomes and solving at $\alpha = .05$ yields

$$r = (b-a)\sqrt{\frac{\log(2m/\alpha)}{2n}}.$$

With $n = 240$, the radii are 0.39948 for both signed outcomes and 0.19974 for cosine. The simultaneous intervals, intersected with the known ranges, are

$$[-0.1579, 0.6411], \qquad [0.1559, 0.5554], \qquad [-0.5672, 0.2317].$$

No identical-distribution or symmetry condition is required (Hoeffding, 1963), but independence of process effects is still assumed. The calculation supplies a conservative sensitivity comparison, not a new prospective experiment. Only the direction interval excludes zero. Thus a positive stratified signed-benefit interval must not be presented as agreement across every assumption set.

All-family omission summaries. All 240 processes remain in the main result. Omitting capacity constraints gives correct-versus-incorrect mean/median $0.429989/0.586536$; omitting level changes gives $0.006949/0$; omitting timed interruptions gives $0.287925/0.045690$. The corresponding direction-cosine mean/median pairs are $0.629700/0.776482$, $0.082492/0.063197$, and $0.354768/0.797308$. Complete values for the no-event comparison and sign counts appear in the numerical supplement. These are descriptions of dependence on the chosen family mixture, not new subgroup significance tests or a prospective prediction of failure mechanisms.

Sign-count alternative. For completeness, the sign-count comparison raised during review targets a different null: a common probability $1/2$ for a positive sign conditional on a nonzero effect. With zeros excluded, the counts are $126/222$, $163/220$, and $104/232$. Raw one-sided probabilities are 0.02568, 2.54×10^{-13}, and 0.94974; Holm adjustment across all three yields 0.05137, 7.62×10^{-13}, and 0.94974. This common fair-sign model is not an assumption-free test of a heterogeneous mean. In particular, the first raw probability alone is not reported as corroborating confirmation.

Joint GPT-2 readouts. The lower block of Table 5 averages each final-state readout over 240 evaluation processes within a training unit, then over the 16 units. It describes update 200 only. Original and permuted embeddings both yield the primary curve-score difference, yet original correct-text error is below the random reference while permuted error is above it. The random source has the largest final signed score and displacement. These complementary readouts demonstrate an ordering difference, not a new hypothesis test or separate attribution to geometry. Saved losses at every training time are included for numerical reconstruction.

L Review-Driven Auxiliary-Context Follow-up

The follow-up used the same fixed 240 processes and the same fixed Qwen2.5-7B-Instruct version. Its evaluation prompts were specified before these outputs were generated. It contains 240 no-auxiliary-context prompts, 240 metadata-only prompts, and the correct and incorrect article-normalized capacity

prompts for 80 processes. Of the latter 160 prompt conditions, 90 contain exactly the same messages as the original prompts and reuse their saved predictions after a repeat evaluation of 24 original prompts matched every forecast value exactly; the other 70 were generated anew. In total, 550 forecasts were newly generated and 90 were reused. All 640 prompt evaluations produced finite 24-step forecasts, with no sample replacement or training.

For no auxiliary context, the all-process mean/median world-averaged score is $-0.128311/-0.089691$; the family means for capacity constraints, level changes, and timed interruptions are 0.010147, -0.218083, and -0.176998. For metadata only, the overall mean/median is $-0.157928/-0.261793$, with family means 0.011631, -0.282788, and -0.202626. The approximate stratified simultaneous intervals are $[-0.238007, -0.015417]$ and $[-0.276363, -0.041840]$. The corresponding simultaneous Hoeffding intervals are $[-0.510499, 0.253877]$ and $[-0.540115, 0.224260]$. Both procedures target the same mean world-averaged score over the specified process population. They differ in their assumptions and coverage guarantees: the stratified percentile intervals are approximate, while the Hoeffding intervals are conservative under independence and boundedness. Both are retained.

Squared-loss decomposition and averaging. Define the within-process centers and half-differences

$$\bar{f}_i = \tfrac{1}{2}(f_i^c + f_i^w), \quad d_i = \tfrac{1}{2}(f_i^c - f_i^w), \qquad \bar{y}_i = \tfrac{1}{2}(y_i^c + y_i^w), \quad e_i = \tfrac{1}{2}(y_i^c - y_i^w).$$

For $H = 24$, write $\langle a, b\rangle_H = H^{-1}\sum_{h=1}^{H} a_h b_h$ and $\|a\|_H^2 = \langle a, a\rangle_H$. Baseline $g_{i,b}$ is the same forecast in both event worlds. The raw world-averaged MSE advantage is exactly

$$\begin{aligned} A_i^{\mathrm{MSE}}(b) &= \tfrac{1}{2}\sum_{k\in\{c,w\}} \left[\|g_{i,b} - y_i^k\|_H^2 - \|f_i^k - y_i^k\|_H^2\right] \\ &= \underbrace{2\langle d_i, e_i\rangle_H}_{\text{event alignment}} \quad \underbrace{-\|d_i\|_H^2}_{\text{response magnitude}} \\ &\quad + \underbrace{\|g_{i,b} - \bar{y}_i\|_H^2 - \|\bar{f}_i - \bar{y}_i\|_H^2}_{\text{center-error difference}}. \end{aligned} \tag{24}$$

The identity follows by expanding the two squares in each pair: the terms linear in the world sign cancel, and the shared $\|e_i\|_H^2$ cancels between the baseline and event-conditioned losses. Each reported component is its arithmetic mean over 240 base processes. The factor H^{-1} is already included in every term; two worlds do not create additional process units. This decomposition concerns raw squared-error differences, not the average of normalized ratios $\bar{S}_i(b)$ or R_i.

The raw two-world average mean squared error (MSE) advantage of event-conditioned over baseline forecasts is negative on average for both controls: -15.037152 against no auxiliary context and -15.078976 against metadata only. In the specified algebraic decomposition, the mean event-alignment component is 3.425341 and the response-magnitude component is -13.001120 under either baseline; the center-error component is -5.461373 and -5.503197, respectively. Numerical decomposition residuals are at floating-point scale. These terms are algebraically related components, not independent confirmations.

The capacity wording check changes a forecast value in 28 of 80 processes. Its role-averaged correct-versus-incorrect score R_i (Eq. (9)), before article correction, has mean/median $-0.219522/-0.112760$; afterward, $-0.190806/-0.076206$. The mean score change is 0.028716. Direction cosine changes from mean/median $-0.192440/-0.283617$ to $-0.161687/-0.168720$. The normalization addresses two literal article errors only and does not isolate semantics from every surface feature.

M Minimal Input and Reproduction Description

Prediction input. The system message is "You are a useful forecasting assistant." The user prompt provides a context block containing the scenario, a history block with 96 timestamp–value pairs, and the 24 requested future labels. It asks for ordered timestamp–value pairs inside forecast tags and no commentary. Numeric history values use six significant digits. The neutral scenario is exactly "No additional event is supplied for the forecast period." Correct, incorrect, paraphrase, irrelevant,

and applicable time-shifted descriptions are preserved verbatim in the numerical supplement. The review-driven follow-up additionally evaluates a prompt with the entire auxiliary context block removed and a prompt whose context contains only history length, horizon length, and ordinal-label metadata. These are new fixed-condition outputs, not values inferred by editing saved forecasts.

Model and computation. The direct-prompt model is Qwen2.5-7B-Instruct; its fixed model version, full source identifier, and tokenizer version are provided in the supplement. Inference used bfloat16 (BF16) precision with greedy constrained generation, a 4,096-token generation budget, and batch size 8. The recorded system used an NVIDIA RTX PRO 6000 Blackwell Server Edition, Python 3.12.3, PyTorch 2.8.0 with CUDA 12.8, Transformers 4.45.2, and tokenizers 0.20.1. These reproducibility details do not change any scientific comparison. The recorded estimate of computation time is not presented as measured evaluation time.

Available data, code, and model parameters. The anonymous numerical supplement contains 240 synthetic histories and future pairs, 1,440 original condition definitions, 1,280 original Qwen forecasts, 640 follow-up inputs and forecasts, exact prompt and parser definitions, and 69,120 GPT-2 loss records. Its code reconstructs loss-derived metrics, curve scores, and the labeled post-review analyses. It does not provide all trained GPT-2 tensors, the combined Time-MMD/Context is Key training corpus, or a complete retraining environment; model re-evaluation and retraining are not tested reproduction levels. Retained Time-MMD file/window records do not establish the upstream revision or redistribution license. The third-party corpus is not redistributed.