Context-Gain Concentration Is Not Evidence of Routable Forecast Value

Anonymous authors
Paper under double-blind review

Abstract

A steep context-gain concentration curve, in which a few forecast origins carry most of the realized gain from adding context, invites the conclusion that context should be routed to those origins. We show that it is not evidence of routable value: two settings can have identical distributions of realized gains, and hence identical curves, while no policy can exploit the gain in one and a simple policy can in the other. The quantity that decides routing is the covariance between a pre-decision policy and conditional context value, which we estimate as matched-budget net gain. Across energy, retail, and finance/news with two zero-shot and one trained forecasting backbone, opening context at the router's top 20% of predicted gain beats random selection of the same number of origins in 4 of 6 zero-shot domain–backbone pairs, covering all 3 domains, and beats selection by forecast uncertainty in 3. These counts come from separate per-pair 95% intervals without multiplicity adjustment, from an exploratory, post-development analysis on data that shaped the method, and the retail intervals rest on few independent time blocks; they are not a confirmed cross-domain effect. The signs hold over time within the evaluation fold and under an online threshold. A simulation with constant context value reproduces a median 0.731 of the observed concentration, so the curve alone says little about where context pays.

1 Introduction

Textual and exogenous context helps some forecasts and hurts others. A natural argument for routing it, buying or using context only at some origins, ranks origins by realized context gain and shows that a small top fraction carries most of the positive gain. That curve cannot carry the argument. Ranking and summing the same noisy gains selects favourable noise, so the curve is steep even when context value is identical at every origin (Figure 1). More fundamentally, the curve is a function of realized gains alone, whereas routing depends on what is known before the outcome: two settings can share every realized-gain statistic and still differ in whether any policy can exploit the gain (Proposition 1).

The practical consequence is an evaluation error. Reading a curve can credit routable value where none exists, and it cannot tell a learned router from a rule that selects on forecast uncertainty. The quantity that decides routing is the gain of a pre-decision policy over random selection at the same budget, which equals the covariance between the decision and conditional context value (Lemma 1). We measure it directly, against random, uncertainty-based, and change-point-based selection at the same budget, with paired intervals that resample time blocks jointly across series.

Findings. Opening context at the router's top 20% beats random selection of the same number of origins in 4 of 6 zero-shot domain–backbone pairs, in all 3 domains, and beats uncertainty-based selection in 3. The intervals are per pair and unadjusted for multiplicity, the retail intervals rest on few independent time blocks, the analysis is exploratory and post-development, and no untouched data exist to confirm it.

Development record. The study began from a preregistered hypothesis that the top 10%–20% of origins would carry a large majority of positive context value across domains and models; it is kept as a benchmark, not as an observed result. Because the original v1 synthetic validation failed, a post-hoc amended-v2 gate replaced it; that gate counts a cell only when a rank-transition signature

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.

and Fold-D predictability both satisfy frozen rules. Failure on either leg remains a negative result. That gate is a diagnostic, not a test of routing value; Appendix C reports it.

Contributions. This is an evaluation-method study with three contributions:

1. **A non-identification result.** Realized-gain concentration cannot identify routable value (Proposition 1), and the decision target is the covariance between a pre-decision policy and conditional value (Lemma 1): the conditional loss difference of conditional predictive ability (Giacomini & White, 2006), read through a budgeted decision.
2. **A matched-budget evaluation.** Net gain against random and pre-decision baselines at equal budget, with paired time-block intervals, in batch, online, and shared-horizon forms.
3. **Evidence with stated scope.** Matched-budget net gain is positive in every domain for at least one zero-shot backbone; the router adds to uncertainty-based selection for energy with Moirai and for both retail pairs, but not for finance/news; and constant-value simulation reproduces much of the observed concentration.

2 Method

2.1 Origin- and Horizon-Specific Context Value

Let $F^U_{t,h}$ and $F^C_{t,h}$ be univariate and context-augmented probabilistic forecasts under a lower-is-better proper score S. Realized gain is $d_{t,h} = S(F^U_{t,h}, Y_{t+h}) - S(F^C_{t,h}, Y_{t+h})$, and conditional context value is $\Delta_h(x) = \mathbb{E}[d_{t,h} \mid X_t = x]$, where X_t is everything known when context must be bought: target history, the univariate forecast, calendar, and availability, but not the context itself. This is the conditional loss difference of conditional predictive ability (Giacomini & White, 2006). A policy $g(X_t) \in \{0, 1\}$ opens context; its budget is $q = \mathbb{E}[g(X_t)]$.

2.2 Concentration and Selective Evaluation

A context-gain concentration curve ranks origins by realized gain and reports the share of total positive gain in the top fraction q: signed capture $G(q) = \sum_{i \in T_q} d_i / \sum_i d^+_i$ with random baseline $q \sum_i d_i / \sum_i d^+_i$, or positive-gain capture $\mathrm{PGC}(q) = \sum_{i \in T_q} d^+_i / \sum_i d^+_i$ with random baseline q, where T_q is the top-q set and $d^+ = \max(d, 0)$. A steep curve reads as "a few origins carry most of the value, so route context there." The curve cannot support that reading.

Proposition 1 (concentration does not identify routable value). Let X and Z be independent, each ± 1 with probability one half; X is observed at decision time and Z is not. Let ε be independent of (X, Z) with mean zero, and $a > 0$. Origins are independent draws of (X, Z, ε). In world A, $d = aZ + \varepsilon$; in world B, $d = aX + \varepsilon$. Then (i) d has the same distribution in both worlds, so for any sample size the realized gains, and every functional of them including G and PGC at every q, have the same distribution; (ii) $\mathbb{E}[d \mid X]$ is 0 in A and aX in B; (iii) the policy $g = \mathbf{1}\{X = 1\}$ has budget one half, and its expected gain over random selection at that budget, $\mathbb{E}[(g - \frac{1}{2})d]$, is 0 in A and $a/2$ in B.

Lemma 1 (the decision target). For any policy $g(X)$ with budget q, $\mathbb{E}[(g(X) - q)\, d] = \mathbb{E}[(g(X) - q)\, \Delta(X)] = \mathrm{Cov}(g(X), \Delta(X))$. Hence some policy beats random selection at its budget if and only if $\Delta(X)$ is not almost surely constant. Appendix A gives both proofs.

Proposition 1 separates what a curve can show from what routing needs. The curve is a statistic of realized gains; routing value is a covariance between the decision and Δ, which depends on what is observable before the outcome. Figure 1 shows the practical size of the gap. We therefore evaluate routing by the sample analogue of Lemma 1 at a fixed budget, against random and against simple pre-decision rules (Section 2.3).

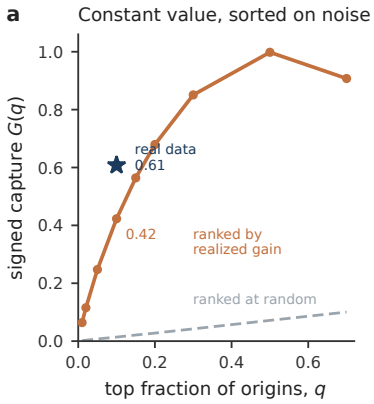

Figure 1: Sorting on noise makes a steep curve. Conditional value is identical at every simulated origin, yet ranking by realized gain gives a steep curve; the plotted curve is one representative setting. Across simulations matched to the origin count and noise of each of 276 real cells, the median signed capture G in the top tenth is 0.422, against 0.013 for random ranking; the star is the median observed capture 0.608. The matched constant-value simulation reaches a median 0.731 of the observed capture ([0.612, 0.835], cross-cell 10th–90th spread) and more than half of it in 273 cells. Real curves are therefore compatible with substantial concentration under constant value; the ratio is not a decomposition of real gains into signal and noise.

2.3 Matched-Budget Policy Value

In units $\tilde{d} = d/L_c$, where L_c is the cell's median univariate loss, the estimate for policy g on n evaluation rows is $V_{\text{rand}} = n^{-1}\sum_i (g_i - \bar{g})\tilde{d}_i$, the gain over random selection of the same number of rows. Against a baseline rule g' with the same budget, $D = n^{-1}\sum_i (g_i - g'_i)\tilde{d}_i$. The primary policy opens context at the top 20% of the router score $\widehat{\Delta}$ within a cell (one target and horizon). This is a batch evaluation: ranking uses the whole evaluation fold's scores. The *online* variant thresholds each origin at a quantile of earlier origins' scores only, and so needs no future scores; the shared-horizon variant makes one decision per origin for all horizons, matching a single purchase of context. Intervals come from one circular block bootstrap per domain–backbone pair: time blocks are shared by all cells and series of the pair, so dependence across series at a common time is kept, and contrasts between policies are paired.

2.4 Cross-Fitting and Statistical Discipline

Temporal partitions separate forecaster training, paired losses, router fitting, and evaluation. Zero-shot forecasters use only paired-loss and evaluation folds; the trained family uses disjoint training windows and is marked training-window confounded. Router labels use Fold C and evaluation uses Fold D.

2.5 Acquisition: Deciding Whether Context Is Needed

Before context is seen, the router estimates $\widehat{\Delta}_h(X_t)$ from target history, baseline uncertainty, regime metadata, horizon, availability, and calendar variables; a feature guard excludes gated content. The gradient-boosted estimator with isotonic recalibration is a measuring device, not a contribution.

2.6 Fusion Control: Deciding Whether Available Context Is Used

Fusion starts after context is available and may use its content; its information set differs from acquisition's, so its metrics cannot support acquisition claims.

2.7 Explaining a Selected Forecast Revision

Explanations target the selected revision $R_{t,h} = F^C_{t,h} - F^U_{t,h}$, not why the future occurred; deletion, permutation, and timestamp shifts test pathway, content, and alignment.

3 Experiments

3.1 Evaluation Scope and Evidence Rules

We evaluate energy (Xu et al., 2025), finance/news (Dong et al., 2024), and retail (Makridakis et al., 2022) across 3 backbones: zero-shot Ansari et al. (2024) and Woo et al. (2024), plus Nie et al. (2022) trained per cell. Zero-shot effects need only paired-loss and evaluation folds; PatchTST also needs a disjoint training window, confounding its context arm with supervised fit, so it is reported separately. CRPS is uniquely minimized in expectation by the true predictive distribution; its context-score difference measures expected-score reduction rather than rewarded misreporting (Matheson & Winkler, 1976; Gneiting & Raftery, 2007). The router, the folds, and every evaluation rule were fixed during method development on the same archives; no data untouched by development exist (Appendix J), so all results below are post-development evaluations on Fold D.

3.2 Acquisition Results

Exploratory. Table 1 is the main result. With context opened at the router's top 20% within each cell, the per-pair interval for gain over random selection at the same budget lies above zero for 4 of 6 zero-shot domain–backbone pairs, covering 3 domains, and below zero for none. The analysis plan was registered before computation, but it was written after the gate results were known, the intervals are per pair, and no multiplicity adjustment is made; the counts are exploratory evidence, not a confirmed cross-domain effect.

Is the router needed? A router that only rediscovers forecast uncertainty adds nothing. Against the top 20% by the univariate predictive-entropy proxy, the paired contrast lies above zero in 3 zero-shot pairs; against change-point probability, in 3. Energy with Moirai and both retail pairs beat both rules. Finance/news with Moirai does not: uncertainty selection collects a similar gain, so its positive V_{rand} is routable value that the router does not add to.

Deployment form. The top-fraction column ranks the whole evaluation fold at once. The online threshold, which uses only earlier scores and therefore spends a data-dependent budget, beats random selection at its own budget in 5 zero-shot pairs. Each horizon cell above decides independently; with one decision per origin shared by all horizons, 4 zero-shot pairs keep an interval above zero (Table 2).

Reliability of the intervals. The point estimates do not depend on block length; only the intervals do. Of the 4 zero-shot pairs with an interval above zero, 4 keep it there with a quarter or half of the frozen block length, which gives more resampling units, and 4 have a positive point estimate in each contiguous third of Fold D. The time thirds show that the sign is stable over the fold; they do not show that the interval widths are reliable. Retail is the weakest case for inference: its Fold D has 87 origin times, only 3 blocks at the frozen length, and longer blocks leave too few units for any interval to be informative. Its intervals, including the large retail–Moirai estimate, therefore rest on very few independent time units and should be read as evidence of sign, not of magnitude. Entries are in units of the cell's median univariate loss per evaluation row, averaged over cells, not relative CRPS reductions. PatchTST shows similar gains over random but is confounded by its training window.

3.3 Post-Development Gate Evaluation

Development used a two-leg gate instead of policy value: a constant-value null on a prespecified rank-transition signature of realized gains and a held-out test that realized gain can be ranked. Appendix C reports it in full; it supports 5 of 9 pairs under its strict rule. It is not a test of routing value. Its first leg is neither necessary nor sufficient for it (Appendix C), and its pair readings differ from Table 1 on 3 pairs. The two target different quantities at different aggregation and power, and an interval that includes zero does not show that value is zero, so these differences are not false positives of

Table 1: Matched-budget net gain of opening context where the router predicts the largest gain, on Fold-D origins. With $\tilde{d} = d/L_c$ the context gain in units of the cell's median univariate loss and $g \in \{0, 1\}$ the decision, $V_{\text{rand}} = n^{-1} \sum (g - \bar{g})\tilde{d}$ for the top-20% of $\widehat{\Delta}$ compares with random selection at the same budget; D_{unc} and D_{cp} are $n^{-1} \sum (g - g')\tilde{d}$ against the top-20% by the univariate forecast's predictive-entropy proxy and by change-point probability, both known before the decision; *online* is V_{rand} for a threshold set at the 80th percentile of earlier origins' $\widehat{\Delta}$. Entries are equal-weight means over a pair's horizon cells, times 100, so $+1.00$ is one percent of median univariate loss, not a relative CRPS change. Brackets are paired 95% circular-block-bootstrap intervals with time blocks shared by all cells and series of a pair; they are per pair and unadjusted for multiplicity. Plans `policy-value-v1` and `policy-value-checks-v1` were registered before computation; the analysis is exploratory and post-development.

domain	backbone	V_{rand}	D_{unc}	D_{cp}	online
zero-shot backbones					
energy	chronos2	+0.48	+0.11	+0.55	+0.71
		[−0.15, +1.32]	[−1.29, +1.30]	[−0.40, +1.64]	[+0.15, +1.40]
	moirai	+1.89	+1.90	+2.66	+1.99
		[+0.94, +2.39]	[+0.11, +3.67]	[+0.57, +4.93]	[+0.96, +2.80]
retail	chronos2	+1.79	+3.42	+4.09	+2.20
		[+1.13, +2.32]	[+2.63, +4.20]	[+3.07, +5.24]	[+1.48, +2.79]
	moirai	+9.66	+5.55	+8.74	+9.51
		[+8.08, +11.14]	[+4.09, +6.91]	[+6.29, +11.16]	[+8.07, +10.91]
finance/news	chronos2	+0.16	−0.34	+0.29	+0.26
		[−0.58, +0.97]	[−0.73, +0.02]	[−0.63, +1.09]	[−0.29, +0.78]
	moirai	+0.84	−0.07	+0.07	+0.63
		[+0.40, +1.22]	[−0.99, +0.68]	[−0.64, +0.84]	[+0.37, +0.81]
PatchTST (trained per cell; training-window confounded)					
energy	patchtst	+2.32	−0.63	+2.27	+2.47
		[+0.45, +4.20]	[−5.08, +2.79]	[−0.41, +4.82]	[+0.66, +4.24]
retail	patchtst	+2.74	+0.60	+3.05	+2.92
		[+2.10, +3.32]	[+0.10, +1.11]	[+2.47, +3.66]	[+2.23, +3.49]
finance/news	patchtst	+0.64	+0.10	+1.57	+0.47
		[−0.08, +1.39]	[−0.74, +0.94]	[+0.42, +2.75]	[−0.15, +1.07]

either analysis. The selection effect behind Proposition 1 also appears inside our own diagnostics: backbones agree on top-decile realized-gain origins at 0.203 against chance 0.1, but the permuted-context arm agrees at 0.197, so the agreement reflects shared target difficulty, not context value.

3.4 Exploratory Horizon and Regime Analyses

Exploratory. Most cells lack a monotone horizon ordering, and event, change-point, and regime associations are small; Appendix G gives the classification.

3.5 Fusion Control Results

Fusion is evaluated after context is available on the same 9 pairs; because its information set differs, results are in Appendix H.

3.6 Forecast-Revision Explanation

For zero-shot families, deleting every covariate recovers the univariate arm exactly, a pathway identity; the trained family passes its arm-gap test on 13.8% of selected origins. Fidelity does not show that context caused the accuracy gain.

4 Related Work and Scope Delineation

Context-aided forecasting. Forecasting systems add text by language-model reprogramming (Jin et al., 2023), exogenous-variable fusion (Wang et al., 2024), or paired numeric–text benchmarks (Liu et al., 2024). Essential-context tasks exclude harmful context by design (Williams et al., 2024), whereas harm is part of our target. Audits find little semantic contribution under empty, constant, or shuffled text (Sridhar et al., 2026) and collapsing gains under spectrum-and-marginal-preserving surrogates (Cakiroglu et al., 2026); a routing pre-training diagnostic (Zhou et al., 2026) targets a different object. These works ask whether context helps on average or survives perturbation; we ask whether a policy acting before the outcome can locate where it helps, and evaluate that claim at a matched budget.

Variable importance is a different question. Attention identifies used inputs (Lim et al., 2021), not their incremental out-of-sample value; shortcuts can absorb a text signal (Geirhos et al., 2020). Our estimand concerns forecast skill. Related assumption-lean significance asks what excess predictability a covariate buys (Lundborg et al., 2024); conditional independence is hard in general (Shah & Peters, 2020), motivating nulls adapted to continuous conditioning variables (Candès et al., 2018; Berrett et al., 2019; Kim et al., 2022).

Selective evaluation and routing. Context acquisition is a reject-option problem (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), or learning to defer when the alternative costs more (Madras et al., 2017; Mozannar & Sontag, 2020). We use this machinery; the threshold router is conventional, and router architecture is not a novelty claim of this work. Perturbation-based behavioural testing is also established (Ribeiro et al., 2020); deletion and timestamp checks are diagnostics.

Positioning: ranking on noise. Classical forecast comparison is unconditional (Diebold & Mariano, 1995), and our proper-score estimand cannot reward miscalibration (Matheson & Winkler, 1976; Gneiting & Raftery, 2007; Gneiting et al., 2007; Ehm et al., 2016). Conditional predictive ability already tests whether a loss difference is predictable from origin information (Giacomini & White, 2006); Lemma 1 reads the same conditional loss difference through a budgeted decision, so its magnitude is a policy's gain rather than a test statistic, and baselines such as uncertainty-based selection enter as competing policies. Ranking and rereading the same noisy effect creates a winner's curse (Efron, 2011), requiring selection-aware intervals or an explicit null (Benjamini & Yekutieli, 2005; Berk et al., 2013; Fithian et al., 2014). Overlapping windows motivate temporal blocks (Kunsch, 1989; Politis & Romano, 1994), within-stratum randomization follows surrogate-data tests (Lotwick & Silverman, 1982; Theiler et al., 1992), and cell families require multiplicity control (Benjamini & Hochberg, 1995). Beyond these tools, we show that realized-gain concentration cannot identify routable value, and we measure routing value as matched-budget net gain against pre-decision baselines. Distribution-free risk control is complementary (Angelopoulos et al., 2025); regime labels enter only descriptive change-point summaries (Truong et al., 2020).

5 Limitations, Negative Results, and Sensitivity

5.1 Limitations

The router, folds, and evaluation rules were developed on the same archives, and no data untouched by development exist yet; the policy-value results are post-development and exploratory, with per-pair intervals and no multiplicity adjustment. Their plans were registered before computation, the robustness plan after the main result was known. Retail's Fold D spans 87 origin times in 3 blocks, so its intervals rest on few independent time units. Appendix J reserves the first forward block for one out-of-design evaluation once the public archive reaches it.

The development gate uses a versioned post-hoc amendment after failed v1 validation; residual seasonality, higher-order dependence, or persistent latent regimes can violate its exchangeability assumptions. Coverage differs by domain: 8 finance/news cells are unavailable and two series are absent, while retail includes known-future events and promotions. Under context latency, 3 of 27

rows miss retention threshold 60%. PatchTST remains secondary because disjoint training windows mix context and training-window effects.

5.2 Negative Results

Energy and finance/news with Chronos-2 show no matched-budget gain over random selection at the fixed budget, and in finance/news the router does not improve on uncertainty-based selection. The unit-weighted Phase-One summary does not support the initial claim that context is nonpositive at most origins, no finance/news backbone meets the development gate's joint support, horizon concentration is mostly nonmonotonic, and regime and event associations are small. PatchTST deletion passes on only 13.8% of selected origins, leaving 86.2% unfaithful.

5.3 Parameter Sensitivity

Exploratory. Appendix I distinguishes executed from unavailable grid cells and reports both population reruns; unavailable combinations remain empty rather than imputed. For the development gate's first leg, zero clipping changes 47 passes to 48 of 72, with 3 changes including 2 zero-shot cells; past-only retail changes 9 to 6, with all 3 losses in the secondary trained family, and the aggregate verdict holds.

6 Conclusion

A context-gain concentration curve is a statistic of realized gains, and routing depends on what is known before the outcome. Two settings can share the curve and differ in whether any policy can exploit it, so routing claims need a decision target: the gain of a pre-decision policy over matched-budget alternatives. Measured that way, opening context at the router's top 20% beats random selection in 4 of 6 zero-shot pairs across 3 domains and beats uncertainty-based selection in 3, in an exploratory, post-development analysis with per-pair intervals. Confirming these readings needs data untouched by development. Acquisition and fusion remain separate decision problems, and neither a conventional router architecture nor explanation fidelity substitutes for evidence of predictable context value.

AI use statement

In this work we used a large language model for the following tasks in the policy's required-disclosure category: implementing methods, in that it helped write the analysis and reporting scripts released with this paper, including the scripts that generate the synthetic validation data from the frozen specification; and providing feedback on research methodology and on the manuscript, in that drafts of the paper and of the decision rule's documentation were reviewed by a model acting as a critical reader, and its comments informed revisions. That feedback went beyond wording: the two-world construction behind Proposition 1, the counterexample showing that the gate's first leg is neither necessary nor sufficient for routable value, the choice to evaluate routing as matched-budget net gain with uncertainty and change-point baselines, and the reorganization of the paper around that result were first suggested in model feedback, and model feedback also informed how several results are interpreted. We checked each proof and every analysis, and the final interpretation is ours. We did not use generative AI to propose the original hypotheses, to choose the estimand, the decision rule, or the amendment; the remaining required-disclosure tasks (translation, dataset cleaning, and qualitative analysis) do not apply to this work. Additionally, we used a model for tasks in the recommended-disclosure category: creating and editing software code, suggesting experimental parameters, drafting and editing prose, including the written proofs, assembling LaTeX, and reviewing code. We have reviewed all AI-assisted work. Model-written code was verified against the frozen specification and is covered by the released test suite; every experimental result, figure, and number was produced by that code rather than typed by a model, and every scalar in the text is rendered from the number registry described in the reproducibility statement, so any value can be traced to the artifact and field it came from. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

We have read the ICLR Code of Ethics and this work conforms to it. The assessment behind that statement follows.

Human subjects. No humans took part in this research. There are no participants, annotators, or crowd workers, and no human judgements enter the pipeline; every label the analysis uses is computed from the series. We assessed whether ethical review was required and concluded that it was not, because the project collects nothing from people and studies no one.

Data provenance and use. We created no new data. All three domains are pre-existing public datasets, recorded as read-only in the source registry, so nothing we ran could alter an input. None of them holds records about identifiable individuals: the energy panel is grid-level power, the retail panel is per-product store sales, and the news panel is financial journalism about listed companies. We made no attempt to re-identify anyone, and the design would gain nothing from doing so. Every input file is recorded with its content digest. The retail data were fetched during this project from public HuggingFace mirrors of the competition release, named in the shipped configuration, without the competition platform's credentialed interface; the acquisition record reports that no platform API was called and lists the content digest of each file. The other two domains were already on disk as read-only archives. Each dataset and each pretrained forecaster is cited, and Appendix K lists where each was obtained and the terms under which it is distributed; the pretrained forecasters are pinned by snapshot identifier.

Risk of overreading. Our verdict is that concentration of context value is weaker and narrower than routing work has assumed, on these domains, under this decision rule, and at this resolution. It is not a finding that context never helps, and it would be a misuse of this paper to retire a deployed context channel on its strength. Appendix C quantifies what each failing cell excludes, and the appendices state where the instrument stops resolving. The same care applies to the work we qualify: the papers whose claims we scope are cited for what they demonstrate, and none of our own failures is presented as their error.

Compute. Compute is this project's main environmental cost, incurred in the forecast sweeps and the null calibrations. We reduced it where the design allowed: forecasts are cached against a content hash so a repeated configuration is read rather than recomputed, and the synthetic validation sweep checkpoints each grid point so an interrupted run resumes instead of restarting.

Research integrity and interests. The negative parts of this record are disclosed deliberately. A preregistered rule that failed validation, the version that replaced it, every amendment, and the decisions we would now make differently are reported with their dates and reasons, because a reader cannot judge a result whose discarded attempts are invisible. No funder or sponsor had any role in the design, the analysis, or the decision to report what we found, and we know of no competing interest bearing on this work.

Reproducibility statement

Every scalar is rendered from a number registry that binds it to a source, SHA256, and an exact field or deterministic selector; the build rejects missing sources and unresolved placeholders, while versioned preregistration files identify each result as prospectively frozen after method development and evaluated once on the designated analysis folds, exploratory, or descriptive. Numbered pipeline stages record seeds, package versions, configuration hashes, and commands in provenance sidecars and write deciding outputs' completion markers last; Fold D is ledger-guarded, and numerical runs pin the BLAS thread count to one. Appendix D documents the temporal design, Appendix A states the assumptions and complete proofs of Proposition 1 and Lemma 1, Appendix B the policy-value robustness checks, Appendix E the constant-value null and failed first version, and Appendix L the registry, provenance, and build checks. Both policy-value analysis plans are recorded verbatim, with their hashes, in the released decision log, and the anonymous supplementary package contains the code that produces every table. Every citation key was verified by a live title-matching query; en-

tries without a retrieved venue are cited under the verified arXiv preprint. This manuscript is a new submission to ICLR 2027; it is not a revision of an existing submission.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19, 2025. doi: 10.1214/24-aoas1998. arXiv:2110.01052.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

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.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005. doi: 10.1198/016214504000001907.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2), 2013. doi: 10.1214/12-aos1077.

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197, 2019. doi: 10.1111/rssb.12340.

Mert Onur Cakiroglu, Mehmet Dalkilic, and Hasan Kurban. The spectrum is not enough: When context helps time-series forecasting. *arXiv preprint arXiv:2607.13006*, 2026.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/tit.1970.1054406.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4918–4927, 2024. doi: 10.1145/3637528.3671629. arXiv:2402.06698.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: Consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):505–562, 2016. doi: 10.1111/rssb.12154.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605—-1641, 2010.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665—-673, 2020. doi: 10.1038/s42256-020-00257-z. arXiv:2004.07780.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74 (6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2): 243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6), 2022. doi: 10.1214/22-aos2233.

Hans R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3), 1989. doi: 10.1214/aos/1176347265.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37:1748—-1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012. arXiv:1912.09363.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-mmd: Multi-domain multimodal dataset for time series analysis. *arXiv preprint arXiv:2406.08627*, 2024.

H. W. Lotwick and B. W. Silverman. Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(3):406–413, 1982. doi: 10.1111/j.2517-6161.1982.tb01221.x.

Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6), 2024. doi: 10.1214/24-aos2447.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. doi: 10.1016/j.ijforecast.2021.11.013.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. doi: 10.1287/mnsc.22.10.1087.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. *arXiv preprint arXiv:2006.01862*, 2020.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 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, 2020. doi: 10.18653/v1/2020.acl-main.442.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), 2020. doi: 10.1214/19-aos1857.

Karthik Sridhar, Atharva Gupta, Nishant Pradhan, Murari Mandal, Dhruv Kumar, and Saurabh Deshpande. Semantics or structure? auditing text sensitivity in multimodal time-series forecasting. *arXiv preprint arXiv:2608.22321*, 2026.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58 (1-4):77–94, 1992. doi: 10.1016/0167-2789(92)90102-s.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. doi: 10.1016/j.sigpro.2019.107299.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

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

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

A Proofs

Throughout, $\mathbb{E}|d| < \infty$ and the policy g is a function of the decision-time information X only.

Lemma 1. Because $g(X) - q$ is a function of X, the tower property gives $\mathbb{E}[(g(X) - q)d] = \mathbb{E}[(g(X) - q)\mathbb{E}[d \mid X]] = \mathbb{E}[(g(X) - q)\Delta(X)]$, and since $\mathbb{E}[g(X) - q] = 0$ this equals $\mathrm{Cov}(g(X), \Delta(X))$. If $\Delta(X) = c$ almost surely, the value is $c\,\mathbb{E}[g(X) - q] = 0$ for every policy. Otherwise some τ has $0 < \Pr(\Delta(X) > \tau) < 1$. Take $g = \mathbf{1}\{\Delta(X) > \tau\}$, whose budget is $q = \Pr(\Delta(X) > \tau)$. Then

$$\mathrm{Cov}(g, \Delta) = q\,(1 - q)\,(\mathbb{E}[\Delta \mid g = 1] - \mathbb{E}[\Delta \mid g = 0]) > 0,$$

because the first conditional mean exceeds τ and the second does not.

Proposition 1. (i) X and Z have the same distribution and each is independent of ε, so (aZ, ε) and (aX, ε) have the same joint law and so does d. Because origins are drawn independently (an assumption of the proposition), the vector of realized gains has the same law in both worlds; $G(q)$ and $\mathrm{PGC}(q)$ are functions of that vector (ties broken by independent randomization), so their distributions coincide at every q. (ii) In world A, $\mathbb{E}[d \mid X] = a\,\mathbb{E}[Z \mid X] + \mathbb{E}[\varepsilon \mid X] = 0$; in world B, $\mathbb{E}[d \mid X] = aX$. (iii) The budget is $\Pr(X = 1) = \frac{1}{2}$. By Lemma 1 the value is $\mathbb{E}[(g - \frac{1}{2})\Delta(X)]$, which is 0 in A and $a\,\mathbb{E}[(\mathbf{1}\{X = 1\} - \frac{1}{2})X] = a/2$ in B. In world A, Lemma 1 also gives zero value for every policy that uses X, although its concentration curve is exactly that of world B.

The sample contrast V_{rand} of Section 2.3 replaces q by the realized budget $\bar{g}$. For the batch top-fraction rule, g depends on the fold's scores but never on realized gains, so the argument applies conditionally on the scores.

Table 2: Robustness of the top-20% V_{rand} in Table 1 (times 100). $T/\ell/K$: Fold-D origin times, block length, and number of time blocks. Intervals are recomputed with block length $\ell/4$ and $\ell/2$, which give more blocks; longer blocks leave too few (two in retail at 2ℓ) for an interval to be informative, and those results are in the artifact only. Even at the frozen length, with few blocks the bootstrap has few independent units, so these widths are indicative. *thirds*: point estimates on three contiguous, equal-count time thirds. *shared*: one decision per origin shared by all horizons (mean $\widehat{\Delta}$ over horizons selects, mean $\tilde{d}$ scores). *cells*> 0: share of horizon cells with a positive point estimate. Plan `policy-value-checks-v1`; exploratory.

pair	$T/\ell/K$	$\ell/4$	$\ell/2$	thirds	shared	cells> 0
energy–chronos2	1463/160/10	[−0.26, +1.30]	[−0.21, +1.32]	+0.28, +0.75, +0.29	[−0.80, +0.46]	0.67
energy–moirai	1463/160/10	[+1.36, +2.36]	[+1.19, +2.41]	+1.83, +0.87, +0.19	[+0.42, +1.85]	0.77
retail–chronos2	87/42/3	[+0.90, +2.65]	[+1.09, +2.41]	+2.12, +1.95, +1.15	[+0.58, +1.18]	1.00
retail–moirai	87/42/3	[+7.53, +11.79]	[+8.01, +11.19]	+11.35, +10.63, +7.11	[+4.73, +7.31]	1.00
fin./news–chronos2	275/74/4	[−0.47, +0.84]	[−0.50, +0.83]	−0.62, +1.19, −0.01	[−0.84, +0.59]	0.50
fin./news–moirai	275/74/4	[+0.17, +1.50]	[+0.24, +1.40]	+1.19, +0.61, +0.59	[+0.37, +1.15]	0.88
energy–patchtst	1463/160/10	[+0.81, +3.85]	[+0.69, +4.05]	+3.45, +3.34, −0.71	[−0.07, +2.65]	0.73
retail–patchtst	87/42/3	[+1.69, +3.73]	[+2.05, +3.60]	+3.24, +2.55, +2.52	[+1.23, +2.24]	1.00
fin./news–patchtst	275/74/4	[+0.01, +1.32]	[−0.04, +1.42]	−0.26, +1.51, +0.43	[−0.09, +0.57]	1.00

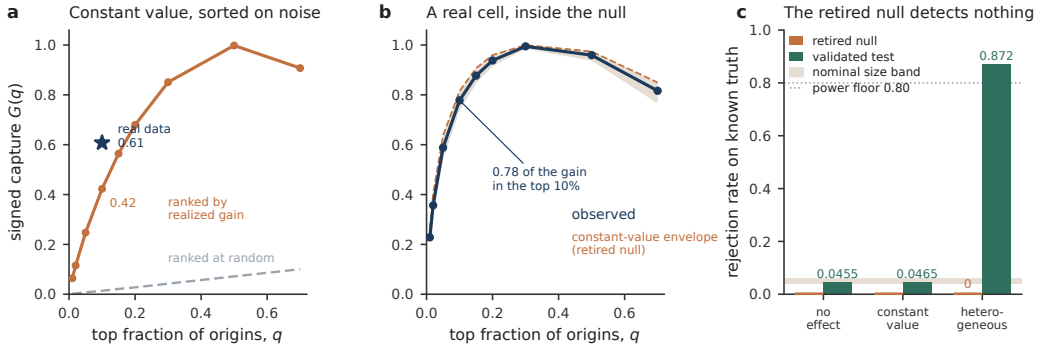

Figure 2: The development-stage null. **(a)** As in Figure 1. **(b)** The rule-selected real cell, like every other cell in its domain, stays inside the first null's constant-value envelope. **(c)** That null rejects neither homogeneous nor heterogeneous synthetic truth; the replacement controls size and clears the power floor over 2000 replicates per class.

B Robustness of the Policy Value

Table 2 reports the checks summarized in Section 3.2; Table 3 gives, for the primary rule and for opening context wherever $\widehat{\Delta} > 0$, the gain over random, over never using context, and over always using it. The bootstrap resamples blocks of consecutive origin times, and every series and cell of a pair uses the same resampled times, so dependence across series at a common time is retained. The number of independent units is therefore the number of time blocks, not the number of rows. Retail has the fewest, which is why its short-block intervals and its time-third estimates are shown next to the frozen-length interval.

C The Development-Stage Gate

Before policy value was adopted as the target, the protocol decided whether context value was heterogeneous with a two-leg rule, frozen after method development and evaluated once on the designated analysis folds. This appendix reports that rule and its results. The original v1 synthetic validation failed (Appendix E); the rule below is the post-hoc amended-v2 replacement.

The rank-transition statistic and its constant-value null. Let q_{lo} and q_{hi} be the preregistered fractions 10% and 20%. A global mean, observed-stratum scale, and clipped first-order autoregressive coefficient are fit on training folds; held-out one-step innovations are mid-ranked within each

Table 3: Policy value: the net gain of acting on the router prediction $\widehat{\Delta}$ on the Fold-D origins, in percent of the cell's median univariate loss L_c. With $\tilde{d} = d/L_c$ the standardized context gain of an origin and $g \in \{0, 1\}$ the policy's decision to use context, $V_{\text{rand}} = n^{-1} \sum (g - \bar{g})\tilde{d}$ compares the policy with random selection at the same budget $\bar{g}$, $V_{\text{never}} = n^{-1} \sum g\tilde{d}$ with never using context, and $V_{\text{always}} = -n^{-1} \sum (1 - g)\tilde{d}$ with always using context; positive values favour the policy. Each entry is the pair mean over cells, times 100; brackets give the 95% circular-block-bootstrap percentile interval, with blocks shared across the pair's cells and no multiplicity adjustment. *use* is the pair mean of the share of origins with $g = 1$; *strict gate* marks the pairs the strict conjunction supports. The plan `policy-value-v1` was registered before computation; the analysis is exploratory and post-development, and is not a gate.

domain	backbone	strict gate	use	V_{rand}	V_{never}	V_{always}
top-20% of $\widehat{\Delta}$ (primary)						
energy	chronos2	✓	0.200	+0.48 [−0.15, +1.32]	+0.40 [−0.59, +1.49]	+0.80 [−1.25, +2.54]
	moirai	✓	0.200	+1.89 [+0.94, +2.39]	+3.31 [+0.90, +5.90]	−3.76 [−6.27, −0.97]
	patchtst	✓	0.200	+2.32 [+0.45, +4.20]	+3.49 [+1.21, +5.94]	−2.37 [−7.34, +1.48]
retail	chronos2	–	0.200	+1.79 [+1.13, +2.32]	+1.25 [+0.57, +1.81]	+3.93 [+3.11, +4.72]
	moirai	✓	0.200	+9.66 [+8.08, +11.14]	+12.54 [+10.57, +14.53]	−1.86 [−3.36, −0.47]
	patchtst	✓	0.200	+2.74 [+2.10, +3.32]	+4.03 [+3.25, +4.76]	−2.41 [−2.99, −1.78]
finance/news	chronos2	–	0.200	+0.16 [−0.58, +0.97]	+0.25 [−0.61, +1.22]	−0.22 [−1.62, +1.16]
	moirai	–	0.200	+0.84 [+0.40, +1.22]	+1.20 [+0.69, +1.84]	−0.61 [−2.64, +1.28]
	patchtst	–	0.200	+0.64 [−0.08, +1.39]	+0.27 [−0.37, +0.85]	+2.11 [−0.19, +4.38]
$\widehat{\Delta} > 0$ *(secondary)*						
energy	chronos2	✓	0.241	+0.75 [+0.00, +1.61]	+0.38 [−0.89, +1.79]	+0.78 [−1.03, +2.21]
	moirai	✓	0.715	+0.77 [+0.16, +1.19]	+10.60 [+6.93, +14.31]	+3.53 [+2.46, +4.98]
	patchtst	✓	0.553	+3.12 [+1.46, +4.59]	+6.40 [+3.69, +8.92]	+0.53 [−2.69, +3.34]
retail	chronos2	–	0.727	+2.72 [+1.85, +3.60]	+0.93 [+0.44, +1.44]	+3.60 [+2.51, +4.66]
	moirai	✓	0.641	+6.50 [+5.55, +7.47]	+16.16 [+13.92, +18.64]	+1.75 [+1.19, +2.31]
	patchtst	✓	0.971	+0.15 [−0.06, +0.38]	+6.45 [+5.58, +7.28]	+0.00 [−0.26, +0.28]
finance/news	chronos2	–	0.997	+0.00 [−0.03, +0.03]	+0.46 [−1.31, +2.37]	−0.01 [−0.03, +0.02]
	moirai	–	0.980	−0.08 [−0.24, +0.05]	+1.69 [−0.55, +4.11]	−0.12 [−0.28, −0.00]
	patchtst	–	0.000	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	+1.84 [−0.69, +4.01]

series and fold, $U_t = (\text{rank}(\varepsilon_t) - 1/2)/m$. For deterministic non-overlapping adjacent pairs and each q, $E_q = \{U_{t-1} \geq 1 - q \text{ and } U_t \leq q\}$. One Pearson X^2 uses the mutually exclusive counts $(E_{q_{\text{lo}}}, E_{q_{\text{hi}}} \setminus E_{q_{\text{lo}}}, \text{neither})$. The randomization holds previous endpoints and all marginal ranks fixed and permutes current endpoints within series and fold $B = 2000$ times; the add-one p-value rejects at $p \leq 0.05$. Mid-ranks make the statistic invariant to continuous monotone transformations. The permutation unit is the current endpoint of a non-overlapping pair within one series and fold; it keeps each series' marginal ranks and previous endpoints, but it does not keep the dependence between pairs that overlapping forecast windows induce. Its validity therefore rests on exchangeability of

Table 4: Both readings of the gate, per domain–backbone pair. *cells* counts evaluated origin–horizon cells and *computable* those on which the predictability leg can be evaluated at all. The two legs are counted independently; *both* requires one cell to clear them at once, and is the count quoted in the text. *pair* is the coarser roll-up that pools a pair's cells instead of demanding a single cell carry both. The two readings disagree on exactly one row.

domain	backbone	cells	computable	constant-value null	predictability	both	pair
energy	chronos2	96	72	65	14	12	✓
	moirai	96	72	40	29	8	✓
	patchtst	96	72	26	40	17	✓
retail	chronos2	8	4	2	2	0	✓*
	moirai	8	4	4	4	4	✓
	patchtst	8	4	3	4	3	✓
finance/news	chronos2	16	16	0	0	0	–
	moirai	16	16	0	2	0	–
	patchtst	16	8	0	0	0	–

* the single row where the pooled reading supports a pair that no one cell carries: two cells clear the constant-value null and two clear predictability, never the same cell.

the current endpoints within a series and fold, conditional on the training-fitted nuisance, and residual dependence beyond the fitted autoregression violates it. Leg 1 is Benjamini–Hochberg corrected within each domain–backbone–horizon family. The second leg passes when the paired 95% circular-block-bootstrap interval for held-out rank association and excess positive-gain capture of $\widehat{\Delta}$ on Fold D clears the null; by preregistered choice it is uncorrected. A cell supports the rule only when both legs pass, and support must cover two domains and two backbones, including energy.

What the first leg does not test. The prespecified signal is excess adjacent transitions from the top tenth to the bottom tenth and from the top fifth to the bottom fifth; because Pearson X^2 is unsigned, deficits can also raise it. Between-series or between-fold levels, persistent high-rank runs, heterogeneity that preserves those counts, and cross-horizon patterns lie outside its scope. It is neither necessary nor sufficient for routable value. If $d_t = aX_t + \varepsilon_t$ with X_t observed before the decision and independent over time, then $\Delta(X_t) = aX_t$ can be routed, yet the filtered innovations carry no transition structure, and the leg rejects at about its nominal rate. Against heterogeneity independent over time its power comes only from serial dependence in the noise, which is the form of the synthetic power alternative; conversely, unmodelled regimes or higher-order dependence can produce transitions with no routable value.

Results. The rule supports 5 of 9 domain–backbone pairs under the strict cell-conjunction rule, in energy and retail. At cell level, the first leg clears 140 of 360 cells across 6 pairs, the second clears 95 of 268 computable cells, and both clear 44. Of the passing cells, 20 use confounded PatchTST; without them, 24 cells span 3 pairs, 2 domains, and 2 backbones. Applying Benjamini–Hochberg to the second leg within the same 42 families changes 95 passing cells to 94; the conjunction retains 43 cells and the same 5 pairs. A global-null simulation with the family and coverage structure returns support on at most 0.0017 of replicates; it omits cross-family dependence and draws the legs independently. The pooled roll-up re-estimates each leg per pair and supports 6 of 9; there all 3 unsupported pairs are finance/news, while under the strict rule retail with Chronos-2 is also unsupported because different cells clear different legs. No finance/news cell passes both legs: of its 40 computable cells, 38 fail the second leg and the other 2 fail the first.

What the failing cells rule out. Exploratory. Among 268 computable cells, 173 fail the second leg. At resolution 0.1, the round number nearest the weakest detected effect, 67 of 173 exclude an effect that large on both quantities (Appendix F). No finance/news predictability failure admits rank association above 0.165 or excess capture above 0.173, against an energy rank ceiling of 0.428.

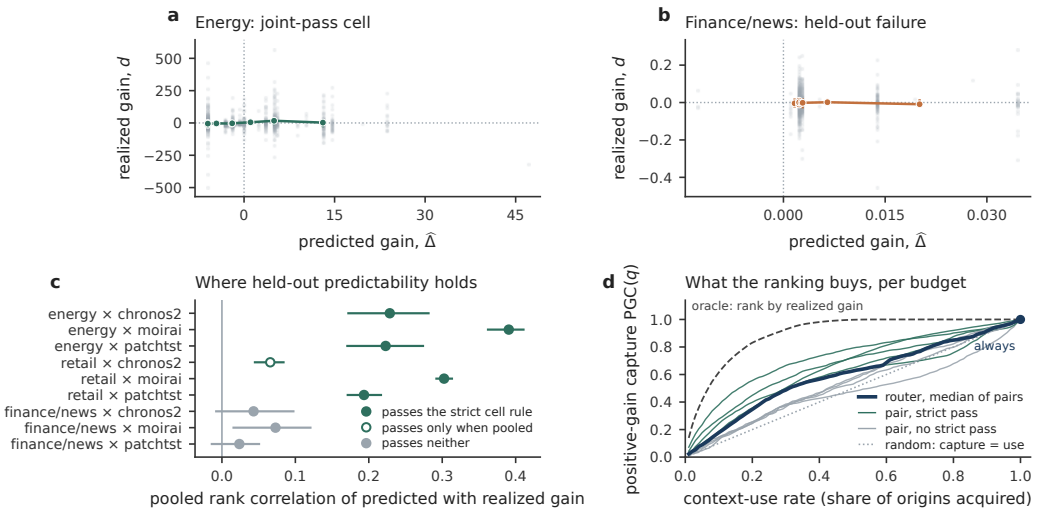

Figure 3: Evidence behind the gate. Panels (a) and (b) show Fold-D origins from the Fold-C-trained $R_{\min}$ predictor and equal-count decile bins after a stable sort, with linked means and within-bin interquartile range. Panel (a) selects the shortest-horizon, lexicographically first target among joint-pass all-origin Fidel cells for the first zero-shot backbone; panel (b) applies that rule to FNSPID cells whose held-out predictability leg is complete and fails. Panel (c) shows pooled rank association and its interval by pair, distinguishing strict, pooled-only, and unsupported pairs. Panel (d) compares positive-gain capture curves and their median with random, always-context, and realized-gain-oracle references.

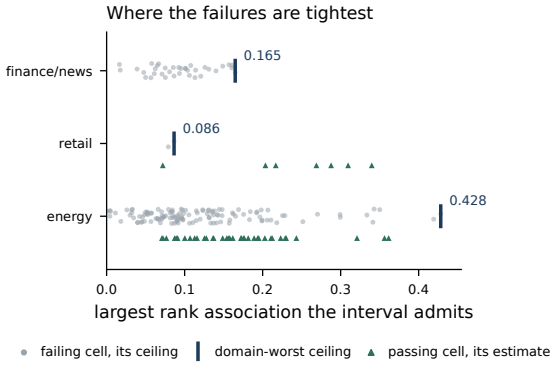

Figure 4: Resolution ceilings for predictability-leg failures. Grey points show the largest rank association each interval permits, domain rules show worst ceilings, and triangles show passing estimates.

D Temporal Design, Folds, and Embargoes

The analysis unit is a quadruple of series, origin timestamp, horizon, and target. In deployment, buying context is one decision per origin shared by every horizon. The main policy-value column evaluates each horizon cell as an independent decision; the shared-horizon column of Table 2 evaluates one purchase per origin.

Four windows run in time order, and each is used for one thing. Forecaster training comes first and is needed only by the supervised backbone. Paired losses are then constructed on a later window. Router labels come from Fold C. Fold D is evaluation only. Consecutive windows are separated by an embargo of the context lag plus the longest horizon, so no forecast window can straddle a boundary; the same quantity is the floor on admissible block length, which is why the shortest candidate block already covers the overlap between neighbouring forecast windows. Table 5 lists the resulting sizes.

Table 5: Fold contract, per domain, at the largest cell in each. Fold C carries router labels and Fold D is evaluation-only. Rows are paired origin–horizon observations; n_{eff} corrects rows for within-series autocorrelation and ignores dependence across series at a common time; *blocks* counts time blocks of the base length, which the bootstrap intervals resample jointly across series (Appendix D). Candidate block lengths ℓ follow the frozen rule $\ell = m \cdot (L + H_{\max})$ over three multipliers, so the shortest admissible block already spans the overlap between forecast windows; the same quantity sets the embargo between folds. Finance/news has fewer rows than energy yet a larger n_{eff}, because its series decorrelate faster. Cells carry one target series in energy, up to four in finance/news, and thirty in retail.

	Fold C (router labels)			Fold D (evaluation only)			
domain	rows	n_{eff}	blocks	rows	n_{eff}	blocks	ℓ
energy	2696	445	17	1463	380	10	160, 320, 640
retail	9270	9269	8	2610	2610	3	42, 84, 168
finance/news	1645	693	8	825	715	4	74, 148, 296

Two features of the table matter downstream. The n_{eff} column divides row count by one plus twice the pooled within-series autocorrelation of d up to the base block length; it ignores dependence between series at the same time, so in retail, whose thirty series show no positive within-series autocorrelation, it equals the row count. It enters only the detectable-effect calculations. The *blocks* column counts time blocks of the base length, and the bootstrap intervals resample those blocks jointly across series, so their number, not n_{eff}, bounds the independent units behind an interval. And because the supervised backbone needs its own training window, its cells carry a context effect and a training-window effect together. We mark every such cell training-window confounded and keep that family secondary, so the headline claims rest on the zero-shot families.

E The Constant-Value Null and Its Failed First Version

The null has to permit the nuisance structure and forbid only the thing being tested. Our target is $\Delta \equiv \mu$: one constant conditional value, with whatever noise the data actually have. Two design choices follow.

Stratification is the first. An unstratified block bootstrap implicitly assumes noise variance is homogeneous across origins, which real data violate strongly: solar at night against solar at noon, or a financial series in earnings season against a quiet week. Under a constant Δ with heteroskedastic noise, ranking by realized gain preferentially surfaces the high-variance stratum, so an unstratified null rejects *because of variance* and the rejection then reads as evidence of heterogeneous value. Resampling residuals inside strata, with the mean held global, permits variance heterogeneity and forbids value heterogeneity. That makes the stratified version primary and the unstratified one a secondary check, and it makes a positive result the stronger claim: concentration beyond what variance heterogeneity alone explains.

The second choice is that each null replicate must re-rank by its own resampled gains and recompute its own denominator. Reusing the observed ordering would drop the step under test, ranking on noise, and produce an envelope that is systematically too narrow, which is the same error as holding the denominator fixed. Blocks span the overlap between forecast windows so that dependence survives resampling, and strata are formed from the context regime channel, horizon, and series.

Why the first version was retired. The rule frozen in v1 added block-bootstrapped centred gains back onto the global mean within strata, and declared an escape when the observed curve cleared the pointwise upper envelope at both preregistered selection fractions. On synthetic data with known truth it rejected nothing: not under a zero effect, not under a constant effect, and not under the heterogeneous alternative it existed to detect. A test that never rejects is degenerate at this cell size, so its non-rejection on real data is uninformative. Table 6 records this alongside everything that followed, and Figure 2 shows what it means for reading a curve.

The diagnosis was pursued four ways before the rule was retired. Block length was raised to twice and then four times its floor. The unstratified secondary variant was checked, confirming the stratified

Table 6: Every rung of the validation ladder, in the order it was run. Size is measured on two null classes (no effect, and a constant nonzero value) and power on a heterogeneous alternative, with a stronger variant where one was declared. The frozen v1 rule rejects nothing anywhere, including under the alternative it exists to detect, so it is degenerate and not merely conservative. Each pilot rung was frozen before its results were read and each was rejected against its own pre-declared acceptance band, for the reason recorded in the decision log; only the last was eligible to freeze. The v2 row is the single full run made after freezing. Every candidate after v1 shares a cross-fitted autoregressive stage, whose common name prefix is omitted. Stages read in run order: v1 as frozen, the first round of candidates, three post-pilot amendments, then v2 as frozen.

			size		power	
stage	statistic	replicates	no effect	constant	heterogeneous	strong
v1	residual constant-Δ	2000	0	0	0	–
pilot 1	pit_decile_g	200	0.015	0.03	0.21	0.505
pilot 1	pit_joint_5bin	200	0.025	0.03	0.255	0.69
pilot 1	pit_legendre4	200	0.035	0.03	0.305	1
amend. 1	pit_qtransition	200	0.005	0.02	0.8	1
amend. 2	pit_qtransition_conditional	200	0	0	0.77	1
amend. 3	rank_qtransition	200	0.035	0.04	0.845	1
v2	rank_qtransition	2000	0.0455	0.0465	0.872	–

"–" marks a class the rung did not declare. Size is acceptable inside the pre-declared Monte Carlo band and power against the preregistered floor; both bands are recorded in the artifacts and neither was widened after any run.

choice was the original primary. An integrated statistic over the lower half of the selection range replaced the pointwise one, and an oracle supplied with the true noise was run as a positive control. None of them moved the rejection count off zero, which established that the failure lay in the rule itself.

What replaced it, and under what discipline. The successor conditions on an autoregressive fit, transforms held-out innovations to mid-ranks within series and fold, pairs adjacent non-overlapping ranks, and tests whether a high previous rank is followed by a low current one at the two preregistered selection fractions, combining both into a single statistic against a single critical distribution. Calibration is by conditional randomization of the current endpoints inside series and fold, holding previous endpoints and all marginal ranks fixed. Because ranks are invariant to any continuous monotone transformation, one randomization covers the Gaussian, Student-t, and right-skewed innovation families at once, which removes the need to combine family-specific p-values. The statistic reads only realized gains and index columns; synthetic truth, the true effect, the regime label, and the innovation family are all withheld from it by construction.

Each rung in Table 6 was frozen before its results were read, and each was rejected against its own pre-declared band for a recorded reason: the first three candidates had too little power against the weak alternative; the next held power but its theoretical reference was too conservative for a fitted transform; the one after that squeezed its own size to zero by combining family-specific p-values redundantly. Only the last cleared both bands and was eligible to freeze, after which exactly one full run was made. Every rung before that freeze is exploratory.

Two limits should be stated explicitly. The synthetic grid's largest sample size is below the largest real Fold-C cell, so calibration verified on the grid does not extrapolate to that cell by measurement, only by assumption. And the amended rule is not an omnibus test: it targets the transition structure it was built for, and unobserved variance regimes, higher-order dependence, or a structural break can still invalidate it.

F Resolution Ceilings for the Failing Cells

Appendix C reports how much each failing cell rules out. This appendix gives the whole curve behind that reading.

Table 7: Absence classification as a function of the resolution δ^*, over the cells that fail the predictability leg. A cell reads as evidence of absence at a given δ^* when both of its confidence ceilings fall below it, so the data exclude an effect that large on either quantity; otherwise the cell is simply uninformative at that resolution. We report the whole curve because choosing one δ^* after seeing which cells failed would be the same move the paper's calibration argument rejects. The arrow marks the value quoted in the text, chosen as the round number nearest the weakest effect the gate did detect. The whole table is exploratory.

δ^*	evidence of absence	absence of evidence	share
0.05	13	160	8%
0.10 ←	67	106	39%
0.15	110	63	64%
0.20	125	48	72%
0.25	141	32	82%
0.30	150	23	87%
0.40	164	9	95%
0.50	171	2	99%

Cells that clear predictability and fail calibration are excluded throughout: their predictability intervals exclude zero, so pricing them as predictability failures would invert the finding. Cells whose leg is not computable are excluded as well, and never imputed.

For every cell that fails the predictability leg we record two ceilings, the upper confidence limit on rank association and the upper limit on excess capture, both from that cell's own paired block bootstrap. A cell reads as evidence of absence at a resolution δ^* when both ceilings fall below it. Because that verdict depends on δ^*, and because picking one value after seeing which cells failed would be an effect-size cut chosen from the results, Table 7 reports the classification across the declared grid and leaves the choice to the reader. The share rises smoothly with δ^*, as a well-behaved set of interval estimates should: nearly every failing cell excludes a large effect, and only a minority exclude an effect as small as the weakest one the gate detected.

Two exclusions bear repeating, since both change the denominator. Cells that clear predictability and fall at the constant-value null are not priced here at all: their predictability intervals exclude zero, so treating them as predictability failures would report the reverse of what was measured, and pricing their constant-value null leg would need that leg's own reference distribution, which belongs to a separate analysis. Cells whose leg is computable are the ones listed. The per-cell values are in the released artifact rather than in print, because the table would run to several hundred rows and its shape is what matters.

G Horizon Ordering

Exploratory. A natural guess is that context should matter more at longer horizons, where the target history is less informative. Our cells do not show this ordering. Figure 5 classifies each base cell's concentration curve by how it moves with horizon: the large majority order neither way, and the signed rank correlation straddles zero with a slightly negative median. We report the classification as measured.

H Acquisition Beside Its Fusion Control

Acquisition and fusion answer different questions over different information sets, so Table 8 lists them side by side. The ordering between the two settings changes with the metric and with the domain, which is the reason we make no pooled claim that fusion is the better of the two.

For acquisition alone, Table 9 summarizes six policies as medians over domain–backbone pairs.

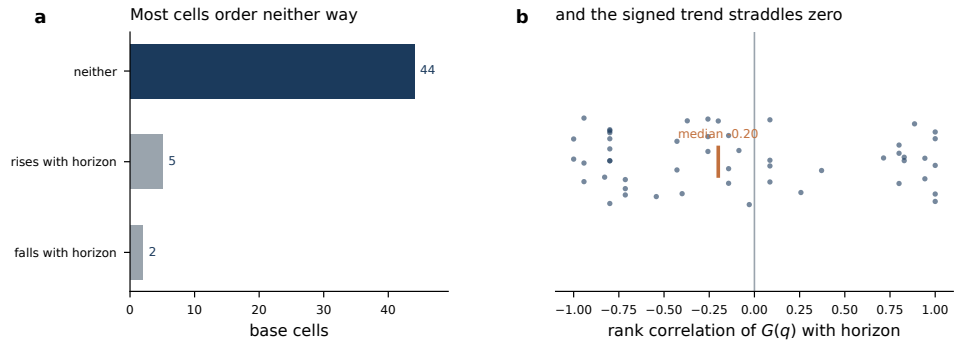

Figure 5: Concentration against forecast horizon, per base cell. **(a)** Trend classification: a cell counts as ordered only when its curve moves monotonically across the horizons evaluated for it. **(b)** The signed version, one point per base cell, with the median marked. Neither panel supports treating horizon as a route to where context helps. Both are exploratory.

Table 8: Acquisition beside its fusion control, per domain–backbone pair. The two answer different questions — whether to obtain context, and whether to use context already in hand — over different information sets, so their rows are reported side by side and never pooled. Which setting looks better changes with the metric and the domain, which is why we make no pooled claim of fusion superiority. Acquisition is the primary setting throughout.

		acquisition		fusion control	
domain	backbone	rank corr.	PGC@.20	rank corr.	PGC@.20
energy	chronos2	0.229	0.695	0.231	0.677
	moirai	0.391	0.741	0.428	0.787
	patchtst	0.223	0.621	0.236	0.614
retail	chronos2	0.066	0.279	0.067	0.264
	moirai	0.303	0.606	0.305	0.607
	patchtst	0.193	0.466	0.189	0.453
finance/news	chronos2	0.043	0.419	0.055	0.420
	moirai	0.073	0.456	0.001	0.424
	patchtst	0.024	0.211	0.082	0.175

I Parameter Sensitivity Grid

Exploratory. The grid crosses domain, block length, selection fraction, cost, embedding width, null variant, and population. Table 10 gives its status breakdown; most of the grid records what was not run.

That is deliberate, and it is also a real limit on how far the word "sensitivity" should be taken here. Just under half the combinations vary an embedding width that was never computed, and roughly a quarter name a population that was never materialized. Neither is imputed from a neighbour, so those axes carry no comparison at all; writing one in would manufacture a robustness check from nothing. What the grid does support is comparison across the block lengths and selection fractions that were actually recomputed.

One population is now materialized. The frozen scope decision `S-1` declared `zero_clipped` a secondary population for the solar targets, on the reasoning that night-time zero inflation and a few source-side negative rows might be inflating how steep the concentration curve looks. Because clipping is a score transform on stored quantiles and not a different forecast arm, it can be evaluated from committed forecasts: clip both arms and the realized values through the one shared function, rescore with the archive's estimator, and re-pair. Across 36 backbone–fold–horizon cells the median change in capture at the preregistered selection fraction is 0.0001, with a range from -0.082 to 0.027. The two zero-shot families are the stable part; the largest single change belongs to the supervised family, which is secondary and training-window confounded, so the conclusion is about the zero-

Table 9: Acquisition policies at zero context cost, median over domain–backbone pairs. All-origins cells are pooled within a pair before taking the median. Use is read from the realized-oracle, zero-cost strategy rows; PGC and harm are recomputed on the same Fold-D snapshots. The median includes PatchTST†.

	named policy			leg-2 top-20%	
strategy	use	PGC	harm	PGC	harm
Never	0.000	0.000	–	–	–
Always	1.000	1.000	0.424	–	–
Random-q	0.200	0.197	0.420	–	–
Uncertainty	0.200	0.343	0.393	–	–
Proposed ($\widehat{\Delta} > 0$)	0.715	0.844	0.433	0.347	0.416
Oracle	0.522	1.000	0.000	–	–

Never, Always, Random-q, Uncertainty, Proposed, and Oracle are `always_univariate`, `always_context`, `random_q`, `uncertainty_routing`, `learned_routing`, and `oracle_routing`. The first PGC and harm columns evaluate each named policy. For Proposed that deployment policy is $g = \mathbf{1}\{\widehat{\Delta} > 0\}$; the last two columns instead use leg-2's per-cell top-20% by $\widehat{\Delta}$ gate, with constant-score cells excluded exactly as in the committed harm artifact. †The artifact marks 3 PatchTST pairs as training-window confounded; they remain secondary and are included in the all-pair median. At realized $c = 0$, the optimism ledger records `room` and `clear_gain=true` for 276 of 276 complete cells. By its stored definition that verdict compares the non-deployable realized oracle only with Always, has no interval, and is not a gain test for Proposed.

shot evidence and not about all 36 cells uniformly. The stability also holds at the decision level, because we re-ran the amended constant-value null leg on the clipped population through the same per-cell code path the headline gate uses. On the 72 solar cells it can evaluate, the leg passes 47 under the base population and 48 under clipping, reaching the same decision on 69 of them. So clipping leaves the descriptive curve where it was and moves the leg's decision only at the margin: 3 cells change, and 2 of those sit in the zero-shot families, one gained and one lost.

The held-out predictability leg reads Fold D, whose two-checkout budget was already spent, so running it here required a versioned amendment (`prereg_v4_fold_d_secondary_amendment`) declaring a new purpose with its own separate budget and ledger; the primary budget and its audit stay byte-identical. We treat the result as exploratory; it is useful because a stable descriptive curve does not imply a stable decision. Clipping is a score transform, so the router inputs and thus the committed $\hat{\Delta}$ do not change; only the realized gain is rescored, and each cell is re-evaluated through the same `paired_leg2_bootstrap` on the same frozen index, its base value reproduced to the bit before the clipped one is read. On the 72 cells the leg passes 45 under the base population and 42 under clipping, the same decision on 69. All 3 that move are lost rather than gained, and all 3 belong to the supervised family; the two zero-shot families reach the same decision on every cell. So both legs, descriptive and held-out, agree: clipping moves only the training-window-confounded arm we had already set aside.

The other preregistered population is also materialized now. Decision `M-1` treats retail promotions and events as known in advance, which is what makes paying to retrieve context a real decision, and it preregisters the opposite arm as a bound: route the same covariates through the past channel so the forecaster sees them only once they have happened. This arm does need a forward pass, because the model receives a different input, but not retraining (same panel, same weights, one channel switched), and it writes a separate arm so the committed ones stay untouched.

Two quantities separate in the result, across 24 backbone–fold–horizon cells. Context is worth less without advance availability: mean gain falls in 18 of them, which is the bound `M-1` asked for. Concentration, though, barely moves, with a median change in capture at the preregistered selection fraction of 0.008. We also re-ran the amended constant-value null leg here, and it is more informative than the curve is. On the 12 cells the leg can evaluate, it passes 9 under the base population and 6 under past-only. All 3 cells that change are lost rather than gained, and all 3 of them belong to the supervised family. The two zero-shot families, which carry the primary evidence, reach the same decision on every cell.

Table 10: What the sensitivity grid contains, by status, over every combination of domain, block length, selection fraction, cost, embedding width, null variant, and population. Most of the grid is a record of what was *not* run, which is the point of publishing it: an unavailable combination is listed with its reason and is never imputed from a neighbouring one. Conclusions may therefore be compared across the block lengths and selection fractions that were computed, and not across embedding widths or populations, where there is nothing to compare.

status	combinations	share
committed statistics, re-read on this axis	550	5%
derived block analysis, re-read on this axis	1250	12%
embedding width never run	4800	47%
population never materialized	2400	24%
statistic unavailable, not fabricated	1200	12%

Of 10200 combinations, 4200 correspond to a recomputed statistic and 0 to a re-run forecasting model. No row varies a model fit: the grid varies the analysis applied to committed forecasts, so it bounds the sensitivity of our conclusions to analysis choices and says nothing about their sensitivity to retraining.

The descriptive curve and the constant-value null decision therefore separate in a specific way. Clipping, a monotone transform of scores, barely touches either: the curve holds and the pass count moves by one, though 3 cells change identity. Rerouting the covariate channel changes what the model sees, and while the curve still barely moves, the decision does, entirely inside the arm whose training window already confounds it. Only the second perturbation concentrates in one family. Clipping's few changes scatter, in both directions and across families, which is the signature of cells sitting near their own threshold. Read narrowly and on this M5 implementation, availability affects how much context is worth and leaves the primary zero-shot leg decisions unchanged.

The residual change should be located rather than averaged away. It lies almost entirely in the supervised family, where all 8 of 8 cells grow steeper, up to 0.073; across the two zero-shot families, which carry the primary evidence, the change never exceeds 0.039 in absolute value. Since the supervised family already mixes a context effect with a training-window effect, the concentration curve on the evidence the paper leans on is the part that does not respond. Taken with the clipping result above, two physically unrelated perturbations of the population leave the curve where it was, which is what the ranking account in Figure 1 predicts.

One boundary still holds for the rest of the grid. No other row re-runs a forecasting model: every entry re-analyses committed forecasts. So the grid bounds how much our conclusions depend on analysis choices, and it leaves retraining out of scope. Population choice is the axis where this limit matters most, because daylight-only solar origins, context-available finance/news origins, and past-only retail covariates are three genuinely different populations. Where a conclusion moves across the materialized choices we report it as sensitivity.

J A Pre-Frozen Held-Out Block

Every result above is measured on folds that were on the table while the amended rule was being written. The amendment chain records this, but recording it is not the same as testing it. What the record does not yet contain is one execution of the frozen rule on origins that could not have shaped it.

One slice of data qualifies by construction. The fold table caps the energy domain at the Fold-D boundary, so no committed forecast, score, paired loss, or cell-index row touches a single origin beyond it, and the block is the first window forward of that boundary. Carrying the design's own inter-fold convention forward from the boundary leaves 196 steps of separation, against the 160 steps the frozen embargo rule asks for, and then 580 steps of block: 548 scored origins in each of 8 series, 4384 together, over the 216 energy cells the committed index already fixes. The block gets its own identifier, its own artifact namespace, and its own ledger, so the fold vocabulary and every hash-bound artifact stay where they were.

The scope is set by arithmetic, not by preference. Finance/news leaves too little forward of its boundary to clear the embargo, short by 10 trading days, and the acquired retail file ends *at* the boundary,

short by 42 days. Since composition requires two domains, this block cannot re-derive the composition verdict; the most it can do is replicate inside energy. That limit is written into the amendment before the read, so that the scope is declared in advance rather than explained afterwards.

Power is declared in the same way and at the same time. The block carries fewer origins per series than Fold D, so at equal per-origin variance the committed detectable effect widens by a factor of 1.634, which is enough to make a non-replication ambiguous between a rule that does not travel and a block too short to tell. The per-cell value is left to the committed function on the block's own residuals; fixing one here would predetermine an outcome.

Access runs through a ledger with a budget of 1 read. Preconditions are verified before that ledger is charged, so an absent asset refuses and spends nothing, and the refusal is recorded without being counted. Four outcomes are enumerated ahead of the run (replicated, not replicated, a powered subset, inconclusive), and a fifth clause commits to reporting whichever one lands, including the case where the read is granted and the run then fails, which spends the block. At submission the budget is unspent and the block has not been read: the public energy archive currently ends at the Fold-D boundary, so the window holds no observation yet. The amendment, execution overlay, access rule, and outcome grid are frozen, and a follow-up amendment records the archive's extent. We report the outcome when the archive reaches the block and the read is taken.

K Data and Model Sources

This appendix lists every dataset and pretrained model the study reads, where each was obtained, and the terms under which it is distributed. No dataset is redistributed with this paper: the released code re-derives every panel from the sources named here, and the supplementary package carries digests of the files that were read rather than the files themselves. All use is non-commercial research. Two entries carry no explicit licence from their distributor. The energy panel is the Germany Renewable Energy Grid subset of the Fidel-TS benchmark (Xu et al., 2025), whose HuggingFace release attaches no licence field; its generation series are the German transmission system operators' statutory publications on Netztransparenz, and the benchmark authors state that the data were sourced under publicly accessible terms. The retail panel is the M5 competition data (Makridakis et al., 2022), obtained from a public HuggingFace mirror rather than from the competition platform; the competition rules under which Walmart released the data permit non-commercial use, including academic research. Caption embeddings for the energy panel use `sentence-transformers/all-MiniLM-L6-v2`, whose model card declares Apache-2.0; the configuration records the model identifier but not the exact revision used, so that revision remains unknown. The PatchTST family is our own implementation trained from scratch inside the released code, so no external model licence applies to it.

Fidel-TS, Germany Renewable Energy Grid (Xu et al., 2025). Energy panel (wind and solar generation for four transmission system operators) and the weather captions. Obtained from `huggingface.co/datasets/fidel-ts`. No licence field on the release; the generation series are public statutory publications of the German transmission system operators.

FNSPID (Dong et al., 2024). Finance/news panel. Obtained from `huggingface.co/datasets/Zihan1004/FNSPID`. Licence: CC BY-NC 4.0.

M5 (Makridakis et al., 2022). Retail panel. Obtained from the public HuggingFace mirror `kashif/M5` of the competition files. Terms: the competition rules, which permit non-commercial use including academic research.

Chronos-2 (Ansari et al., 2024). Zero-shot backbone. `amazon/chronos-2`, snapshot `29ec3766`. Licence: Apache-2.0.

Moirai-1.0-R-small (Woo et al., 2024). Zero-shot backbone. `Salesforce/moirai-1.0-R-small`, snapshot `f5bd5d01`. Licence: CC BY-NC 4.0.

all-MiniLM-L6-v2. Caption-embedding model for the energy panel. `sentence-transformers/all-MiniLM-L6-v2`. Licence: Apache-2.0. The exact revision used is not recorded in the project configuration and remains unknown.

PatchTST (Nie et al., 2022). Supervised backbone, trained from scratch in the released code; no external model licence applies.

Table 11: Reported quantities, median over cells within each domain. These are descriptive summaries of the selective-evaluation protocol and carry no gate decision: the gate reads the two legs in Table 4, never these medians. Cost-adjusted score and regret are divided by the cell's median univariate loss, because in raw score units they differ by three orders of magnitude between these domains and cannot be compared down a column. The last column is the harm rate of the router's own selection, which replaces a false-activation rate that is zero by construction (see the note). Cells where a quantity is undefined are dropped from that quantity's median and never imputed.

domain	cells	PGC@.10	cost-adj. score	regret vs oracle	calib. deviation	harm@.20
energy	216	0.603	1.44	0.0842	-0.000492	0.427
retail	12	0.646	1.59	0.0708	-0.00763	0.325
finance/news	48	0.624	1.4	0.0586	0.0312	0.477

Context use is pinned at the selection fraction on every cell (range 0.1–0.1), and the policy's false-activation rate (`policy_false_activation_by_construction`) is identically 0 on all 276 cells because that policy keeps the hindsight top 10% of origins by realized d and so never activates where $d < 0$. Both are fixed by the policy's construction, so they are stated here and not tabulated as measurements. *PGC*@.10 is the positive gain in the hindsight top 10% of origins by realized d, over total positive gain. *harm*@.20 is `harm_rate_selected`: the share of origins with $d < 0$ among those the router's own top-20% selection by $\hat{\Delta}$ activates on Fold D ($R_{\min}$, fit on Fold C), the selection the predictability leg scores. The 8 cells whose $\hat{\Delta}$ is constant leave that selection undefined and are dropped from its median.

Snapshot identifiers for the two zero-shot forecasting backbones are the leading characters of the HuggingFace commits pinned in the released configuration.

L Reported Quantities

Table 11 summarizes the quantities the selective-evaluation protocol reports, as a median over the cells of each domain. None of them enters the gate, which reads only the two legs.

Context use is pinned at the selection fraction on every cell. The quantity `policy_false_activation_by_construction` is zero because the hindsight policy keeps the top tenth by realized d and therefore never activates where $d < 0$. The measured replacement asks how often the router's own selection harms: at the pair median, $d < 0$ in 0.416 of selected cases for $R_{\min}$ and 0.399 for the full router, as shown in the last column of Table 11. Cost-adjusted score and regret carry a different hazard. In raw score units they are incomparable across these domains, whose baseline losses differ by three orders of magnitude, so both are divided by the cell's median univariate loss before any row is placed beside another.

Registry, provenance, and build checks. Each registry entry binds a scalar to a source output, its SHA256, and the exact field or deterministic selector used. The build checks declared sources and rejects unresolved placeholders. Numbered deciding stages attach sidecars with the seed, package versions, configuration hash, and command, then write a completion marker last; upstream preregistration, fold, and feature records use the same chain, while simulation and secondary-population reports neither read Fold D nor write a decision. Roll-ups and manuscript assembly verify that the Fold-D access ledger is unchanged, and numerical entry points pin the BLAS thread count to one. Citation records store each live query and title match; when no venue was retrieved, the bibliography uses the verified arXiv preprint.