WHAT A CONTEXT-BLIND PROPOSAL COSTS: STRUCTURED OPERATORS VERSUS SELECTION IN EVENT-ANNOUNCEMENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-conditioned forecasting predicts a series from its numerical history together with natural-language context describing events the history cannot reveal. The dominant training-free recipe keeps a time-series foundation model as the numerical engine and lets a language model *select* among sampled trajectories. We show this carries a cost whenever the proposal is *context-blind*: a selector can only return a trajectory the proposal happened to draw, and the budget needed before an announced deviation appears grows exponentially in how surprising that deviation is—a statement about sample complexity, not impossibility. On the Context-is-Key subset whose context announces a dated, signed, magnitude-specified deviation (13 families, 195 instances), a ground-truth-assisted selector over 256 candidates reaches only 0.864 of the context-blind error in median, while on the same samples a five-parameter operator reaches 0.264, a richer edit schema 0.033, and a frozen 14B model that writes that schema zero-shot, never touching the ground truth, reaches 0.098, beating the selection ceiling on $169/195$ instances. Pooling instead of taking medians moves the levels but not the ordering (0.743, 0.186, 0.061, 0.139). Two oracle-guided sequential proposals do no better (0.782, 0.884). The regime is narrow: across all 71 families the operator ceiling no longer dominates (0.382 against 0.471), and a surprise statistic computed without consulting either branch classifies Time-MMD as out of scope, where the ordering reverses. We have no forecasting-time scope detector; three attempts fail. The contribution is an action-space analysis with a characterised regime, not a general text-conditioned forecaster.

1 INTRODUCTION

Numerical history alone cannot reveal a plant shutdown, a promotion, or a heat-wave advisory. Text-conditioned forecasting (Williams et al., 2025; Liu et al., 2024) supplies that information in natural language and asks a forecaster to use it. The difficulty is a division of labour: time-series foundation models (TSFMs) (Ansari et al., 2024; Woo et al., 2024; Das et al., 2024; Shi et al., 2025; Liu et al., 2025a) produce calibrated numerical futures but cannot read; language models read but are unreliable numerical extrapolators (Merrill et al., 2024; Tan et al., 2024).

The prevailing training-free resolution is to let the language model *select* rather than generate. Nguyen et al. (2026) formalize this: a frozen TSFM acts as a simulator producing candidate continuations, and a frozen LLM supplies a policy and value function that rank and judge them via Monte Carlo tree search. Numerical values stay entirely under TSFM control. Best-of-N reranking and ensemble orchestration (Cao et al., 2025) are special cases of the same move.

This paper argues that the move carries a cost that grows out of reach. A selector returns one of the trajectories it was given, so it inherits whatever the proposal happened to produce. When the context announces an effect the TSFM rarely samples—"the meter will be offline between 08:00 and 14:00, which results in zero readings"—the number of draws needed before such a trajectory appears grows exponentially in how surprising the announcement is under the proposal. We are careful not to claim impossibility: the proposal may well place positive mass there, and our statements are about budgets, not supports. Hua et al. (2026) formalize the corresponding support-inclusion property for

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.

unconditional TSFM inference and show that input perturbations strictly enlarge the support. We take their proposition as our starting point and ask the conditional question they do not: when the required direction of expansion is *stated in text*, is blind expansion still the right instrument?

Our answer is no, and the margin is large. We measure three ceilings under a matched oracle—each branch is given N candidates and the ground truth picks the best—so that the comparison isolates the action space rather than the policy. We then show that the winning action space is not merely reachable in principle: a frozen 14B model reads the context and writes down the parameters directly, with no training and no oracle.

What this paper does and does not claim. The dichotomy we study is not selection versus operators, it is a *context-blind proposal followed by selection* versus a context-aware modification of that proposal. Everything we prove and measure sits on the first side: our propositions assume draws from a fixed proposal, and every baseline we run, including an oracle-guided rollout that stands in for tree search with a perfect judge (Section 5.1), uses a TSFM that never reads the text. A proposal that did read the text is not covered, and building one is the alternative we advocate, with structured operators as one cheap way to do it. Our result is a statement about what a context-blind proposal costs, not a bound on selection as such.

Contributions.

- We show empirically that selection-based text conditioning is bounded well away from the operator ceiling at practical budgets on event-announcement forecasting. On 195 in-scope instances an oracle selector over 256 candidates reaches a median 0.864 against 0.264 for a five-parameter additive operator and 0.033 for the richer edit schema; a separate 12-family budget study reaches 0.634 at $N{=}4096$ against 0.356 for the additive operator on those instances. Every comparison we report is within a single matched instance set (Sections 5, C.2).
- We show the operator parameters are recoverable from text zero-shot, reaching a median 0.098 and beating the selection ceiling on $169/195$ instances; we verify the gain is text-driven with a context-mismatch control, and measure how much of the residual gap is extractor capability with a within-family scaling control (Section 6).
- We show the conclusion replicates across two substantially different TSFM backbones, and that operator parameters—unlike emitted values—transfer between TSFMs unchanged (Section 7).
- We delimit the claim on all 71 Context-is-Key families and show the binding constraint is the text-to-parameter map rather than the action space. A surprise statistic computed independently of either conditioning branch separates the regimes and classifies Time-MMD as out of scope, which its results confirm. Turning that diagnostic into a forecasting-time predicate is unsolved: three detectors fail (Section 7.1 and Appendix D, Appendix F.1).

2 Related work

Text-conditioned forecasting. Context-is-Key (CiK) (Williams et al., 2025) contains 71 tasks whose context carries information absent from the numerical history, scored by a region-of-interest-weighted scaled threshold-weighted CRPS (Allen et al., 2023); Time-MMD (Liu et al., 2024) pairs series with textual reports. The numerical engines these methods build on are a fast-moving family—beyond the backbones we use, Sundial (Liu et al., 2025b) and Chronos-2 (Ansari et al., 2025) are recent entries—and the agentic wrappers around them, such as TimeCopilot (Garza & Rosillo, 2025), TimeSeriesScientist (Zhao et al., 2025) and TS-Reasoner (Yu et al., 2026), orchestrate tools and models rather than changing what the proposal can produce, so our analysis applies to them as well. Methods split three ways: *LLM-as-forecaster* emits numbers directly (Gruver et al., 2023), though controlled evaluations put LLM temporal reasoning barely above chance (Merrill et al., 2024; Tan et al., 2024); *fusion* trains multimodal encoders; and *selection* keeps a frozen TSFM and lets the LLM choose among its samples, of which LAFP (Nguyen et al., 2026) is the strongest instance and the object of our analysis.

Revision, post-training, and inference scaling. PostTime (Liu et al., 2026) post-trains an LLM with SFT and RL to revise, preserve, or ignore a TSFM prior, but emits raw `(timestamp, value)` pairs, so the numbers are produced by the language model; Liao et al. (2026) describe an industrial agent converting business context into forecast edits, evaluated by case study. Our operators sit between these: like PostTime we intervene on a TSFM prior, but the LLM emits a low-dimensional *transformation*, never a value. Hua et al. (2026) study test-time compute for TSFMs, prove that perturbation-based diversification strictly enlarges the sampling support, and propose an oracle-selection metric for headroom. Their setting is unconditional; we adopt their perturbations as our strongest selection baseline and ask what changes when text specifies the direction of expansion.

3 SETUP

Task and metric. A CiK instance provides a history $y_{1:T}$, a context c in natural language, a forecast horizon H, and a region of interest. A forecaster returns $K = 25$ sample trajectories, scored by the benchmark's region-of-interest-weighted scaled CRPS, written RCRPS. All results are reported as RCRPS divided by the *context-blind* score of the same TSFM, so 1.000 means "no better than ignoring the text" and smaller is better.

Matched comparisons. Our experiments use different instance sets (15 seeds in scope, 3 for the sweeps, one for the budget study); every pair of numbers we place side by side comes from a single set and each table says which. The in-scope selection ceiling has a median of 0.864 over 195 instances and a mean of 0.798 over the 39-instance subset, and we report both.

Unparseable extractor output. An extractor is a forecasting component, so it does not get to decline a case. When its output does not parse as the edit schema we apply the identity edit, giving that instance a ratio of exactly 1.000, and every aggregate uses the full denominator. This costs us a little—Qwen2.5-Coder-7B (Hui et al., 2024) fails to parse on $7/195$ in-scope instances and the zero-shot held-out extractor on $24/240$—but excluding those cases would report only the instances where the extractor happened to succeed.

Aggregation. The per-instance ratio is unstable when the context-blind score is near zero: the base RCRPS spans 0.00067 to 5.96, and our largest observed ratio, 2659, occurs on an instance with base 0.00067. Instances with ratio above 2 have a median base of 0.080 against 0.181 overall, so the effect is the denominator, not implausible forecasts. We therefore report the *median* ratio and, where a mean is wanted, the pooled quantity $\sum_i \mathrm{RCRPS}_i / \sum_i \mathrm{base}_i$, which is insensitive to near-zero denominators. Scoring every branch on the same instances against the same base does not by itself make a mean of ratios safe—a small denominator still gets outsized weight—so we give the median alongside it throughout and the pooled figure wherever a mean would otherwise carry the argument; Section F spans the whole benchmark and reports both. We flag this because the two aggregations can rank methods differently, and reporting only one invites the suspicion that it was chosen after the fact.

Backbones. Chronos-T5-base (Ansari et al., 2024) unless stated otherwise; Moirai-1.1-R-small (Woo et al., 2024) for the backbone-independence study. Both produce genuine sample trajectories, which selection baselines require.

The four branches we compare—two oracle-assisted selection ceilings, an oracle operator fit, and the oracle-free text-derived operator—are defined in Appendix A.

3.1 STRUCTURED CONTEXT OPERATORS

Let $\hat{y}^{(k)}_{1:H}$ be the k-th context-blind sample. A structured context operator applies one shared transformation to every sample,

$$\tilde{y}^{(k)}_t \;=\; a\,\hat{y}^{(k)}_t \;+\; b \;+\; m\,\mathbf{1}[s \le t < s + d] \;+\; m_2\,\mathbf{1}[t \ge s_2], \tag{1}$$

with $\theta = (a, b, m, s, d)$ for the 5-parameter *pulse* family and $\theta = (a, b, m, s, d, m_2, s_2)$ for the 7-parameter *pulse_step* family; the 2-parameter *affine* family keeps only (a, b). Sharing θ across samples is essential: it edits the location and shape of the predictive distribution while preserving

the TSFM's uncertainty structure, and it prevents the operator from fitting each sample to the target individually.

The parameters carry meaning that matches how events are described: s is when the event starts, d how long it lasts, m how large it is, a and b a global rescaling. This is what makes them writable from text, and it is the property that distinguishes the action space from emitting values.

Two families, kept distinct. Eq. 1 is purely *additive*: convenient for the analysis of Section 4, but unable to express "four times the usual consumption" or "zero readings". The schema we ask a language model to write is strictly richer—a global scale and offset plus events typed `add`/`multiply`/`set_absolute`—so the additive fit is *not* a bound on it (median 0.264 against 0.033 in scope). All `text_op` numbers use the schema and should be read against the schema oracle.

4 Why selection fails: a sample-complexity account

Let P be the context-blind predictive law of a frozen TSFM over trajectories in $\mathbb{R}^H$, and let $X_1, \dots, X_N \overset{iid}{\sim} P$ be the candidate pool. A *selector* returns a sub-multiset of the pool; the forecast is the empirical law $\hat{F}$ of the selected trajectories, scored per timestep by CRPS. This covers reranking, best-of-N, and our `sel_div` branch, whose perturbation-and-temperature mixture is itself a fixed law that the draws are i.i.d. from. It does *not* cover adaptive proposals: the tree search of Nguyen et al. (2026) expands continuations conditioned on prefixes it has already kept, so its draws are neither independent nor identically distributed, and our propositions say nothing about it directly. The intuition carries—if every proposal in play gives the target region small mass, reaching it is still expensive—but making that precise for sequential proposals is future work, and we do not claim the theory bounds the selection paradigm as a whole.

Proposition 1 (Envelope floor, for a realised pool)**.** *Fix a timestep t and a* drawn *pool $X_1, \dots, X_N$ in which every candidate lies below the truth by at least $\delta_t > 0$, i.e. $X_{i,t} \le y^*_t - \delta_t$ for all $i \le N$. Then for* every *selector,* $\mathrm{CRPS}(\hat{F}_t, y^*_t) \ge \delta_t$.

Proof in Appendix B.

The statement is conditional on the pool actually drawn: an empirical envelope is not the support of P, and unless that region has zero mass a larger N may reach it. Proposition 1 says only that a pool missing the announced effect cannot be repaired by selection; what such a pool costs to draw is answered by Proposition 2. Hua et al. (2026) prove the analogous support-inclusion result for unconditional inference; this is its pointwise metric form and we claim no novelty for it.

Proposition 2 (Selection budget grows exponentially in the announced deviation)**.** *Let $B = \{x : \|x - y^*\|_\infty \le \varepsilon\}$ and $p_\varepsilon = P(B) > 0$. The smallest budget at which some candidate lands in B with probability at least $1 - \eta$ is exactly*

$$N^*(\eta) \;=\; \left\lceil \frac{\log(1/\eta)}{\log\!\big(1/(1-p_\varepsilon)\big)} \right\rceil, \qquad \text{and for } p_\varepsilon \le \tfrac{1}{2}, \quad \frac{\log(1/\eta)}{2\,p_\varepsilon} \;\le\; N^*(\eta) \;\le\; \frac{\log(1/\eta)}{p_\varepsilon} + 1.$$

If $P = \mathcal{N}(\mu, \Sigma)$, write $z = \|y^* - \mu\|_{\Sigma^{-1}}$ for the deviation in proposal standard deviations and $w = \Sigma^{-1}(y^* - \mu)/z$, so $w^\top \Sigma w = 1$. Then $p_\varepsilon \le \exp\!\big(-\tfrac{1}{2}(z - \|w\|_1 \varepsilon)^2\big)$ and hence

$$N^*(\eta) \;\ge\; \tfrac{\log(1/\eta)}{2}\, \exp\!\big(\tfrac{1}{2}(z - \|w\|_1 \varepsilon)^2\big).$$

For isotropic log-concave P—or any log-concave P after normalising the variance along w—the same argument gives $N^*(\eta) \gtrsim e^{c(z - \|w\|_1 \varepsilon)}$.

Proof in Appendix B.

Proposition 2 is what makes the ceiling practical rather than merely asymptotic. The cost of reaching an announced event by sampling is exponential in how surprising that event is under the proposal, so the regime in which text is most informative—an event the history cannot foreshadow—is exactly the regime in which selection is most expensive. An operator family $\{g_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ with $y^* \in \{g_\theta(x) : \theta \in \Theta,\, x \in \operatorname{supp} P\}$ escapes this: once θ is known, a single candidate suffices, and k does not depend on z.

Proposition 3 (Mixture floor, and the bridge to hitting probability)**.** *Fix a timestep t and $\delta > 0$. Let q_t be the $\hat{F}$-weight the selector places on candidates with $|X_{\cdot,t} - y^*_t| \le \delta$, and suppose every other selected candidate lies on one side of y^*_t at distance $> \delta$. Then*

$$\mathrm{CRPS}(\hat{F}_t, y^*_t) \;\ge\; (1 - q_t)^2\, \delta.$$

Consequently $\mathrm{CRPS}(\hat{F}_t, y^*_t) \le \tau$ *forces* $q_t \ge 1 - \sqrt{\tau/\delta}$*. If this holds at every t in a region of interest R with $\gamma := \sqrt{\tau/\delta} < 1/|R|$, then $\hat{F}\big(\bigcap_{t\in R}\{|X_{\cdot,t} - y^*_t| \le \delta\}\big) \ge 1 - |R|\gamma > 0$, so at least one* drawn *candidate lies in the sup-norm δ-ball around y^* restricted to R, and Proposition 2 applies to that event with $\varepsilon = \delta$.*

Proof in Appendix B.

Proposition 3 joins the per-timestep floor to the hitting-time bound and disposes, in the one-sided miss regime relevant to the announced deviations we study, of the objection that a selector can blend several mediocre trajectories instead of drawing one good one. The regime is the common one here: on 17 of 21 in-scope instances the context-blind candidates fall entirely on one side of the truth at every step of the region of interest. Where the misses straddle the truth, a mixture can cancel them and the bound does not apply. Blending is not cheap: the floor falls only as the *weight* on δ-accurate candidates approaches one, and once the demanded accuracy is fine relative to the width of the region of interest, some single drawn trajectory must be accurate throughout it—the event Proposition 2 prices. It bounds the budget needed to reach a given τ, not the RCRPS attainable at a given budget, and δ is ours to choose, so it bites hardest when the announced deviation is large relative to the accuracy demanded.

Directed versus blind expansion. Reaching y^* to resolution ε by uniform search costs $\Theta((r/\varepsilon)^H)$ samples when the proposal is expanded isotropically to radius r, and $\Theta((\operatorname{diam}\Theta/\varepsilon_\theta)^k)$ evaluations when the operator family is searched; if the context determines θ, one evaluation suffices (Appendix C.1). Both search costs are exponential in the dimension of whatever is searched, so a smaller Θ is not by itself an advantage: our budget-matched random operator search over $k{=}5$ reaches only 1.087 (Section 5), *worse* than diversified sampling at the same budget, because P already supplies the trajectory shape a blind operator search must rediscover. The family pays off only when θ is supplied rather than searched—which is what reading the text does.

Testable prediction. Proposition 2 predicts that the selection ceiling degrades towards 1.0 as the announced deviation grows in proposal standard deviations, while the operator ceiling does not. Section 4.1 tests this directly on all 213 instances.

4.1 Testing the prediction

Proposition 2 makes a falsifiable prediction: the selection ceiling should degrade as the announced deviation grows relative to the proposal's spread, whereas the operator ceiling should be indifferent to it. We measure, for each of the 213 instances, the surprise of the truth under the context-blind proposal,

$$z \;=\; \max_{t\in \mathrm{ROI}} \frac{|y^*_t - \mathrm{med}_t|}{\max(\sigma_t,\, 0.05\,\hat{\sigma}_{\mathrm{hist}})},$$

where med_t and σ_t are the pointwise median and standard deviation of 256 context-blind samples. The floor on σ_t matters: without it, timesteps where the proposal is nearly deterministic send z to 10^8 and the statistic reports on $\sigma_t \to 0$ rather than on the announced deviation. Quintile means of the selection ceiling rise monotonically with z $(0.190, 0.376, 0.440, 0.629, 0.670)$ while the operator ceiling does not $(0.313, 0.401, 0.458, 0.325, 0.383)$.

Figure 2 in the appendix plots both relationships.

The prediction holds (Figure 2). Selection correlates with surprise, Spearman $\rho = 0.627$ ($p = 1 \times 10^{-24}$), while the operator ceiling does not, $\rho = 0.030$ ($p = 0.66$)—an effect and a non-effect on the same instances with the same base, which is what separates a mechanism from a confound. The envelope-coverage statistic that Proposition 1 depends on falls with z at $\rho = -0.695$, so the two propositions are consistent with one another rather than merely both true.

Table 1: In-scope families, 13 families × up to 15 seeds = 195 instances, Chronos-T5-base at N=256. Entries are RCRPS relative to the context-blind TSFM under both aggregations of Section 3, with 95% bootstrap intervals. **sel** is ground-truth-assisted, not a proven optimum; the operator rows are oracle fits in the additive family of Eq. 1 and in the extractor's schema; **text_op** uses no ground truth, and an output that does not parse counts as the identity edit.

Branch	median	pooled	improves on base
context-blind base	1.000	1.000	—
sel* (ground-truth-assisted, N=256)	0.864 [0.815, 0.898]	0.743 [0.673, 0.801]	195/195
operator ceiling, additive (Eq. 1)	0.264 [0.229, 0.292]	0.186 [0.159, 0.221]	195/195
operator ceiling, extractor schema	**0.033** [0.031, 0.046]	**0.061** [0.051, 0.077]	195/195
`text_op`, Phi-4 (14B), no oracle	**0.098** [0.088, 0.116]	**0.139** [0.114, 0.173]	185/195
beats the selection ceiling on			**169/195**
`text_op`, Qwen2.5-Coder (7B)	0.200 [0.144, 0.230]	0.357 [0.244, 0.482]	162/195
`text_op`, Qwen2.5-Coder (32B)	0.103 [0.088, 0.120]	0.158 [0.124, 0.205]	185/195
beats the selection ceiling on			165/195

This is correlational evidence for a mechanism, not a measurement of the exponential rate: CiK does not let us vary z with the task held fixed, so we cannot fit $\log N$ against z^2. It does establish that the two branches respond to surprise in opposite ways, the observable signature the account predicts.

5 How far selection actually gets

We first isolate the 13 CiK families whose context announces a dated, signed, magnitude-specified deviation inside the forecast window—seven ElectricityIncrease variants, three ATM scenarios, SensorMaintenanceInPrediction, DecreaseInTraffic, and BoundedPredConstraints. Section 7.1 treats the remaining 58 families and explains the split.

Table 1 reports the result. Giving a selector the ground truth and 256 candidates—a ground-truth-assisted ceiling no practical reranker attains—improves the context-blind forecast only to a median of 0.864, or 0.743 pooled. On the seven ElectricityIncrease variants it barely moves at all: the context announces a fourfold consumption spike for two hours, and no unconditional Chronos sample contains it. Five additive parameters on exactly the same 25 samples reach 0.264, and the richer edit schema 0.033; both beat the selection ceiling on 195/195 instances, under either aggregation.

The operator space is large, not merely better. A budget-matched *random* operator search—256 Sobol draws in the same five-dimensional additive box, oracle-selected—reaches only 1.087, worse than doing nothing. The action space is not easy to search blindly. What makes it usable is that the text names the parameters. This is the precise sense in which text supplies the *direction* of support expansion that Hua et al. (2026) must obtain by brute force.

Figure 1 in the appendix shows three instances where the truth lies outside the sample envelope entirely.

Is the ceiling an artefact of greedy search? No. On 21 in-scope instances we reran the selection branch with three much stronger searches over the same pools—iterated local search, random restarts, and simulated annealing over multisets. At N=64 the best of them improves on greedy by 0.17% on average and 0.61% at worst, moving the branch from 0.752 to 0.751 of the context-blind base, against operator ceilings of 0.033–0.264 in the same units (Appendix C.3).

5.1 Two sequential proposals, neither of which closes the gap

Our propositions assume draws from a fixed proposal, and our selection branch draws its pool up front. The strongest published selection method does neither: LAFP expands continuations conditioned on prefixes it has already kept. We cannot bound that analytically, so we test two oracle-guided instances of it, both standing in for their learned counterparts by replacing the judge with

Table 2: Neither sequential scheme beats drawing the pool up front (21 in-scope instances, 7 families, 3 seeds). All three are oracle-guided and all three use a context-blind proposal.

Branch	draws	mean	median
context-blind base	—	1.000	1.000
ground-truth-assisted beam over blocks (W=25)	472	0.884	0.961
ground-truth-assisted block-greedy rollout	1600	0.782	0.857
ground-truth-assisted selection, non-adaptive	256	**0.677**	**0.688**
operator ceiling, additive	—	**0.319**	**0.339**

the ground truth. **Block-greedy** splits the horizon into B=4 blocks and at each block draws M=16 continuations conditioned on the prefix kept so far, keeping the best. **Beam** maintains W=25 prefixes, expands each with M=8 continuations, and keeps the best 25 of the 200 extensions; its final beam is the forecast. Beam exists because block-greedy commits irrevocably, and a scheme that loses from premature commitment would say little about adaptivity as such.

Both sequential schemes lose to the non-adaptive ceiling, which uses the smallest budget of the three; beam beats it on $0/21$ instances and block-greedy on $3/21$. Two mechanisms are visible. Sequential expansion returns correlated trajectories, beam's final 25 agreeing on every prefix that survived, and CRPS penalises a narrow predictive distribution, whereas subset selection over an i.i.d. pool can assemble a spread one. Both schemes also score prefixes, a surrogate for the metric they are judged by.

We are careful about what this licenses. Both weaknesses belong to our implementations rather than to sequential proposals in general, and a scheme scoring completed rollouts might do better; we did not find that adaptivity fails as a concept. What we did find is that neither scheme comes within a factor of two of the operator ceiling, for a reason that is not implementation-specific: conditioning on a prefix changes *where* the proposal is sampled, not *what* it is willing to produce, and the announced spike is as rare in Chronos's conditional law as in its marginal one. A proposal that reads the text would change the latter, which is why we advocate building one rather than searching harder inside one that does not.

Increasing the search budget. Pushing the non-adaptive branch to N=4096 on 12 families moves `sel_div` from 0.878 to 0.634, about 0.041 per doubling, against an additive operator ceiling of 0.356 on the same samples; a log-linear extrapolation puts the crossover near $N \approx 4 \times 10^5$, three to four orders of magnitude beyond the tens to hundreds of simulations LAFP uses. This study has one seed per family and the crossover is a projection rather than a bound, so we treat it as supporting evidence (Appendix C.2).

6 Operator parameters are recoverable from text

We give a frozen Phi-4 (Abdin et al., 2024) the context, the numbered future timestamps, and a summary of the history, and ask for a JSON object containing a baseline scale, a baseline offset, and a list of events, each with a start index, end index, type (`set_absolute`, `add`, `multiply`) and value. The result is applied to the context-blind samples and scored. Nothing in this pipeline observes the ground truth.

On the in-scope families this reaches a median of 0.098 and a pooled 0.139 (Table 1), a tenfold reduction in median RCRPS, and beats the selection ceiling on $169/195$ instances. How much of the residual gap is the extractor rather than the text? Comparing Phi-4 to a 32B model would confound family with size, so we scale within one family: Qwen2.5-Coder reaches a median 0.200 at 7B and 0.103 at 32B, and pooled 0.357 against 0.158, with unparseable output on $7/195$ against $0/195$. Capacity clearly helps. Across families it is not the whole story, since Phi-4 at 14B (0.098) matches Qwen at 32B (0.103), so family and training data matter at least as much as parameter count—but the honest reading is that the distance from 0.098 to the schema ceiling at 0.033 mixes two causes: information the context never carried, and extraction a stronger model would get right. We cannot apportion them, and we do not claim, as an earlier version of this work did, that scale is irrelevant here. The two numbers answer different questions and should not be conflated: 0.033 says

the operator family, which includes primitives as strong as `set_absolute` and `multiply`—can be fitted to the truth, whereas 0.098 says how much of that the context actually determines. Only the latter is attainable at forecasting time, and it is the figure a deployment should be judged on.

Was the schema designed around the benchmark? A fair worry is that primitives as strong as `multiply` and `set_absolute` were chosen after looking at CiK's event families. The experiment that would settle it, freezing the schema and evaluating on an event-announcement benchmark whose task design we never saw, is one we cannot run because no such benchmark exists to our knowledge. Three things bound the concern. The conclusion does not depend on the suspicious primitives: refitting with `add` alone reaches 0.313 in mean ratio, still far below the selection ceiling at 0.677, and the richer primitives take it to 0.171 without creating the gap. Each primitive then earns its keep exactly where the announcement's wording uses that operation, as a schema built around announcement language rather than around this benchmark should: on ELECTRICITYINCREASE, whose context says "four times the usual", `multiply` moves the fit from 0.445 to 0.027 while `set_absolute` does nothing (0.438); on CASHDEPLETED and SENSORMAINTENANCE, whose contexts say "no withdrawals" and "zero readings", the ordering reverses (0.260 and 0.189 with `set_absolute` against 0.287 and 0.214 with `multiply`). A schema fitted to the benchmark rather than to the language would not show that alignment. Finally, the same schema applied unchanged to Time-MMD does not help at all, and on the 58 out-of-scope CiK families the zero-shot extractor is net harmful.

Context-mismatch control. A skeptical reading is that the operator merely reshapes the predictive spread in a way the metric rewards. We test this by transplanting the operator emitted for instance i onto instance j with the same horizon and a different task family. Matched operators score 0.550; mismatched operators score $\mathbf{1.414}$, and are worse than applying no edit at all on $38/42$ instances, with matched better than mismatched on $41/42$. The gain comes from reading the text.

Cross-seed control: the edit is not a per-family constant. A second reading is that each family admits one good edit, so the operator merely encodes a family-level template and the text is incidental. CiK re-randomises the event time and magnitude per seed, which lets us test this directly: we fit an oracle edit on seed i and apply it to seed j of the *same* family. Fitted in place the edits reach 0.291; transplanted across seeds within the family they reach a median of 2.132, and are worse than no edit at all in 65 of 71 families. The operator parameters are instance-specific, and the per-instance context is what determines them.

7 GENERALITY AND SCOPE

We repeat the entire sweep with Moirai-1.1-R-small at a budget of $N{=}128$ matched between the two backbones. Because the emitted parameters describe a transformation rather than values, we reuse the *same* LLM outputs; only the samples they act on change. Every qualitative claim survives the swap and two get stronger. In scope the operator ceiling beats the selection ceiling on $39/39$ Moirai instances against $38/39$ on Chronos, and the zero-shot text operator again lands far below selection (0.248 against 0.763) using parameters produced for a different model. On the full benchmark it wins on $189/213$ instances against $121/213$, because Moirai's proposal is less dispersed and selection therefore extracts less from the same budget while the operator ceiling is unaffected, the asymmetry Proposition 2 predicts.

The scope boundary replicates too: on both backbones the extractor is net harmful out of scope (1.698 and 1.298) and strongly helpful in scope (0.145 and 0.248), so the partition belongs to the tasks rather than to Chronos. Parameter portability is itself a property of the action space, since a method that emits values, as PostTime does (Liu et al., 2026), must regenerate them for each backbone while one that emits "multiply by 4.0 over steps 8–10" need not. Figure 4 in the appendix gives the per-family comparison.

7.1 WHERE OPERATORS DO NOT APPLY

The full sweep over all 71 families (213 instances) delimits the claim. Outside the announcement regime the operator ceiling merely ties selection (0.384 against 0.398 at $N{=}256$); over all 71 it

Table 3: The predicate classifies Time-MMD as out of scope from z alone, without consulting either conditioning branch, and the ordering of the branches duly reverses. All three rows are Chronos at $N{=}256$ with the additive operator family, so the budget is matched across datasets; CiK uses 3 seeds per family.

	measured from the data alone		method results		
Dataset / subset	median z	envelope cov.	sel*	op_fit$^+$	text_op
CiK in-scope (39)	**41.24**	**0.095**	0.798	**0.374**	**0.145**
CiK out-of-scope (174)	2.19	0.665	**0.398**	0.384	1.698
Time-MMD (218)	1.98	0.838	**0.363**	0.425	1.000

reaches 0.382 against 0.471, winning on $40/71$ families and $121/213$ instances (Table 3, Figure 3). The families it loses on are coherent: analogical retrieval across Montreal-fire incident types, cross-series transfer for county unemployment, physical priors for solar irradiance, and ORACLEPREDUNIVARIATECONSTRAINTS, where projection is the right tool. None is a localized edit to a trajectory, which is all Eq. 1 can express. The zero-shot extractor degrades further still, reaching 1.698 because it edits whenever it can find an edit to make. A gate suppressing the harmful edits would recover the mean to 0.815 and fire on 37% of instances, but a naive heuristic keyed on whether the edit is localized fails outright (1.539 against 1.055), so knowing *when* to edit remains unsolved.

That the extractor degrades faster than the action space is the point. On families where it fails we fitted the operator family to the ground truth to see whether the required edit was expressible at all (Appendix C.4). For two it plainly was: SENSORTRENDACCUMULATION admits a fit at 0.035 against the extractor's 0.991, and SENSORPERIODICMAINTENANCE 0.261 against 1.014. The emitted JSON explains the second. Its context says the sensor *was* offline daily and "will not be in maintenance in the future", and the model applies the maintenance instead of removing it, misreading a negation rather than meeting an inexpressible edit. Ramp and gate-fill primitives do not help either, so expressivity is not the binding constraint. Only SENSORSPIKE and ATMUNDERPERIODICMAINTENANCE show genuinely limited headroom.

An independent dataset the predicate rules out in advance. A scope claim is worth little if it can only be applied after seeing the results. Ours can be applied without consulting either branch, since the regime is characterised by the surprise statistic z of Section 4.1, which reads the context-blind proposal and the realised target and is therefore a dataset diagnostic rather than a forecasting-time predicate. Time-MMD (Liu et al., 2024) pairs series from nine domains with qualitative commentary rather than dated, signed, magnitude-specified announcements, and its surprise distribution is indistinguishable from CiK's out-of-scope families (median z 1.98 against 2.19), with the truth inside the sample envelope 84% of the time against 9.5% in scope. The predicate rules it out of scope, and the results follow: over 218 instances the selection ceiling is at its strongest anywhere (0.363), the operator ceiling is worse than selection and wins on only $66/218$, and the zero-shot text operator has median ratio exactly 1.000. A scope condition that only ever explains successes would be unfalsifiable; this one predicted that our own approach would fail on an unseen benchmark and that the baseline we argue against would win there. Full protocol and per-domain results are in Appendix D.

8 CONCLUSION

Selection cannot represent what it was never handed. Where the context announces a dated, signed, magnitude-specified deviation, a ground-truth-assisted selector with a 4096-candidate budget stays far from what a handful of interpretable parameters achieve, and a frozen language model that writes those parameters zero-shot beats it by a wide margin. What buys this is moving the model's output from values to a low-dimensional, portable transformation. Where the idea stops is equally clear: a surprise statistic marks the regime, but it reads the realised target and so diagnoses datasets rather than gating forecasts, and deciding when to edit remains open. We offer an action-space result with a characterised regime, and a precise statement of what a deployable method still needs.

REPRODUCIBILITY STATEMENT

Every number in this paper comes from a script we ran, and the pipeline is deterministic given the seeds we report. Task instances are generated by the public Context-is-Key and Time-MMD releases at the seeds stated in each table; forecasts come from the public Chronos-T5 and Moirai checkpoints; the extractors are the public Phi-4 and Qwen2.5-Coder releases used without modification, decoded greedily except in the self-consistency study, where the temperature is stated. Oracle branches consume the realised target and are labelled as such wherever they appear, so no reported method number depends on information unavailable at forecasting time. The branches are specified in Appendix A, the operator families in Section 3.1, and the fitting procedure and search settings in Appendix C; the prompt, the edit schema and the exact parameter bounds and optimiser budgets are in the released code. Code and the generated artefacts will be released.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Sam Allen, David Ginsbourger, and Johanna Ziegel. Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):906–940, 2023. doi: 10.1137/22M1532184.

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. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Defu Cao, Michael Gee, Jinbo Liu, Hengxuan Wang, Wei Yang, Rui Wang, and Yan Liu. Conversational time series foundation models: Towards explainable and effective forecasting. *arXiv preprint arXiv:2512.16022*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024.

Azul Garza and Renée Rosillo. Timecopilot. *arXiv preprint arXiv:2509.00616*, 2025.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ruijin Hua, Zichuan Liu, Kun Zhang, and Yiyuan Yang. Diversified scaling inference in time series foundation models. *arXiv preprint arXiv:2601.17376*, 2026.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Yuhua Liao, Zetian Wang, Qiangqiang Nie, and Zhenhua Zhang. Bridging the last mile of time series forecasting with LLM agents. *arXiv preprint arXiv:2606.02497*, 2026.

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

Haoxin Liu, Yichen Zhou, Rajat Sen, B. Aditya Prakash, and Abhimanyu Das. Rethinking post-training recipes for multimodal time-series forecasting. *arXiv preprint arXiv:2605.29401*, 2026.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3512–3533, 2024. doi: 10.18653/v1/2024.findings-emnlp.201.

Huu Hiep Nguyen, Dung Nguyen, Minh Hoang Nguyen, Dai Do, and Hung Le. LLM as forecasting planner: Training-free text conditioning for time-series foundation models. *arXiv preprint arXiv:2607.24892*, 2026.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations (ICLR)*, 2025.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Andrew R. Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *International Conference on Machine Learning (ICML)*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Fangxu Yu, Hongyu Zhao, and Tianyi Zhou. TS-reasoner: Aligning time series foundation models with LLM reasoning. *Transactions on Machine Learning Research*, 2026.

Haokun Zhao, Xiang Zhang, Jiaqi Wei, Yiwei Xu, Yuting He, Siqi Sun, and Chenyu You. Timeseriesscientist: A general-purpose AI agent for time series analysis. *arXiv preprint arXiv:2510.01538*, 2025.

A BRANCHES, PROMPT, AND EDIT SCHEMA

Branches. All branches operate on samples from the same frozen, context-blind TSFM. Given a budget N:

- **sel_std**(N): greedy forward selection of K trajectories out of N standard draws, *with the ground truth in the objective*. We call this oracle-assisted rather than optimal: the subset objective mixes a fidelity term with the pairwise dispersion term of the CRPS and we do not prove greedy is its global minimiser. We check below with much stronger combinatorial search that the reported gap is not a greedy artefact.
- **sel_div**(N): the same, over N trajectories drawn with the task-agnostic perturbations of Hua et al. (2026) (Gaussian noise, random offset, prefix/suffix padding, missing data) across temperatures $\{1.0, 1.3, 1.6\}$.
- **op_fit**: a structured operator with 2, 5, or 7 parameters fitted to the ground truth by differential evolution, applied identically to all K base samples.
- **text_op**: the same operator family, with parameters produced by a frozen Phi-4 (Abdin et al., 2024) reading only the context, timestamps, and history summary. *No ground truth enters this branch.*

B PROOFS

Proof of Proposition 1.

Proof. $\hat{F}_t$ is supported in $(-\infty, y_t^* - \delta_t]$, so its CDF satisfies $\hat{F}_t(x) = 1$ for $x \geq y_t^* - \delta_t$. Using $\mathrm{CRPS}(F, y) = \int (F(x) - \mathbf{1}\{x \geq y\})^2 \, dx$,

$$\mathrm{CRPS}(\hat{F}_t, y_t^*) \;\geq\; \int_{y_t^* - \delta_t}^{y_t^*} \big(\hat{F}_t(x) - 0\big)^2 dx \;=\; \delta_t. \qquad \square$$

Proof of Proposition 3.

Proof. Write $G = \{x : |x - y_t^*| \leq \delta\}$ and suppose the selected mass outside G sits below $y_t^* - \delta$ (the mirror case is identical). For $x \in [y_t^* - \delta,\, y_t^*)$ we have $\mathbf{1}\{x \geq y_t^*\} = 0$ and $\hat{F}_t(x) \geq 1 - q_t$, because all mass above x is confined to G and carries weight at most q_t. Hence

$$\mathrm{CRPS}(\hat{F}_t, y_t^*) = \int \big(\hat{F}_t(x) - \mathbf{1}\{x \geq y_t^*\}\big)^2 dx \;\geq\; \int_{y_t^* - \delta}^{y_t^*} (1 - q_t)^2 \, dx = (1 - q_t)^2 \delta.$$

Solving $(1 - q_t)^2 \delta \leq \tau$ gives $q_t \geq 1 - \sqrt{\tau/\delta} = 1 - \gamma$. For the last claim, let S_t be the set of selected candidates in G at time t, so $\hat{F}(S_t) \geq 1 - \gamma$; then $\hat{F}(\bigcap_{t \in R} S_t) \geq 1 - \sum_{t \in R} (1 - \hat{F}(S_t)) \geq 1 - |R|\gamma$, which is positive when $\gamma < 1/|R|$, so the intersection contains a candidate that was actually drawn. $\square$

Proof of Proposition 2.

Proof. $\Pr[\text{no candidate in } B] = (1 - p_\varepsilon)^N$, which is non-increasing in N, so the smallest admissible N is the least integer with $(1 - p_\varepsilon)^N \leq \eta$, i.e. $N \geq \log \eta / \log(1 - p_\varepsilon) = \log(1/\eta) / \log(1/(1 - p_\varepsilon))$, giving $N^*(\eta)$. (We stress the direction: the familiar bound $(1 - p)^N \leq e^{-Np}$ yields only the *sufficient* condition $N \geq \log(1/\eta)/p_\varepsilon$, which is an upper bound on N^*, not a lower one.) For the bracket, $-\log(1 - p) = \sum_{k \geq 1} p^k / k$ satisfies $p \leq -\log(1 - p) \leq p/(1 - p)$, and $p/(1 - p) \leq 2p$ when $p \leq \frac{1}{2}$; substituting the two sides into $N^*(\eta)$ gives the stated inequalities.

For the Gaussian claim, $\langle w, X - \mu \rangle \sim \mathcal{N}(0, 1)$ by the choice of w, and $\langle w, y^* - \mu \rangle = z$. Every $x \in B$ obeys $\langle w, x - \mu \rangle \geq z - \|w\|_1 \|x - y^*\|_\infty \geq z - \|w\|_1 \varepsilon$ by Hölder, so B lies in that half-space and $p_\varepsilon \leq \bar{\Phi}(z - \|w\|_1 \varepsilon) \leq \exp(-\frac{1}{2}(z - \|w\|_1 \varepsilon)^2)$ for $z > \|w\|_1 \varepsilon$. Combining with the lower bracket $N^*(\eta) \geq \log(1/\eta)/(2p_\varepsilon)$ gives the claim. Log-concave measures, once the variance along w is normalised to one, satisfy $\Pr[\langle w, X - \mu \rangle \geq t] \leq e^{-ct}$ for a universal c, which replaces the Gaussian tail in the same argument. $\square$

C SUPPORTING ANALYSES

C.1 DIRECTED VERSUS BLIND EXPANSION

Proposition 4 (Directed versus blind expansion)**.** *Reaching y^* to resolution ε costs, by uniform search, $\Theta\big((r/\varepsilon)^H\big)$ samples when expanding the proposal isotropically to radius r, and $\Theta\big((\operatorname{diam}\Theta/\varepsilon_\theta)^k\big)$ evaluations when searching the operator family. If the context determines θ, one evaluation suffices.*

Proposition 4 is a warning as much as a promise. Both search costs are exponential in the dimension of whatever space is searched, so a smaller Θ is not by itself an advantage: our budget-matched random operator search over $k{=}5$ reaches only 1.087 (Section 5), *worse* than diversified sampling at the same budget, because P already supplies the trajectory shape a blind operator search must rediscover. The family pays off only when θ is supplied rather than searched—which is what reading the text does.

C.2 BUDGET STUDY TO $N{=}4096$

The natural objection is that $N{=}256$ is small. We push one seed of each of 12 event families to $N{=}4096$.

Averaged over those families, `sel_std` falls $0.908 \to 0.879 \to 0.817 \to 0.780$ and `sel_div` $0.878 \to 0.856 \to 0.794 \to 0.634$ as N goes $64 \to 256 \to 1024 \to 4096$, against an additive operator ceiling of 0.356 on the same samples. A $64\times$ budget increase moves the stronger branch from 0.878 to 0.634, about 0.041 per doubling. Log-linear extrapolation places the crossover with the operator ceiling near $N \approx 4 \times 10^5$, three to four orders of magnitude beyond the tens-to-hundreds of simulations used by LAFP (Nguyen et al., 2026). We report this as a practical, not an information-theoretic, statement: the gap is not closed by any budget a test-time system would spend.

At $N{=}4096$ the operator still wins on 8 of 12 families. The four exceptions (DECREASEINTRAFFIC, SENSORSPIKE, ATMUNDERPERIODICMAINTENANCE) share a signature: either the context-blind error is already small, so there is little headroom, or the required change is to *remove* a pattern the TSFM extrapolated rather than to add an announced event.

C.3 STRONGER SEARCH FOR THE SELECTION BRANCH

Our selection branch minimises the CRPS of the selected multiset greedily, and that objective mixes fidelity with the CRPS dispersion term, so the headline gap could in principle come from a weak search rather than from the action space. We tested this on 21 in-scope instances (7 families, 3 seeds, $N{=}64$, $K{=}25$): best-improvement local search over single replacements iterated to a local optimum, the same search from three random restarts, and simulated annealing over multisets.

Relative to greedy, the best of the three improves the selected CRPS by 0.21% on average (0.12% median, 0.89% maximum) under standard sampling and 0.17% (0.10% median, 0.61% maximum) under diversified sampling; no instance improves by more than 1%. In the units used throughout the paper the branch mean moves from 0.752 to 0.751 of the context-blind base under diversified sampling (0.822 to 0.821 under standard sampling); the raw scores behind those ratios are $0.2132 \to 0.2128$ and $0.2434 \to 0.2431$. The budget here is $N{=}64$, which is why the level sits above the 0.677 reported at $N{=}256$ in Table 2. Operator ceilings on these families lie between 0.033 and 0.264 of the same base in median, so a sub-percent search gain cannot account for the difference. We still avoid the word optimal: three qualitatively different searches agreeing with greedy to within a fraction of a percent is evidence, not a proof that greedy attains the minimum.

C.4 OPERATOR-FAMILY ABLATION

Oracle fits on the families where the zero-shot extractor fails, comparing the additive pulse family against a ramp variant, a window-replacing *gatefill* variant, and their union. Expressivity is not the binding constraint on four of the five.

Table 4: Operator-family ablation, oracle fits, RCRPS relative to context-blind base (3 seeds). Zero-shot text_op on these families is ≈ 1.0 or worse.

Family	pulse	ramp	gatefill	full
SensorTrendAccumulation	0.035	0.038	0.035	0.035
SensorMaintenanceInPrediction	0.265	0.486	0.191	0.271
SensorPeriodicMaintenance	0.261	0.446	0.297	0.309
ATMUnderPeriodicMaintenance	0.664	0.729	0.670	0.637
SensorSpike	0.842	0.933	0.807	0.877

D TIME-MMD: THE PREDICATE APPLIED TO AN UNSEEN DATASET

A scope claim is worth little if it can only be applied after seeing the results. Ours can be applied without using the outcomes of either the selection or the operator branches: the regime is characterised by the surprise statistic z of Section 4.1, computed from the context-blind proposal and the realised target. Because z uses the target it is a dataset diagnostic, not a forecasting-time predicate; a deployable per-instance gate is a different object and we do not have one (Appendix F.1).

Time-MMD (Liu et al., 2024) pairs series from nine domains with textual reports whose text is qualitative commentary—"gasoline prices may stabilise or decrease slightly"—rather than a dated, signed, magnitude-specified announcement. The predicate therefore rules Time-MMD out of scope, and Proposition 2 says selection should do *well* there. We built 218 instances across the nine domains (context 96, horizon 12, text restricted to reports published strictly before the forecast start) and ran the identical branches with a scaled CRPS in place of RCRPS.

Time-MMD's surprise distribution is indistinguishable from CiK's out-of-scope families (median z 1.98 against 2.19) and its truth lies inside the sample envelope 84% of the time, against 9.5% in scope. The results follow: the selection ceiling is *strong* (0.363, its best showing anywhere), the operator ceiling is worse than selection and wins on only $66/218$ instances, and the zero-shot text operator has median ratio exactly 1.000 and helps on $37/218$. Its mean, 44.8, is meaningless—one Economy instance contributes a ratio of 238 on a near-zero base, the instability documented in Section 3.

A scope condition that only ever explains successes is unfalsifiable. This one predicted, from a statistic that consults neither branch, that our own method would fail on an independent benchmark and that the baseline we argue against would win there; both happened. It also sharpens the claim: not that structured operators are a better way to do text-conditioned forecasting in general, but that they are the right instrument in a regime that CiK isolates and Time-MMD essentially lacks.

E ADDITIONAL FIGURES

The figures below support Sections 5–7.1; each is referenced from the main text where its numbers are discussed.

Qualitative instances. Three in-scope instances drawn to scale, showing directly that the truth can lie entirely outside the sample envelope, which is the condition of Proposition 1.

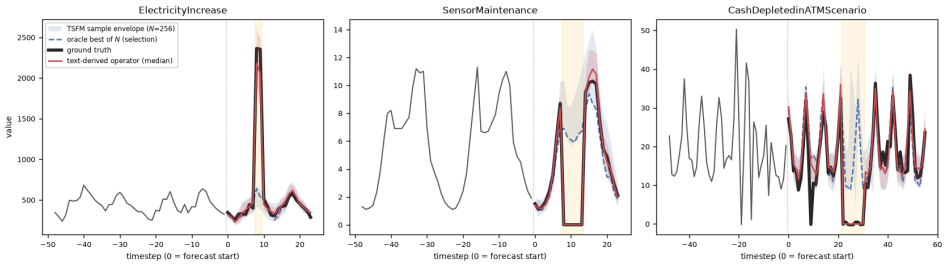


Figure 1: Three in-scope instances. Shaded blue is the envelope of 256 context-blind Chronos samples; the dashed blue line is the best single trajectory an *oracle* selector could pick from them. On the left the context announces a fourfold two-hour spike: the truth reaches 2363 while the envelope of these 256 draws tops out at 715, so by Proposition 1 no selector recovers it *from this pool*. Proposition 2 is what says a pool that does contain it is exponentially expensive to draw; we do not claim the proposal assigns the region zero mass. The red line is a five-parameter operator whose parameters were written by a frozen language model reading only the context; it lands on the truth. Shaded orange marks the metric's region of interest. The contexts are, left to right: "a heat wave … for 2 hours … and 4 times the usual electricity being consumed", for which the model emits `multiply by 4.0 over steps 8--10`; "the meter will be offline for maintenance between 08:00 and 14:00, which results in zero readings"; and "cash is depleted in the ATM … for 9 days, resulting in no withdrawals during that period".

Surprise statistic against the two ceilings. The per-instance data behind the correlations reported in Section 4.1, over all 213 CiK instances.

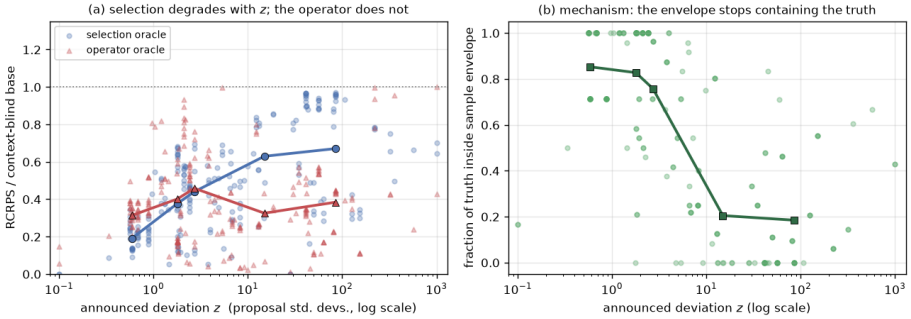


Figure 2: Testing Proposition 2 on all 213 instances. **(a)** The selection ceiling degrades systematically as the announced deviation z grows, while the operator ceiling is flat. **(b)** The mechanism of Proposition 1: the fraction of the truth lying inside the sample envelope collapses as z grows, which is exactly the condition under which the floor binds. Points are instances; lines are quintile means.

Budget curves and per-family comparison (Chronos). Budget curves split by family type, and the per-family operator-versus-selection comparison at $N{=}256$ on Chronos.

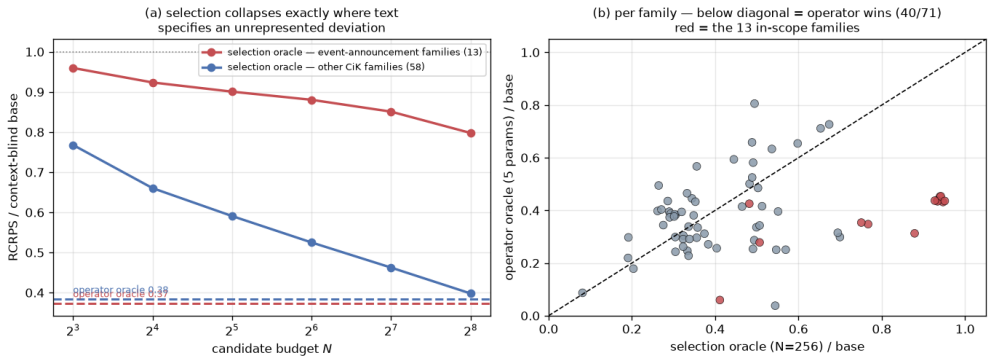


Figure 3: **(a)** Oracle selection ceiling versus budget, split by family type. On event-announcement families (red) selection stalls near 0.80 while the operator ceiling sits at 0.37; on the remaining families (blue) selection descends to meet the operator ceiling. **(b)** Per-family comparison across all 71 CiK families; points below the diagonal are families where the operator oracle beats the selection ceiling. Red points are the 13 in-scope families.

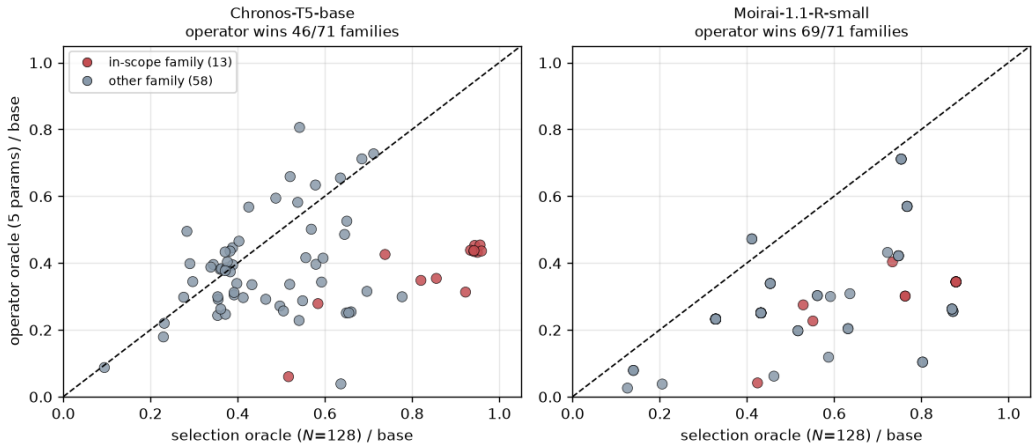


Figure 4: Per-family operator ceiling against selection ceiling at a matched budget $N{=}128$, on both backbones. Points below the diagonal are families where five fitted parameters beat everything a ground-truth-assisted selector can extract from 128 candidates: $46/71$ on Chronos and $69/71$ on Moirai. (At $N{=}256$ the Chronos count is $40/71$; a larger budget helps selection, not the operator.) Red points are the 13 in-scope families, which sit far below the diagonal on both backbones.

F TRAINING THE EXTRACTOR, AND LIMITATIONS

Since the binding constraint is the text-to-parameter map, we tested the obvious remedy: supervised fine-tuning on oracle-fitted edits with LoRA (Hu et al., 2022) and a family-level split. The two aggregations disagree about it, which is itself the finding. Training improves the median ratio from 1.000 to 0.764 and the number of instances it helps from $109/240$ to $150/240$, but it worsens the pooled ratio from 1.010 to 1.632, because a few large absolute failures dominate the pooled sum; a value guard that clips implausible edits brings pooling to 0.783 at a small median cost. We therefore do not claim training helps outright: it helps on most instances and hurts in aggregate unless the tail is controlled. Full protocol and tables are in Appendix F.1. The ceiling results rest on oracles and do not depend on any extractor, so what this adds is that closing the gap is an open problem, not an engineering detail.

Our oracle-assisted branches are empirical ceilings rather than proven optima, and none of them is a method; the only branch we propose as a method is `text_op`, which uses no ground truth. An operator fit consumes more oracle information than a selection index, hence the budget-matched random-operator control (1.087) and the oracle-free number. The budget study uses one seed per family at $N{=}4096$ and its extrapolated crossover is a projection, not a bound.

The scope set is the weakest part of the methodology. We identify it by the regularity in Section 7.1, a context that names a dated, signed, magnitude-specified deviation, but we have no test that applies that description at forecasting time: z separates the regimes and classifies an unseen dataset correctly, yet it reads the realised target and so cannot gate individual forecasts. Three detectors that do not read it fail: a supervised probe (AUROC 0.654, against 0.5 for chance), the operator's own predicted deviation, and extractor self-consistency. We report every out-of-scope family individually (Section 7.1) so a reader who rejects our description can audit the split.

F.1 Full protocol and results

Section 7.1 locates the binding constraint in the text-to-parameter map. We test the obvious remedy directly. Using the oracle fit as supervision, we build 1065 instances across all 71 families—each a context paired with the ground-truth-fitted edit in the exact JSON schema the extractor emits—and fine-tune Qwen2.5-Coder-7B-Instruct (Hui et al., 2024) with LoRA (Hu et al., 2022) ($r{=}32$, 3 epochs, final training loss 0.236). The split is at *family* level: 825 training instances from 55 families, 240 held-out instances from 16 families that never appear in training, chosen to span both regimes. The zero-shot branch is the same base model without the adapter, isolating what training buys.

Table 5: Held-out families (16 families, 240 instances, none seen in training). "guard" clips the edited trajectory to the historical range widened by 3σ and rejects implausibly large moves; it uses no ground truth.

	median ratio	pooled $\sum$RCRPS/ $\sum$base	improves on base
zero-shot	1.000	1.010	109/240
zero-shot + guard	1.000	1.051	108/240
trained	**0.764**	1.632	149/240
trained + guard	0.796	**0.783**	**150/240**
oracle edit	0.223	0.167	—

Training generalises across families. On the 16 unseen families the trained extractor leaves no output unparseable ($0/240$ against $24/240$ zero-shot), improves on the context-blind forecast on $150/240$ against $109/240$, and lowers the median ratio from 1.000 to 0.764. The gains appear on families where the zero-shot model simply does nothing: ZenithInfoHalfDaySolar $1.007 \to 0.353$, MontrealFireTrashAndBicycle $2.157 \to 0.742$, UnemploymentCountySingleState $1.000 \to 0.776$, SensorPeriodicMaintenance $1.005 \to 0.806$.

It also does not solve the problem. Three findings temper the result. First, the trained extractor is *blunter* in scope: on held-out in-scope families the median rises from 0.098 to 0.173, and it regresses on ElectricityIncreaseWithSplitContext ($0.038 \to 0.140$) and SensorMaintenanceInPrediction ($0.114 \to 0.269$). Training on oracle targets it cannot infer appears to teach caution, trading sharpness where the text is explicit for safety where it is not. Second, the two aggregations disagree: `trained` wins on the median but loses badly under pooling (1.632), because a handful of large absolute failures dominate the pooled sum; adding the guard reverses this (0.783) at a small median cost. Third, and most important, every configuration remains far from the oracle edit (0.167 pooled). The extractor recovers a minority of the gain its own action space makes available.

We report this as a negative-leaning result on purpose. The action-space claims of Sections 5–7 are established by oracles and do not depend on any extractor; what this section shows is that closing the remaining gap is a genuine open problem rather than an engineering detail, and in particular that supervised imitation of oracle edits is not sufficient.

F.2 CAN THE SCOPE BE RECOVERED FROM THE CONTEXT ALONE?

Two of our open issues are the same question: an oracle gate that suppresses harmful edits recovers much of the achievable gain, and our in-scope partition is currently defined by outcomes rather than read off the text. Both would be settled by a classifier that decides, from the context alone, whether the text licenses an executable edit. We train one. Labels come from the training families (825 instances, 71% positive): $y = 1$ when the extractor's edit improves RCRPS. Inputs are the context with all numbers, dates and the history summary stripped, so the classifier judges the language and not the series. Evaluation is on the same 16 unseen families.

Table 6: Learned scope gate on held-out families. Neither probe separates the classes well, and no threshold dominates always-editing on both aggregations.

	pooled $\sum$RCRPS/ $\sum$base	median ratio	improves on base
never edit	1.000	1.000	0/240
always edit	1.632	**0.764**	**149/240**
learned gate (embeddings, τ=0.3)	**0.680**	1.000	108/240
oracle gate	0.525	0.764	149/240

A TF-IDF probe reaches AUROC 0.620 on the unseen families, against 0.5 for chance; mean-pooled hidden states from the same 7B model reach 0.654. Average precision tells the same story against its own baseline, the positive rate of 0.621: 0.773 for the surface probe and 0.716 for the semantic one. The property is therefore only weakly decidable from the context by these probes, and semantic features buy little over surface ones. The downstream picture (Table 6) is genuinely two-sided: under pooled RCRPS the learned gate cuts 1.632 to 0.680, capturing roughly 86% of the distance to the oracle gate, because it suppresses precisely the large absolute failures the pooled sum is sensitive to. Under the median it is a regression, from 0.764 to 1.000, and the number of improved instances falls from 149 to 108: the gate buys tail safety by declining edits that would have helped. No threshold we tried dominates always-editing on both aggregations.

Two further detectors that do not need the ground truth. The classifier above is trained on labels that require y^*. Two statistics do not. The first, $\hat{z}$, is the deviation the emitted operator itself predicts, $\max_t |\mathrm{med}(g_\theta(X))_t - \mathrm{med}(X)_t|/\sigma_t$, computable at forecast time from the proposal and θ alone. It does track the true surprise (Spearman 0.386, $p = 7 \times 10^{-9}$) but is useless as a gate: a hallucinated large edit has a large $\hat{z}$ too, and no threshold brings the pooled ratio below 1.000 (best 1.025 at $\hat{z} \geq 50$, against 1.209 for always-editing and 0.826 for the oracle gate).

The second asks whether the context *determines* θ, which is the scope predicate itself: we sample the extractor k=5 times at temperature 0.8 and measure the mean pairwise disagreement D of the induced median forecasts. The prediction is that text-determined parameters agree and invented ones scatter. It fails outright. In-scope instances do have median $D = 0$, but so do $68/170$ out-of-scope instances—the extractor is exactly as self-consistent when it is inventing an edit as when it is reading one off the text. $-D$ predicts whether the edit helps at AUROC 0.510, indistinguishable from chance (Spearman -0.017, $p = 0.80$), and gating on D peaks at a pooled 1.014, worse than never editing. This is a concrete statement about structured-output hallucination: decoding determinism carries no signal about whether the input licensed the output.

We draw two conclusions. The gating problem is real and resists three quite different attacks—a supervised classifier, a magnitude statistic, and a self-consistency statistic—so we do not claim to have solved it. And because the classifier cannot recover the partition, the in-scope set in Section 7.1 remains a description rather than a decidable predicate, which we state plainly below rather than paper over.

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.