WATERMARKING VALUES, NOT TOKENS: DISTRIBUTION-FREE VERIFICATION FOR LLM TIME-SERIES FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can generate probabilistic time-series forecasts as numerical ensembles that let downstream users account for predictive uncertainty rather than rely on a single point forecast. When such forecasts are released and reused, verifying their provenance becomes important. Existing watermarking methods, however, are poorly aligned with this setting in two respects. First, watermark evidence is typically defined over generated tokens, whereas forecast releases are manipulated as numerical values through operations such as rounding, reordering, subsampling, and cropping. Second, watermarking an autoregressive forecast feeds modified values back into later predictions, so local edits propagate through the forecast horizon. To address these two limitations, we propose V-band, a generation-time value-space watermarking framework for frozen autoregressive language-model forecasts. V-band uses the owner's secret key to label numerical bands at each forecast step and tilts the model's conditional digit probabilities so that sampled values are more likely to fall within preferred bands. We further introduce value-feedback-free steering, which preserves unmarked history across forecast steps while conditioning lower-order digits within each value on the higher-order digits that will appear in the released value. The resulting watermark is verified directly from released numerical values by ranking the owner key against independent decoy keys. We establish finite-sample false-positive control under key exchangeability, exact invariance to row reordering, and theoretical connections between numerical displacement, accessible band diversity, and watermark retention. Experiments on 1,455 units across five evaluation sets, together with a transfer study of 26 compatible checkpoints from six model families under the same band-width rule and verifier, show evidence retention under the nine registered release operations with small pooled CRPS changes relative to the original unwatermarked forecasts.

1 INTRODUCTION

Forecasting APIs increasingly support inventory, staffing, and financial planning (Amazon Web Services, 2026; Nixtla, 2026), producing numerical predictions that are stored, shared, and reused in downstream decision pipelines. Once these forecasts leave the provider, however, API authentication and server-side logs can no longer establish the origin of a particular released ensemble. This creates a need for provenance that travels with the forecast itself (Yuan et al., 2025). Unlike ordinary text, forecasts are consumed through their numerical values and predictive distributions: users may round values, reorder trajectories, subsample draws, or crop horizons while preserving the forecast's utility. A useful watermark must therefore survive legitimate numerical reuse without materially changing the predictive distribution that carries forecast uncertainty and supports downstream decisions.

We study this problem for digit-serialized probabilistic forecasts produced by frozen autoregressive LLMs (Gruver et al., 2023). Production inference servers already expose generation-time watermarking and logit-processing controls (Hugging Face, 2026), making conditional digit probabilities available for provider-side insertion without retraining. Yet existing token-level mechanisms face two difficulties: (1) Numerical-reuse experiments show that token identities and sequence positions are fragile carriers: rounding changes digit representations, while trajectory reordering changes token alignment without changing the ensemble distribution. (2) Controlled feedback interventions further

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

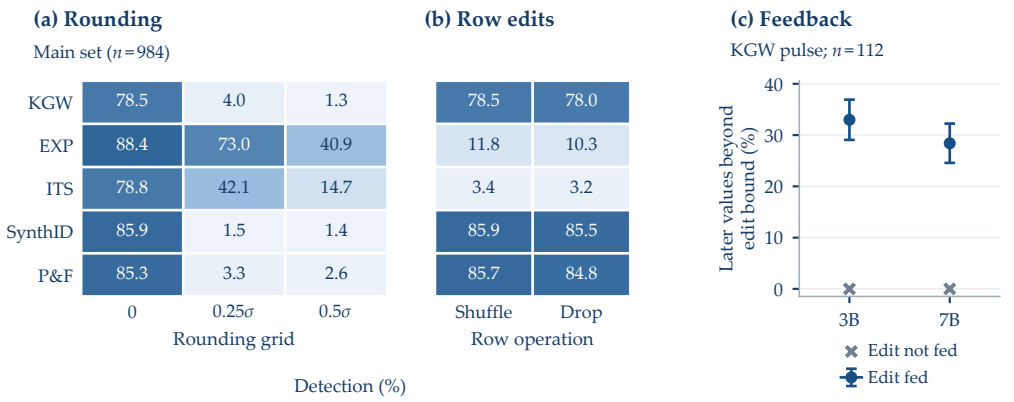

Figure 1: **Numerical reuse and forecast feedback.** (a) Detection under rounding for five token-watermark configurations on 984 main units at rank ≤ 40 among 4,096 keys; controls supply each step's σ, and EXP/ITS use aligned detectors. (b) Detection under row shuffling and 30% row removal using the same configurations, units, and detection protocol. (c) A single KGW units-digit edit on 112 windows per model: later values exceeding the initial nine-quanta edit bound, with the edit fed or withheld from model history and all later steering off. Error bars give 95% series-bootstrap intervals.

show that a local digit edit can propagate through the autoregressive history and alter many later predictions. Provenance must therefore remain attached to reusable numerical values while keeping watermark perturbations local to each individual forecast value throughout the forecast horizon.

Motivated by these findings, we propose V-band, a generation-time value-space watermark for digit-serialized probabilistic forecasts. To the best of our knowledge, V-band is the first such method for autoregressive LLM forecasts. To embed watermark evidence in numerical values, it partitions the numerical range into bands labelled by the owner's secret key and uses conditional digit probabilities to favour values in preferred bands. Value-feedback-free (VFF) steering keeps the unmarked trajectory as the predictive history, preventing local watermark edits from influencing later forecasts. Verification operates directly on released values and forecast steps, comparing the owner's aggregated score with those of independent decoy keys without querying the generator or accessing its weights.

Our contributions are fourfold: (1) Numerical-reuse experiments and controlled feedback interventions reveal two overlooked problems: evidence misalignment and propagation of local watermark edits through future predictions. (2) We propose V-band, which combines keyed numerical bands with value-feedback-free steering to create reusable numerical evidence while isolating watermark edits from future model inputs. (3) We prove finite-sample false-positive control under key exchangeability, exact row-order invariance, and horizon-independent distortion bounds, and characterize evidence creation and retention through numerical geometry. (4) We demonstrate strong detection and robustness across forecast datasets, model families and model sizes at low cost, with ablations explaining the design choices. This makes forecast provenance verifiable from released values.

2 Revisiting Watermarking for Numerical Forecasts

A forecast release $X \in \mathbb{R}^{B \times H}$ contains B sampled trajectories over H forecast steps from a shared context and origin, with rows indexing trajectories and columns indexing steps; we examine how token-based evidence and autoregressive insertion interact with numerical reuse and forecast fidelity.

2.1 Token Evidence under Numerical Reuse

Nearby values can have different digit strings: a decimal carry may change several digits after a small displacement. Token partitions need not respect this geometry, while position-aligned randomness ties evidence to sequence addresses. Rounding and row operations probe these distinct dependencies.

Figure 1a evaluates Qwen2.5-3B-Instruct (Qwen et al., 2024) on 984 units from 14 datasets, with $B = 64$ and $H \in \{24, 64, 128\}$. Rounding at 0.5σ reduces detection by KGW (Kirchenbauer

et al., 2023), SynthID-Text (Dathathri et al., 2024), and Permute-and-Flip (Zhao et al., 2025) from 78.5–85.9% to 1.3–2.6%. EXP and ITS (Kuditipudi et al., 2023) retain 40.9% and 14.7%, respectively.

Row operations expose the second dependence (Figure 1b). EXP and ITS fall to 11.8% and 3.4% under shuffling, and 10.3% and 3.2% after row removal, using their aligned detectors. KGW and SynthID retain their identity decisions under shuffling, so row sensitivity is configuration-dependent. Across all nine release operations, joint survival is at most 0.4% for these five configurations. Table 20 examines an alternative edit-distance detector.

Row exchangeability and numerical tolerance require separate checks. Reordering preserves the empirical forecast distribution exactly; rounding changes values and must be assessed against numerical displacement and scale. Full experimental settings are given in Appendices A, C, and I.3.

2.2 Local watermark edits propagate through forecast feedback

Autoregressive forecasts use previously sampled values as model inputs (Gruver et al., 2023). This creates a feedback path from local sampling decisions to later predictions (Ross et al., 2011).

We isolate this path with a KGW intervention (Kirchenbauer et al., 2023) at $\gamma = 0.5$, $\delta = 2$ on 112 windows from 14 datasets, using Qwen2.5-3B and 7B, 64 trajectories, and a 64-step horizon. The intervention changes only the units digit of the first forecast value, by at most nine serializer quanta; two branches publish the same edited value but feed either the original or the edited value to the model, with all later steps decoded unmarked under shared sampling randomness.

Feeding the edit moves 33.0% of later values on 3B and 28.4% on 7B beyond the nine-quanta bound relative to unmarked forecasts (Figure 1c). Mean future displacement is 97.5 and 3531.4 times the mean initial edit, respectively. Amplification occurs in 94.6% and 88.4% of windows. Retaining the original input preserves the subsequent trajectories. Feedback thus converts a one-digit insertion into a trajectory-level perturbation: released values move and dispersion grows, while the paired fair-CRPS difference between feeding and withholding the same published edit stays within sampling uncertainty (Appendix N). Feedback control therefore buys deterministic displacement and dispersion guarantees rather than a measured accuracy gain.

These observations motivate evidence aligned with numerical values and forecast steps, together with insertion that controls the temporal feedback path. The next section develops a construction around these requirements using keyed bands, value-feedback-free steering, and decoy-key verification.

3 Methodology

Forecasting-service setting. The provider controls digit sampling from a frozen autoregressive LLM, with access to conditional probabilities and forward passes. Insertion needs no parameter updates or gradient-based perturbation search and uses this model access before publication; VFF can evaluate unmarked control trajectories in a separate pass. Consumers verify using a candidate key, numerical values, and public metadata: context-derived scaler, serializer precision and digit width, band width, nonce, and original step indices. The construction is public; the owner key is secret. Reported robustness concerns specified reuse operations; V-band does not provide general protection against deliberate watermark removal. Figure 2 summarizes the numerical bands, VFF sender, and decoy-key verifier.

3.1 Keyed numerical bands

The observed context fixes an affine map $A(x) = (x - o)/s$, with $s > 0$. At serializer quantum $q = 10^{-3}$, a valid scaled value is represented by an integer V of declared width n in $[0, 10^n)$. The verifier re-serializes each released value to $V = \lfloor A(x)/q \rfloor$ and retains finite, nonnegative integers of emitted width exactly n; excluded cells are counted, with rates in Appendix A. Let σ_h be the standard deviation across scaled control rows at forecast step h. We choose the public integer band width:

$$\Delta = \mathrm{clip}\left(\mathrm{round}\left(0.40\,\frac{\mathrm{median}_h\,\sigma_h}{q}\right), 10, 90\right). \tag{1}$$

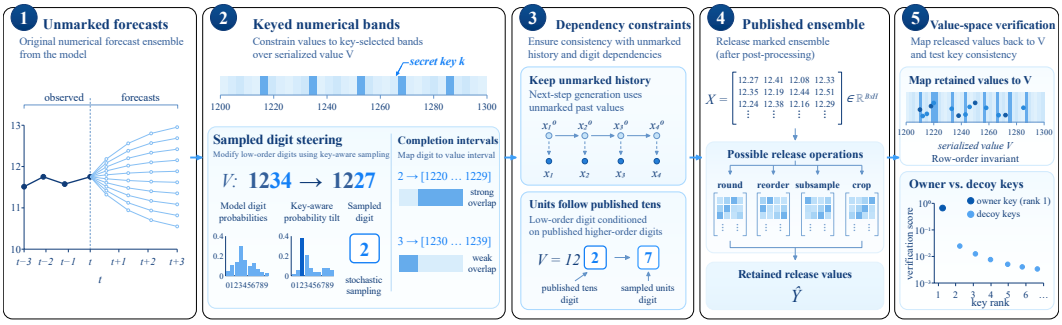

Figure 2: **V-band watermarking.** Keyed numerical bands guide digit sampling towards preferred value regions via probability tilting. VFF retains unmarked history across forecast steps while conditioning lower-order digits on the published prefix within each value. After numerical reuse, verification ranks the owner key against independent decoy keys using the retained numerical values.

For key k, a pseudorandom phase $\varphi_k(h) \in \{0, \ldots, \Delta - 1\}$ shifts the grid and a pseudorandom table $C_k(h, b) \in \{0, 1\}$ colours its bands. Both are keyed by the release nonce:

$$c_k(h, V) = C_k\left(h, \left\lfloor \frac{V + \varphi_k(h)}{\Delta} \right\rfloor\right). \tag{2}$$

The nonce separates releases, and the original step index separates forecast coordinates. Nearby values within a band share evidence. A single horizon-level Δ keeps the metadata to one integer, at the cost of fewer occupied bands at early, low-spread steps (Section 4). Row indices never enter Equation 2; cropping robustness instead requires the retained original step indices, since re-indexed steps desynchronize φ_k and C_k.

3.2 Value-feedback-free digit steering

At decimal place $p = 10^j$, let P be the fixed higher-order prefix. Choosing digit d admits integer completions $I_d = \{P + dp, \ldots, P + (d + 1)p - 1\}$. For owner key 0, the sender computes each candidate's green completion fraction and adds its centred value to the model logit ℓ_d:

$$w_d = \frac{1}{p} \sum_{v \in I_d} c_0(h, v), \qquad \widetilde{\ell}_d = \ell_d + \lambda_j(2w_d - 1). \tag{3}$$

We set $\lambda_j = \lambda$ for $10^j \leq \Delta$ and zero at coarser places; because Equation 1 caps $\Delta \leq 90$, the steered places are the tens and units digits, $j \in \{0, 1\}$, and higher digits are copied. Units weights are binary; tens weights can be fractional across band boundaries. After tilting, we apply temperature $T = 0.7$ and nucleus probability 0.9 over allowed tokens, then renormalize surviving digits. When the weights are equal or the surviving digit mass is at most 10^{-9}, the sender retains the control digit.

For the example in Figure 2, an owner-green band $[1223, 1233)$ gives tens candidates 2 and 3 green fractions 0.7 and 0.3; sampling tens 2 and then units 7 releases 1227, while the unmarked 1234 remains in future history (Appendix B.9).

VFF first generates X^0 and teacher-forces each control row to obtain tens conditionals. Let $Z_{r,h}$ contain the observed context, preceding unmarked values, and copied higher digits. Writing $\mathcal{T}_{h,j}$ for Equation 3 and its sampling rule at place 10^j, the published digits follow:

$$d^{\text{pub}}_{r,h,1} \sim \mathcal{T}_{h,1}[p_\theta(\cdot \mid Z_{r,h})], \tag{4}$$

$$d^{\text{pub}}_{r,h,0} \sim \mathcal{T}_{h,0}\left[p_\theta(\cdot \mid Z_{r,h}, d^{\text{pub}}_{r,h,1})\right]. \tag{5}$$

Here p_θ is the model's digit law. A local forward conditions units on the published tens digit while masking the original tens and later tokens. Batched updates share the teacher-forced cache; the next step again uses unmarked history. VFF is thus a two-pass sender: it needs the saved control ensemble and masked edits of a standard causal attention cache, so it runs as a provider-side insertion pass rather than a streaming logits processor, and models without maskable caches fall back to in-loop FF (Section 5.3). Algorithm 1 gives the sender and verifier, with full cache operations in Appendix B.

Algorithm 1 V-band embedding with VFF and verification from released values

Require: Model p_θ, context, owner key k_0, nonce, B rows, horizon H, band width Δ, declared width n, strength λ, independent decoy keys $k_1, \ldots, k_K$, and rank threshold t

1: Serialize X^0 to V^0; set $V^{\mathrm{pub}} \leftarrow V^0$ and build $G_h(v) = \sum_{u<v} c_0(h, u)$.
2: Teacher-force each control row once; retain its logits and attention cache.
3: **for** each eligible cell (r, h) **do**
4: Set $P = 100\lfloor V^0_{r,h}/100 \rfloor$; read control tens logits ℓ_1 with history $X^0_{r,<h}$.
5: Sample published tens d_1 using $\textsc{Steer}(\ell_1, h, P, 10, d_1^0)$.
6: Obtain units logits ℓ_0 with published d_1 and a mask hiding the original tens and later tokens.
7: Sample published units d_0 using $\textsc{Steer}(\ell_0, h, P + 10d_1, 1, d_0^0)$; set $V^{\mathrm{pub}}_{r,h} = P + 10d_1 + d_0$.
8: **end for**
9: Publish $X^{\mathrm{pub}} = A^{-1}(A(X^0) + q(V^{\mathrm{pub}} - V^0))$ and public metadata.
10: On retained release values, compute S_k for owner and decoys by Equation 6; return $\rho \le t$.

3.3 Verification with decoy keys

For the valid released cells $\mathcal{I}$, verification applies the same colour map to the owner and $K = 4{,}095$ independent decoy keys. It sums the colours and computes a conservative owner rank:

$$S_k = \sum_{(r,h)\in\mathcal{I}} c_k(h, V_{r,h}), \qquad \rho = 1 + \sum_{k=1}^{K} \mathbf{1}\{S_k \ge S_0\}. \tag{6}$$

The default decision $\rho \le 40$ has nominal level $40/4096 = 0.977\%$; $\rho \le 4$ gives $4/4096 \approx 0.098\%$. Ties count against the owner, and verification needs no model queries. To accommodate public decimal unit changes, the verifier can test the five conversions 10^j, $j \in \{-2, -1, 0, 1, 2\}$, at threshold 8; the union bound then preserves the nominal $40/4096$ level (Appendix L). General continuous affine rescaling lies outside this discrete search family.

4 Theoretical analysis

Our analysis links the verifier's calibration, the carrier's numerical geometry, and VFF's control of forecast distortion. Appendix B gives proofs, sampling-law extensions, and implementation details.

Theorem 1 (Finite-sample null validity)**.** *If a null release and its public metadata are independent of the evaluated key functions, and these functions are independent and identically distributed in the ideal pseudorandom-function model, then*

$$\Pr(\rho \le t \mid X, \textit{metadata}) \le \frac{t}{K+1}. \tag{7}$$

This bound permits arbitrary dependence among forecast values.

Key exchangeability gives conservative ranks. Row permutations leave scores unchanged; key-independent selection preserves null validity with original step indices (Theorem 4).

When the numerical carrier survives. Write $x = (h, V)$ and let $\kappa(x, z) = \mathbb{E}_k[(2c_k(x) - 1)(2c_k(z) - 1)]$ measure shared keyed evidence. Independent band colours and a uniform integer phase give

$$\kappa\big((h, V), (h', V')\big) = \mathbf{1}\{h = h'\} \left(1 - \frac{|V - V'|}{\Delta}\right)_+. \tag{8}$$

Thus a small numerical change preserves part of the carrier even when its digit representation changes. To isolate this effect, consider a key-independent candidate law π after temperature scaling, an untruncated tilt $\pi_{k,a}(x) \propto \pi(x)e^{a(2c_k(x)-1)}$ with $a = \lambda/T$, and a key-independent release operation $Q(z \mid x)$.

Theorem 2 (Surviving evidence under local steering)**.** *Let X, X' be independent draws from π and $Z \sim Q(\cdot \mid X)$. If $\mu_Q(a)$ denotes the expected owner-green fraction after tilting and release, minus $1/2$, then*

$$\mu'_Q(0) = \tfrac{1}{2}\left\{\mathbb{E}\kappa(X, Z) - \mathbb{E}\kappa(X', Z)\right\}. \tag{9}$$

Moreover, $|\mu_Q(a) - a\mu'_Q(0)| \le a^2/2$.

The first term measures survival of a candidate's own carrier; the second measures overlap with unrelated candidates. Fresh, disjoint token-key addresses remove both terms. Numerical bands retain the first term under small displacement, while diverse accessible bands keep the second small. This gives a conditional mechanism for the aligned-detector failures in Section 2. It also explains the digit-scale ablation: for any units-only sender whose copied prefix is key independent, $\Delta \geq 10$ implies expected green excess at most $9/(4\Delta)$, regardless of steering strength (Theorem 7). Theorem 2 itself is a weak-tilt certificate for a single untruncated tilt; at the operational strengths $a \approx 4.3$–11.4, behaviour is instead governed by saturation over the accessible bands, and Appendix B analyses the implemented two-stage, nucleus-truncated sampling law.

Why VFF prevents accumulated edits. VFF retains the unmarked trajectory as the predictive context for subsequent forecast steps, so watermark-induced edits do not enter future forecast states. Under the default sender, only the final two digits of each value are editable.

Proposition 3 (Bounded distortion under VFF)**.** *Under VFF with the default two-digit editing schedule, watermark-induced changes do not propagate through future forecast states, and each released value satisfies* $|V^{\mathrm{pub}}_{r,h} - V^{0}_{r,h}| \leq 99$. *With scaler* $A(x) = (x - o)/s$, *the corresponding physical displacement is therefore at most* $99sq$. *For finite, equally sized, row-paired ensembles, the implemented fair CRPS satisfies*

$$\left|\mathrm{CRPS}_{\mathrm{f}}(X^{\mathrm{pub}}, y) - \mathrm{CRPS}_{\mathrm{f}}(X^{0}, y)\right| \leq \frac{2}{BH} \sum_{r,h} |x^{\mathrm{pub}}_{r,h} - x^{0}_{r,h}|. \tag{10}$$

The bound is absolute: with the default schedule the mean fair-CRPS change is at most $198sq = 0.198s$ at any horizon. Relative percentages remain sensitive to baseline sharpness, since a bounded absolute change can dominate a near-zero control score; we therefore report pooled and per-unit utility summaries together (Section 5.2).

5 Experiments

5.1 Experimental setup

The main comparison uses Qwen2.5-3B-Instruct (Qwen et al., 2024) on 984 units, with 64 trajectories per unit. Its 14 datasets comprise four ETT series (Zhou et al., 2021), Electricity (Trindade, 2015), Exchange Rate (Lai et al., 2018), Traffic (California Department of Transportation, n.d.), Weather (Max Planck Institute for Biogeochemistry, n.d.), and six M4 frequencies (Makridakis et al., 2020). Four additional sets, covering finance/M4, ETT, four long-sequence benchmarks (LSF4), and Monash/Darts (Godahewa et al., 2021; Herzen et al., 2022), use 128 trajectories per unit and bring the total to 1,455 units. Each unit has context length $2H$ and horizon $H \in \{24, 64, 128\}$. The band-width rule in Equation 1 and default $\lambda = 5$ are shared. Appendix A.2 lists all dataset sources.

The main baselines are KGW (Kirchenbauer et al., 2023), EXP and ITS (Kuditipudi et al., 2023), SynthID-Text (Dathathri et al., 2024), and Permute-and-Flip (Zhao et al., 2025); EXP and ITS use aligned detectors. Extended comparisons include DiPmark (Wu et al., 2024), Unigram (Zhao et al., 2023), post-hoc value-space methods, and fixed-origin edit-distance detection. Appendices C and H give configurations, sources, and denominators.

The registered numerical reuse protocol comprises identity, row sorting, row shuffling, rounding at 0.25σ and 0.5σ, Gaussian noise at 0.05σ and 0.1σ, removal of 30% of rows, and cropping of 30% of the horizon. Operations are fixed without optimizing against watermark evidence. Here σ is the control ensemble's stepwise standard deviation, shared across methods when setting transformation amplitudes. Each operation is applied separately to the same release; joint survival requires all nine single-operation decisions on a unit, so it reports the worst registered operation, not a composite chain. Appendix A gives the full protocol.

Detection uses owner-rank thresholds 40 and 4. Appendix J.3 explains why repeated band evidence can invalidate cellwise z-score calibration. Forecast quality is measured by fair CRPS against future outcomes (Ferro, 2014) and 10–90% interval coverage. Our primary utility statistic pools the scores:

$$U = 100 \left(\frac{\sum_i \mathrm{CRPS}(X_i, y_i)}{\sum_i \mathrm{CRPS}(X^0_i, y_i)} - 1 \right) \%, \tag{11}$$

Table 1: **Numerical-forecast watermarking on the 984-unit main set.** (a) Aggregate detection and utility. (b) Dataset-level identity detection and V-band joint survival and utility. I_t: identity at rank t; Q_a/G_a: rounding/noise; Joint: all nine transformations. Detection and pooled CRPS change are percentages; Med. is the median per-unit relative CRPS change; coverage changes (ΔCov.) are percentage points. Dataset sources: Appendix A.2.

(a) Aggregate comparison on the 984-unit main set

Method	$I_{40}\uparrow$	$I_4\uparrow$	$Q_{.25}\uparrow$	$Q_{.5}\uparrow$	$G_{.1}\uparrow$	Drop↑	Crop↑	Joint↑	CRPS↓	Med.↓	ΔCov.
V-band $\lambda = 5$	**99.6**	**98.8**	**94.6**	**76.1**	**94.1**	**99.5**	**99.1**	**75.9**	+0.26	+0.39	+0.22
V-band $\lambda = 4$	98.6	97.2	91.8	72.3	90.7	98.4	98.4	72.0	+0.27	+0.28	+0.21
V-band $\lambda = 3$	94.4	90.7	81.6	61.4	81.7	94.4	94.2	60.8	+0.19	+0.17	+0.14
KGW	78.5	75.4	4.0	1.3	7.5	78.0	79.2	0.4	+0.51	+0.64	-0.90
EXP	88.4	87.1	73.0	40.9	72.3	10.3	88.6	0.2	-0.55	+0.00	-0.13
ITS	78.8	76.7	42.1	14.7	50.8	3.2	81.4	0.1	+0.48	+0.00	+0.35
SynthID-Text	85.9	83.5	1.5	1.4	1.5	85.5	85.5	0.0	-0.97	-0.53	+2.09
PF	85.3	83.2	3.3	2.6	1.9	84.8	85.6	0.2	+0.36	+0.48	-2.62

(b) Dataset-level comparison at rank threshold 40

Dataset	n	KGW	EXP	ITS	SynthID	PF	V-band	Joint↑	CRPS↓	Med.↓	ΔCov.
ETTh1	72	81.9	**100.0**	94.4	**100.0**	**100.0**	**100.0**	79.2	-0.18	+0.33	+0.02
ETTh2	72	83.3	83.3	61.1	81.9	80.6	**98.6**	79.2	+0.37	+0.52	+0.89
ETTm1	72	63.9	97.2	79.2	88.9	88.9	**100.0**	87.5	-0.06	+0.23	-0.42
ETTm2	72	69.4	79.2	62.5	72.2	69.4	**98.6**	81.9	+0.99	+0.90	-0.60
Electricity	72	90.3	95.8	95.8	95.8	94.4	**100.0**	66.7	-0.04	+0.35	+0.98
Exchange rate	72	94.4	**100.0**	90.3	**100.0**	95.8	**100.0**	91.7	+0.39	+0.48	+0.93
M4 daily	72	83.3	94.4	90.3	93.1	94.4	**100.0**	81.9	+0.40	+0.27	-0.41
M4 hourly	72	94.4	93.1	87.5	91.7	91.7	**100.0**	65.3	+0.17	+0.76	-0.29
M4 monthly	72	69.4	87.5	72.2	80.6	80.6	**100.0**	73.6	+0.19	+0.20	+0.12
M4 quarterly	72	54.2	62.5	54.2	56.9	56.9	**98.6**	76.4	+0.30	+0.23	-0.41
M4 weekly	72	80.6	88.9	80.6	91.7	90.3	**100.0**	84.7	+0.34	+0.04	+0.39
M4 yearly	48	45.8	54.2	45.8	50.0	50.0	**100.0**	72.9	+0.59	+0.70	+0.95
Traffic	72	97.2	**98.6**	97.2	**98.6**	**98.6**	**98.6**	45.8	+0.41	+0.46	+0.02
Weather	72	79.2	91.7	80.6	88.9	90.3	**100.0**	75.0	+0.45	+0.23	+1.37

where each score averages over forecast steps. Thus U measures the change in the total forecasting score. Utility intervals use 2,000 series-clustered bootstrap draws, and coverage changes are reported in percentage points.

5.2 Detection and forecast quality

V-band retains evidence through numerical processing, with 75.9% joint survival compared with at most 0.4% for the tested token configurations (Table 1a). Identity detection is 99.6%, with Wilson 95% interval [99.0,99.8]%. Coarse rounding accounts for most of the remaining loss: detection after 0.5σ rounding is 76.1%, close to the joint rate. Traffic shows this failure mode at its strongest: the 90-quantum cap binds on 62% of its units, the median 0.5σ rounding grid reaches 125 quanta, 1.4 times the band width, and joint survival drops to 45.8% (Appendix M). Reordering preserves every identity rank (984/984 units); row removal and cropping each retain over 99% detection. V-band also matches or exceeds the tested token configurations' identity detection on all 14 datasets (Table 1b). Pooled CRPS increases by +0.26% [95% CI: +0.10,+0.44], and coverage changes by +0.22 points [−0.09,+0.54]. The pooled ratio weights each unit by its control score; per-unit summaries complete the picture: median +0.39%, 5%-trimmed mean +0.64%, and unweighted mean +7.41%, falling to +2.55% once 23 units with control CRPS below 10^{-3} are excluded. Series with small absolute error scales thus carry larger relative changes than the pooled figure alone suggests (Appendix I).

Value-space comparison. On the main set, WGTD reaches 89.7% joint survival at +2.26% pooled CRPS, the Ngo et al. adaptation reaches 64.4% at +0.57%, and a bounded-noise stand-in for Waltz reaches 76.8% at +1.82%. That stand-in applies unguided uniform perturbations rather than the encoder-guided optimization of the published Waltz method, so it measures a perturbation budget, not that method's frontier (Appendix H). V-band offers a favourable measured tradeoff, though the ordering varies by dataset (ETT WGTD: 90.5% at −0.7% CRPS); these are measured configurations, not a matched-distortion comparison.

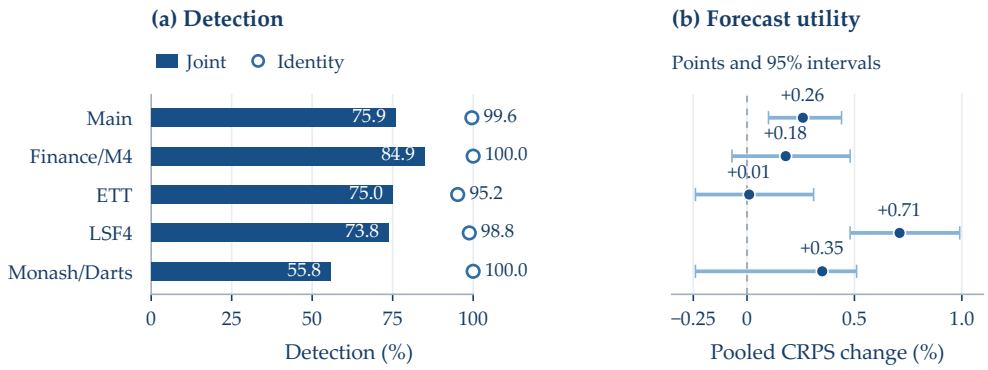

Figure 3: **Five sets under one rule.** (a) Identity detection and joint survival across all nine transformations. (b) Pooled CRPS changes with 95% series-bootstrap intervals. Both panels use the same dataset rows at $\lambda = 5$.

5.3 Transfer across datasets and generators

The shared rule maintains high identity detection across the four additional sets, while survival after processing varies more widely (Figure 3). Identity detection is 95.2–100%, joint survival is 55.8–84.9%, and pooled CRPS changes are +0.01–+0.71%. The confirmation sets publish 128 rows per unit against 64 on the main set, so cross-set detection comparisons involve different amounts of released evidence (Theorem 8). Monash/Darts illustrates the distinction between identity capacity and reuse survival: every unmodified release is detected, yet joint survival is lowest.

The transfer study evaluates 26 checkpoints spanning 0.5B–32B parameters: Qwen2.5 (Qwen et al., 2024), Qwen3 (Yang et al., 2025), Qwen3.5 (Qwen Team, 2026), Gemma 3 (Gemma Team, 2025), Llama 3 (Grattafiori et al., 2024; Meta, 2024), and Mistral (Jiang et al., 2023; Mistral AI, 2024; 2025). On the 199-unit finance/M4 set, identity detection is 99.0–100%, joint survival is 57.3–95.0%, and pooled CRPS change is -0.05% to +0.80%; identity detection remains at least 96.5% at the stricter rank-4 threshold (Table 9). Figure 5 shows variation within each family. Twenty-two checkpoints use VFF; the four Qwen3.5 checkpoints fall back to independently sampled in-loop FF for their hybrid caches, which costs 0.75–1.0 extra generations against the 0.22–0.23 VFF pass and leaves their ensembles outside the row-paired bound of Proposition 3, widening their utility intervals. Appendices D and K give the exact releases, full checkpoint results, and numerical-format checks.

5.4 Ablations of V-band design choices

Sender conditioning. On paired dev-64 at $\lambda = 1.5$, coupled, FF, and VFF have mean per-unit CRPS changes of +1.68%, +0.38%, and +0.37%, with paired differences including zero; coupled generation has higher detection and greater dispersion (Figure 7). FF and VFF share identity and rounding detection but differ in local conditioning law (Appendix I.3).

Digit schedule. More editable digits improve detection at a forecasting cost. On dev-64 at $\lambda = 3$, units-only steering gives 4.7% joint survival, versus 75.0% for the keyed, band-capped schedule. Editing all digits raises survival to 89.1%, but median per-unit CRPS change reaches +9.31%. Appendix I reports the width, coarse-digit, and alternative-schedule comparisons.

Band phase. Keyed phases raise dev-64 rounding-0.5σ detection from 67.2% to 75.0%, with identity detection at 96.9% for both configurations (Figure 6b).

Steering strength. Stronger steering yields diminishing detection gains (Figure 6a). Increasing λ from 3 to 5 raises main-set joint survival from 60.8% to 75.9%, versus 77.5% at $\lambda = 8$. Pooled CRPS changes remain 0.19–0.31% across this sweep. These measurements support $\lambda = 5$ near the detection knee (Appendices E and I).

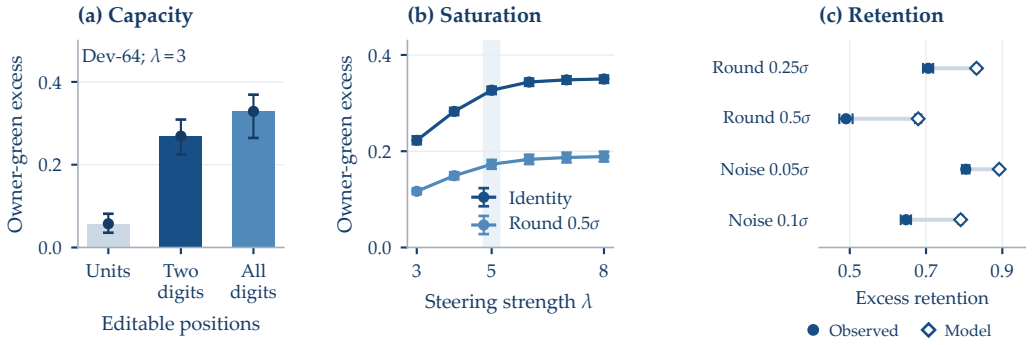

Figure 4: **How evidence is created and retained.** (a) Owner-green excess across digit schedules on dev-64 at $\lambda = 3$. (b) Identity and rounding excess across steering strengths on 984 main units; shading marks $\lambda = 5$. (c) Paired observed and predicted mean retention on 983 units with identity excess ≥ 0.05. Error bars show 95% intervals.

5.5 Mechanism Analysis: Evidence Creation and Retention

To distinguish evidence creation from retention even when detection saturates, we use owner-green excess, the scored green fraction minus $1/2$: identity excess measures creation, and the transformed-to-identity excess ratio measures retention.

Evidence creation depends on accessible numerical alternatives. On paired dev-64, units-only steering produces mean owner-green excess 0.057, compared with 0.269 for two-digit steering and 0.329 when all digits are editable (Figure 4a). Units-only edits span at most nine quanta and often remain in one keyed band, whereas tens-digit edits expose alternatives with different colours, as Theorem 2 requires; all-digit editing buys further capacity at higher utility cost.

Evidence creation saturates with steering strength. Main-set identity excess rises from 0.223 to 0.327 as λ increases from 3 to 5, but reaches only 0.350 at $\lambda = 8$ (Figure 4b): detection saturation reflects diminishing signal gains.

Evidence retention depends on displacement relative to band width. The conditional model in Equation 36 correctly orders mean retention under all four perturbations (Figure 4c), with across-unit correlations of 0.69–0.78. It overestimates the magnitude: rounding at 0.5σ gives predicted retention 0.679 versus observed 0.490 [95% CI: $0.473, 0.507$] (Appendix J).

5.6 Computational Cost

VFF adds 0.22–0.23 times the unmarked generation cost on the measured 3B–7B models (1.22–1.23 total including the control ensemble); token watermarks embed in a single pass at 1.05–1.15, and in-loop FF costs 1.75–2.0. On one pinned CPU core, verifying a 64-row, 128-step release against 4,096 keys takes 5.1 seconds including key-table construction. On the same block, Unigram takes 4.2 seconds, aligned EXP and ITS take 15.3 and 15.9, and SynthID-Text, KGW, Permute-and-Flip, and DiPmark take 89.4–1108.3 (Figure 6c); Appendix F reports the complete two-block grid.

6 Conclusion

We studied provenance for probabilistic forecasts from autoregressive language models, where token-level watermarks are poorly aligned with numerical outputs that are transformed and reused. V-band addresses this mismatch with keyed numerical bands, value-feedback-free steering, and value-space verification, with finite-sample false-positive control and bounded distortion. Across five evaluation sets and 26 compatible checkpoints, it retains substantial evidence under the registered reuse operations at small forecasting cost. Our mechanism analysis shows that effective watermarking requires accessible numerical alternatives at embedding and evidence retention under displacement: watermark carriers should align with the objects users manipulate, not their serialization.

AI USE STATEMENT

Generative AI assisted manuscript organization, English drafting, figure preparation, development and checking of theoretical derivations and pseudocode, and consistency checks between prose and the reported results.

REPRODUCIBILITY STATEMENT

Sections 3 and 5 give the construction and evaluation. Appendices A–N provide dataset and model sources, protocols, proofs, implementation details, full comparator and checkpoint results, ablations, timing measurements, public scale recovery, and distributional analyses. Supplementary code and data support reproduction of the figures and statistical analyses.

REFERENCES

Amazon Web Services. What Is Amazon Forecast? Amazon Forecast Developer Guide, 2026. URL https://docs.aws.amazon.com/forecast/latest/dg/what-is-forecast.html. Accessed 16 September 2026.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the Language of Time Series, 2024. URL https://arxiv.org/abs/2403.07815v3.

Binance. Binance Public Data. Historical market data, n.d. URL https://github.com/binance/binance-public-data.

California Department of Transportation. Performance Measurement System (PeMS). Traffic data service, n.d. URL https://pems.dot.ca.gov/.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. LLM4TS: Aligning Pre-Trained LLMs as Data-Efficient Time-Series Forecasters, 2023. URL https://arxiv.org/abs/2308.08469v6.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable Watermarks for Language Models. In *Conference on Learning Theory*, 2024. URL https://arxiv.org/abs/2306.09194v1.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2023. URL https://arxiv.org/abs/2310.10688v4.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634:818–823, 2024. doi: 10.1038/s41586-024-08025-4. URL https://www.nature.com/articles/s41586-024-08025-4.

Federal Reserve Bank of St. Louis. FRED: Federal Reserve Economic Data. Economic data service, n.d. URL https://fred.stlouisfed.org/.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923, 2014. doi: 10.1002/qj.2270. URL https://doi.org/10.1002/qj.2270.

Gemma Team. Gemma 3 Technical Report, 2025. URL https://arxiv.org/abs/2503.19786v1.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Discovering Spoofing Attempts on Language Model Watermarks, 2024. URL https://arxiv.org/abs/2410.02693v2.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102:359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/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*, 69:243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. URL https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash Time Series Forecasting Archive. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2105.06643v1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783v3.

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*, 2023. URL https://arxiv.org/abs/2310.07820v3.

Hengzhi He, Peiyu Yu, Junpeng Ren, Ying Nian Wu, and Guang Cheng. Watermarking Generative Tabular Data, 2024. URL https://arxiv.org/abs/2405.14018.

Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, et al. Darts: User-Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning Research*, 23:1–6, 2022. URL https://arxiv.org/abs/2110.03224v3.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased Watermark for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.10669v2.

Hugging Face. Text Generation Inference. Official software documentation, 2026. URL https://github.com/huggingface/text-generation-inference. Accessed 16 September 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825v1.

Ming Jin, Shiyu Wang, Lintao Ma, et al. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01728v2.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.19361v2.

Dongbin Kim, Geonwoo Shin, Yujin Choi, Soyeon Park, and Jaewook Lee. A Locally Tokenized Generative Model for Robust Time-Series Watermarking, 2026. URL https://arxiv.org/abs/2608.19727.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2301.10226v4.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the Reliability of Watermarks for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.04634v4.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust Distortion-free Watermarks for Language Models, 2023. URL https://arxiv.org/abs/2307.15593v3.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018. URL https://arxiv.org/abs/1703.07015v3.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36:54–74, 2020. doi: 10.1016/j.ijforecast.2019.04.014. URL https://doi.org/10.1016/j.ijforecast.2019.04.014.

Max Planck Institute for Biogeochemistry. Weather Station at the Max Planck Institute for Biogeochemistry, Jena. Jena climate data archive, n.d. URL https://www.bgc-jena.mpg.de/wetter/.

Meta. Llama 3.2 Model Card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Mistral AI. Ministral-8B-Instruct-2410 Model Card, 2024. URL https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.

Mistral AI. Mistral-Small-24B-Instruct-2501 Model Card, 2025. URL https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501.

Dung Daniel Ngo, Archan Ray, Akshay Seshadri, Daniel Scott, Saheed Obitayo, Niraj Kumar, Vamsi K. Potluru, Marco Pistoia, and Manuela Veloso. Adaptive and Robust Watermark for Generative Tabular Data. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4796–4831. PMLR, 2026. URL https://proceedings.mlr.press/v337/ngo26a.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730v2.

Nixtla. TimeGPT: Forecasting API and Pretrained Time-Series Model. Official software documentation, 2026. URL https://github.com/Nixtla/nixtla. Accessed 16 September 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL https://arxiv.org/abs/2508.10925.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 Technical Report, 2024. URL https://arxiv.org/abs/2412.15115v2.

Qwen Team. Qwen3-4B-Instruct-2507 Model Card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, et al. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting, 2023. URL https://arxiv.org/abs/2310.08278v3.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated Text be Reliably Detected?, 2023. URL https://arxiv.org/abs/2303.11156v4.

Zhi Wen Soi, Chaoyi Zhu, Fouad Abiad, Aditya Shankar, Jeroen M. Galjaard, Huijuan Wang, and Lydia Y. Chen. TimeWak: Temporal Chained-Hashing Watermark for Time Series Data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=akhhwQh6UV.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C58C86.

Unit8. Darts Dataset Loaders and Source References. Dataset documentation, n.d. URL https://unit8co.github.io/darts/generated_api/darts.datasets.html.

N. J. I. van Schaik. Time's Up! Robust Watermarking in Large Language Models for Time Series Generation. Bachelor thesis, Delft University of Technology, 2024. URL https://resolver.tudelft.nl/uuid:dbbf9d34-08ba-43d6-8ec4-0240ec00a2b7.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting, 2021. URL https://arxiv.org/abs/2106.13008v5.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A Resilient and Accessible Distribution-Preserving Watermark for Large Language Models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.07710v2.

Hao Xue and Flora D. Salim. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting, 2022. URL https://arxiv.org/abs/2210.08964v5.

Yahoo Finance. Apple Inc. (AAPL) Historical Data. Historical daily equity prices, n.d. URL https://finance.yahoo.com/quote/AAPL/history/.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388v1.

Wei Yuan, Chaoqun Yang, Yu Xing, Tong Chen, Nguyen Quoc Viet Hung, and Hongzhi Yin. Watermarking Large Language Model-based Time Series Forecasting, 2025. URL https://arxiv.org/abs/2507.20762.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable Robust Watermarking for AI-Generated Text, 2023. URL https://arxiv.org/abs/2306.17439v2.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2402.05864v3.

Yihao Zheng, Haocheng Xia, Junyuan Pang, Jinfei Liu, Kui Ren, Lingyang Chu, Yang Cao, and Li Xiong. TabularMark: Watermarking Tabular Datasets for Machine Learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3570–3584, 2024. doi: 10.1145/3658644.3690373. URL https://doi.org/10.1145/3658644.3690373.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2012.07436v3.

Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.11939v6.

Chaoyi Zhu, Jiayi Tang, Jeroen M. Galjaard, Pin-Yu Chen, Robert Birke, Cornelis Bos, and Lydia Y. Chen. TabWak: A Watermark for Tabular Diffusion Models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=71pur4y8gs.

CONTENTS OF THE APPENDIX

A Forecast and transformation protocol **15**

A.1 Evaluation units and serialization . . . 15

A.2 Dataset sources . . . 16

A.3 Registered transformations and utility . . . 16

A.4 Results by evaluation set and dataset . . . 17

B Formal guarantees and the mechanism of numerical evidence **18**

B.1 Assumptions and the implemented interfaces . . . 18

B.2 Finite-sample calibration and release selection . . . 18

B.3 Carrier covariance and the effect of a release operation . . . 19

B.4 Accessible bands, correlated colours, and the units-only limit . . . 21

B.5 Soft completion weights and the actual sampling law . . . 22

B.6 Perturbation retention and its empirical interpretation . . . 23

B.7 Repeated bands and effective evidence . . . 24

B.8 VFF displacement, forecast-score stability, and feedback . . . 25

B.9 Worked digit-steering example . . . 26

B.10 Implementation-level pseudocode . . . 26

C Generation-time comparator protocols **27**

D Transfer across 26 generator checkpoints **27**

D.1 Checkpoint identities and release references . . . 27

E Steering strength and local design **30**

F Embedding and verification efficiency **31**

G Related work **31**

H Supplementary comparator results **33**

H.1 Post-hoc value-space comparators . . . 33

H.2 Paired identity comparisons and diagnostic subsets . . . 34

H.3 Value-space operating points on additional sets . . . 35

H.4 Edit-distance detection for EXP and ITS . . . 35

I Component ablations and utility heterogeneity **36**

I.1 Operating-point selection and aggregation . . . 36

I.2 Separating temporal feedback and local conditioning . . . 37

I.3 Controlled sender comparison . . . 37

I.4 Digit schedule, phase, and width . . . 39

I.5 Evidence quantity and common perturbation scale . . . 40

J Empirical tests of the local mechanism **40**

J.1 Control NLL and evidence creation 41
J.2 Numerical perturbations and retained evidence 41
J.3 Why cellwise z-scores are not calibrated 42
J.4 Null observations across evaluation sets 42

K Generator diagnostics and additional transfer experiments **42**
K.1 Transfer beyond finance/M4 43
K.2 Numerical-format compatibility 43

L Public scale recovery **44**
L.1 Public scale search and its null observations 44

M Distributional and paired views of the experiments **45**
M.1 Numerical reuse across datasets and detection thresholds 45
M.2 Reuse survival and forecast quality 45
M.3 Paired sender effects 45

N Controlled experiment on forecast feedback **45**

A Forecast and transformation protocol

A.1 Evaluation units and serialization

The main set contains 984 forecast units, each with a context of length $2H$ and a future of length H. Its horizon counts are 336 at $H = 24$, 336 at $H = 64$, and 312 at $H = 128$. Thirteen datasets contribute 24 units at each horizon; M4 yearly contributes 24 at each of the first two horizons. The long multivariate datasets use nonoverlapping windows of length $3H$ in the final 20% of each series. The M4 units use a series' final $3H$ observations. Eligible windows have finite values and a nonconstant context. Series and channels remain distinct when constructing the bootstrap clusters.

The finance/M4 confirmation set contains 199 units. Selection requests eight units for each dataset–horizon combination, orders eligible candidates deterministically by a hash, and cycles through horizons 128, 64, and 24 while excluding overlapping windows. Eight dataset groups contribute 24 units each, while Apple and TSMC contribute five and two. The ETT and LSF4 confirmation sets each contain 84 units, and Monash/Darts contains 104 units from 25 datasets. The main set publishes 64 sampled rows per unit; the four additional sets publish 128.

For context $z = (z_1, \ldots, z_{2H})$, the affine serializer uses

$$o = \min z - 0.3(\max z - \min z), \qquad s = Q_{0.95}(z) - o, \qquad A(x) = \frac{x - o}{s}, \tag{12}$$

with $s = 1$ when the calculated scale is zero. The serializer's precision is $q = 10^{-3}$ in scaled units. Precision fixes the quantum. The declared integer digit width n is a separate unit-level quantity: each unit declares the modal emitted digit width of its control ensemble, and generation requires that mode to cover at least half of the emitted values. Verification re-serializes every released value with the same fixed-point rule, $V = \lfloor A(x)/q \rfloor$, and retains only finite, nonnegative serialized integers of width exactly n; excluded cells are counted and never repaired. Across the nine registered operations, the mean per-unit exclusion rate is 9.7% on the main set and 8.6–11.9% on the four confirmation sets; identity means are 8.4–11.8%, Gaussian noise at 0.1σ raises them by 0.4–1.2 points, and unit-level maxima reach about 50%. Public metadata carries the affine transformation, n, Δ, the nonce, and the original forecast-step indices. Consequently, a consumer can verify a release without access to the control ensemble used to set Δ.

The main generator uses temperature 0.7 and top-p 0.9. The population standard deviation across control rows is calculated separately at each step; the median of these standard deviations sets Δ through Equation 1. All comparisons use the same control-derived perturbation scale for a given unit.

The preliminary schedule and phase ablations use a 64-unit development set, denoted dev-64, with 128 forecast rows per unit, public-label development keys, and 4,095 decoys. These experiments compare local construction choices before the larger evaluation sets. Tables below identify the strength and setting of each retained ablation.

A.2 Dataset sources

Tables 2–4 identify the datasets used in the five evaluation sets. The main set combines the ETT benchmark (Zhou et al., 2021), four long-sequence datasets also used in forecasting benchmarks (Lai et al., 2018; Wu et al., 2021), and the M4 competition (Makridakis et al., 2020). The additional archive set contains 17 Monash datasets (Godahewa et al., 2021) and eight Darts examples (Herzen et al., 2022; Unit8, n.d.). Archive record links distinguish individual datasets within those collections.

Table 2: Sources for the main, ETT, LSF4, and finance/M4 sets. Multiple horizons are sampled from each dataset according to Appendix A.

Datasets	Source / benchmark	Reference
ETTh1, ETTh2, ETTm1, ETTm2	Electricity Transformer Temperature	Zhou et al. (2021)
Electricity	UCI ElectricityLoadDiagrams	Trindade (2015)
Exchange Rate	LSTNet exchange-rate benchmark	Lai et al. (2018)
Traffic	Caltrans PeMS, benchmark preprocessing	California Department of Transportation (n.d.); Lai et al. (2018)
Weather	MPI-BGC Jena weather station	Max Planck Institute for Biogeochemistry (n.d.); Wu et al. (2021)
M4 daily, hourly, monthly, quarterly, weekly, yearly	M4 competition train series	Makridakis et al. (2020)
FRED foreign exchange and rates	Federal Reserve Bank of St. Louis	Federal Reserve Bank of St. Louis (n.d.)
ETH hourly	Binance ETHUSDT market data	Binance (n.d.)
Apple daily	Yahoo Finance historical prices	Yahoo Finance (n.d.)
TSMC daily	LLMTIME memorization dataset	Gruver et al. (2023)

The finance/M4 source series are Apple, TSMC, ETHUSDT, three exchange-rate series (JPY/USD, USD/EUR, and USD/GBP), four rates series (federal funds, US 2-year and 10-year yields, and the term spread), and the five M4 daily-to-yearly groups used in that set. The TSMC file is the 2022 daily series distributed with LLMTIME.

A.3 Registered transformations and utility

Let d_h denote the control-ensemble standard deviation in the original numerical units. The registered rounding operation with strength a maps $x_{r,h}$ to $ad_h \,\mathrm{round}(x_{r,h}/(ad_h))$, for $a \in \{0.25, 0.5\}$. The Gaussian perturbation adds independent $\mathcal{N}(0, (ad_h)^2)$ errors for $a \in \{0.05, 0.1\}$. Standard deviations below 10^{-30} are floored at that value for these operations. Sorting orders complete rows lexicographically, while shuffling applies a random row permutation. Row removal retains $\mathrm{round}(0.7B)$ randomly chosen rows. Horizon cropping retains the first $\mathrm{round}(0.7H)$ steps and their original indices. Identity completes the nine transformations. Each transformation produces its own release of a unit, so joint survival requires nine separate single-operation detections on that unit; it reports the worst registered operation, not a composite chain of operations. Removal acts on trajectories, not on time: temporal decimation that re-indexes forecast steps changes the public step metadata and lies outside the registered set.

Table 3: The 17 Monash datasets in the 104-unit archive set. The archive citation is Godahewa et al. (2021); links identify the individual Zenodo records. "Variants" links include the archive's versions with and without missing values. Counts are evaluated units.

Dataset	Units	Archive record
Australian electricity demand	9	4659727
Bitcoin	9	5121965 / 5122101 (variants)
CIF 2016	3	4656042
COVID deaths	6	4656009
FRED-MD	9	4654833
Hospital	3	4656014
NN5 daily	6	4656110 / 4656117 (variants)
NN5 weekly	3	4656125
Pedestrian counts	9	4656626
Saugeen river flow	3	4656058
Sunspot	3	4654773 / 4654722 (variants)
Tourism monthly	6	4656096
Tourism quarterly	3	4656093
Traffic hourly	9	4656132
Traffic weekly	3	4656135
US births	3	4656049
Weather	9	4654822

Table 4: The eight Darts examples, one evaluated unit each. The implementation and source references are documented by Herzen et al. (2022); Unit8 (n.d.). Loader names uniquely identify the released example series.

Dataset / loader	Series
AirPassengersDataset	Monthly airline passengers, 1949–1960
AusBeerDataset	Quarterly Australian beer production
GasRateCO2Dataset	Gas rate and carbon dioxide measurements
HeartRateDataset	Instantaneous heart rate, 1,800 measurements
MonthlyMilkDataset	Monthly milk production, 1962–1975
SunspotsDataset	Monthly sunspot numbers, 1749–1983
WineDataset	Monthly Australian wine sales
WoolyDataset	Quarterly Australian woollen yarn production

For ensemble $x_1, \ldots, x_B$ and outcome y at one forecast step, the fair ensemble score is

$$\mathrm{CRPS}_{\mathrm{f}}(x, y) = \frac{1}{B}\sum_{r=1}^{B}|x_r - y| - \frac{1}{2B(B-1)}\sum_{r=1}^{B}\sum_{r'=1}^{B}|x_r - x_{r'}|. \tag{13}$$

The unit score averages Equation 13 across its evaluated steps. Equation 11 then compares sums of these unit scores. The 10–90% interval uses the corresponding empirical ensemble quantiles; pooled coverage concatenates the per-step inclusion indicators, giving each evaluated step equal weight. Confidence intervals use 2,000 bootstrap draws over dataset/series clusters. Post-transformation utility compares the transformed ensemble with the unmarked control on the same retained steps.

A.4 Results by evaluation set and dataset

Tables 5 and 6 give the complete five-set detection and pooled-utility results under the shared numerical-release protocol. They report both identity detection and survival after the registered release operations, together with the corresponding forecast-quality changes, allowing reuse survival and utility to be compared under a common evaluation setting.

The complete main-set dataset comparison is reported in Table 1b. The dataset-level and set-level results use the same numerical band rule, original forecast-step indices, and control-derived transformation amplitudes throughout. The resulting variation therefore reflects performance under a common watermarking and release protocol rather than differences in how the evaluation is configured.

Table 5: Detection across five evaluation sets. Rates are proportions. I_{40} and I_4 denote identity detection at the two rank thresholds; Q_a and G_a denote rounding and Gaussian noise in control-standard-deviation units. The main set publishes 64 rows per unit and the four confirmation sets publish 128, so the sets differ in released evidence per unit (Theorem 8).

Set	n	I_{40}	I_4	$Q_{.25}$	$Q_{.5}$	$G_{.1}$	Drop	Crop	Joint
Main	984	0.996	0.988	0.946	0.761	0.941	0.995	0.991	0.759
Finance/M4	199	1.000	1.000	0.955	0.849	0.970	1.000	0.990	0.849
ETT	84	0.952	0.917	0.893	0.762	0.881	0.952	0.940	0.750
LSF4	84	0.988	0.976	0.905	0.738	0.929	0.988	0.988	0.738
Monash/Darts	104	1.000	0.990	0.865	0.567	0.827	0.990	0.990	0.558

Table 6: Forecast utility across evaluation sets. CRPS changes and their 95% intervals are percentages; coverage changes are percentage points.

Set	Pooled CRPS [95% CI]	Coverage [95% CI]
Main	+0.26 [+0.10, +0.44]	+0.22 [-0.09, +0.54]
Finance/M4	+0.18 [-0.07, +0.48]	+0.78 [-0.15, +1.48]
ETT	+0.01 [-0.24, +0.31]	+0.31 [-0.53, +1.14]
LSF4	+0.71 [+0.48, +0.99]	-0.99 [-2.03, -0.13]
Monash/Darts	+0.35 [-0.24, +0.51]	-0.54 [-1.16, +0.15]

B Formal guarantees and the mechanism of numerical evidence

B.1 Assumptions and the implemented interfaces

The verifier receives a multiset of valid pairs (h, V), where h is an original forecast-step index and V is an integer at serializer quantum $q = 10^{-3}$. The public scaler, declared width, band width, and nonce are fixed before verification. In the ideal key-function model, the phase at each step is uniform on $\{0, \dots, \Delta - 1\}$, the band colours are independent fair bits, and distinct keys have independent copies of these variables. The implementation separates the phase and colour domains and obtains an exactly uniform integer phase by rejection sampling. A concrete pseudorandom function provides the computational interpretation of these ideal assumptions.

The following results have different scopes. Rank validity, row invariance, decoy-score variance, and the VFF displacement and CRPS bounds apply to the stated verifier or sender directly. The local carrier-response theorem studies a single exponential tilt with a key-independent candidate law before nucleus truncation. Its role is to identify the geometric terms that a release operation preserves or destroys. We then give the exact local formulas for correlated colours, explain their relationship to the two-stage sender, and separate them from the conditional retention approximation used to compare the measured transformations.

B.2 Finite-sample calibration and release selection

Proof of Theorem 1. Set $N = K + 1$ and condition on a null release and its metadata. Every score is the same measurable function of this fixed release and an independent, identically distributed key function, so $(S_0, \dots, S_K)$ is exchangeable. Introduce independent continuous tie breakers $U_0, \dots, U_K$ and rank (S_k, U_k) in decreasing lexicographic order. Exchangeability makes the owner's randomized rank R_0 uniform on $\{1, \dots, N\}$. The conservative rank counts every tied decoy ahead of the owner, hence $\rho \geq R_0$ and

$$\Pr(\rho \leq t \mid X, \text{metadata}) \leq \Pr(R_0 \leq t \mid X, \text{metadata}) = t/N. \tag{14}$$

The bound is attained when the scores have no ties. No independence assumption on forecast cells enters this argument. □

Theorem 4 (Row-order invariance and validity after selection)**.** *Every row permutation leaves the complete score vector, its ties, and ρ unchanged. Under the null assumptions of Theorem 1, a*

key-independent rule retaining rows R' and original steps H' gives

$$S'_k = \sum_{(r,h)\in\mathcal{I}\cap(R'\times H')} c_k(h, V_{r,h}), \qquad \Pr(\rho' \le t \mid \textit{selected release}) \le \frac{t}{K+1}. \tag{15}$$

The selection rule may depend on the null values and independent randomness.

Proof. For any row permutation π, reindexing the finite sum gives $\sum_{r,h} c_k(h, V_{\pi(r),h}) = \sum_{r,h} c_k(h, V_{r,h})$ for every key. A selected null release remains independent of all evaluated key functions when the selection uses only key-independent information. Theorem 1 then applies to the restricted sum. Keeping the original step labels ensures that every retained value is scored at the same numerical coordinate. □

For fixed public transformations $T_1, \dots, T_m$ that preserve null key independence, the union bound gives

$$\Pr\Big(\min_{1\le j\le m} \rho_j \le u \mid H_0\Big) \le \sum_{j=1}^{m} \Pr(\rho_j \le u \mid H_0) \le \frac{mu}{K+1}. \tag{16}$$

The five-factor scale search uses $m = 5$ and $u = 8$, preserving the nominal $40/4096$ level. The candidate tests may be dependent. The number of public hypotheses is part of this calibration; a larger search requires corresponding multiplicity accounting.

B.3 Carrier covariance and the effect of a release operation

Let $\mathcal{X}$ be a finite candidate space and write $b_k(x) = 2c_k(x) - 1 \in \{-1, 1\}$. Assume $\mathbb{E}_k b_k(x) = 0$ for each fixed x, and define

$$\kappa(x, z) = \mathbb{E}_k[b_k(x) b_k(z)]. \tag{17}$$

This kernel records whether two objects consult shared keyed evidence. It permits correlations between nearby candidates and does not assume independent colours for individual values. Fix a candidate distribution π independent of the key and a key-independent stochastic release map $Q(z \mid x)$. For the local tilted law, set

$$\pi_{k,a}(x) = \frac{\pi(x) e^{a b_k(x)}}{Z_k(a)}, \qquad Z_k(a) = \sum_x \pi(x) e^{a b_k(x)}, \qquad r_k(x) = \sum_z Q(z \mid x) c_k(z). \tag{18}$$

The expected post-release excess is $\mu_Q(a) = \mathbb{E}_k \mathbb{E}_{\pi_{k,a}} r_k - 1/2$. The channel may describe deterministic rounding, fixed changes of numerical units, or independent noise. The channel is fixed independently of the evaluated key; operations chosen to suppress evidence from an entire marked ensemble require a separate model.

Proof of Theorem 2. For fixed k, differentiating the finite exponential family yields

$$\frac{d}{da} \mathbb{E}_{\pi_{k,a}} r_k = \mathrm{Cov}_{\pi_{k,a}}(r_k, b_k). \tag{19}$$

At $a = 0$, expand the covariance using independent $X, X' \sim \pi$ and $Z \sim Q(\cdot \mid X)$. Since $c_k(z) = (1 + b_k(z))/2$ and every fixed b_k has mean zero over keys,

$$\mathbb{E}_k \mathbb{E}_\pi[r_k(X) b_k(X)] = \tfrac{1}{2} \mathbb{E}\kappa(X, Z), \tag{20}$$

$$\mathbb{E}_k\big[\mathbb{E}_\pi r_k(X)\, \mathbb{E}_\pi b_k(X')\big] = \tfrac{1}{2} \mathbb{E}\kappa(X', Z). \tag{21}$$

Subtracting proves Equation 9. Marginal fairness also gives $\mu_Q(0) = 0$.

To bound the Taylor remainder, write $f_k(a) = \mathbb{E}_{\pi_{k,a}} r_k$. A second differentiation gives the mixed central moment

$$f''_k(a) = \mathbb{E}_{\pi_{k,a}}\big[(r_k - \mathbb{E}_{\pi_{k,a}} r_k)(b_k - \mathbb{E}_{\pi_{k,a}} b_k)^2\big]. \tag{22}$$

Because $0 \le r_k \le 1$ and $b_k \in [-1, 1]$, its absolute value is at most $\mathrm{Var}_{\pi_{k,a}}(b_k) \le 1$. Integrating this bound twice gives $|\mu_Q(a) - a\mu'_Q(0)| \le a^2/2$ for $a \ge 0$. □

Define the self-retention term $R_Q = \mathbb{E}\kappa(X, Z)$ and the background-overlap term $B_Q = \mathbb{E}\kappa(X', Z)$. A positive difference $R_Q - B_Q$ gives a positive initial response. The remainder bound additionally certifies $\mu_Q(a) > 0$ for $0 < a < R_Q - B_Q$. Because $R_Q \leq 1$ and $B_Q \geq 0$, this certificate is confined to weak tilts $a < 1$: at the operational strength $a = \lambda/T \approx 7.1$ the quadratic remainder $a^2/2$ exceeds the entire range of μ_Q, and positive excess instead follows from the accessible-band laws below (Theorems 6 and 7). For identity release, $R_Q = 1$ and $B_Q = \mathbb{E}\kappa(X, X')$, so a carrier that assigns the same colour to every accessible candidate has zero local capacity. If a channel maps every candidate to the same fixed value, the two terms coincide and the expected score is exactly $1/2$ at every strength: the release has removed the choice through which evidence was encoded.

The band kernel is a geometric quantity. Condition on a phase. Two values at the same step have product $b_k(h, V)b_k(h, V')$ of expected value one when their shifted band indices agree, and zero otherwise. Different steps use independent colours. For integers V, V' with $d = |V - V'| < \Delta$, exactly $\Delta - d$ of the Δ possible phases put them in the same band. If $d \geq \Delta$, no phase does. Averaging proves Equation 8 without any assumption that a marked sample is uniform inside its band. Consequently, small displacement raises R_Q, while spreading candidate probability across multiple bands lowers B_Q. Increasing Δ can improve the former and worsen the latter; numerical reuse tolerance and editable capacity therefore impose distinct requirements on the band scale.

An exact condition for loss of token-key evidence. Consider a watermark whose embedding depends on a random key table at a fixed source-address set $\mathcal{A}$. Suppose that, after a transformation, the verifier consults a fixed disjoint set $\mathcal{B}$, and that the corresponding owner-table entries are independent of all embedding randomness and entries in $\mathcal{A}$. Conditional on the release, the owner's table on $\mathcal{B}$ then has the same law as each independent decoy's table. The complete score vector is exchangeable and its detection probability is at most $t/(K + 1)$, even though embedding used the owner key. This follows from the rank proof with the unused part of the owner table serving as a fresh key.

This condition formalizes a particular loss of alignment. Absolute token positions or changed token contexts can move an aligned detector toward this regime; recovery procedures can also reuse source evidence. The condition is not automatically satisfied by a row deletion or by every token watermark. The reported aligned and edit-distance EXP/ITS comparisons test those concrete procedures empirically. In contrast, a row permutation in V-band preserves (h, V) exactly, and its invariance follows without a fresh-address assumption.

Theorem 5 (A sufficient finite-strength regime for surviving evidence)**.** *Suppose the local candidate law in Equation 18 has its entire support at a common original step h, with at least two supported values, and every distinct pair is separated by at least 2Δ quanta. Let the key-independent release preserve that step and satisfy $|Z - X| \leq \varepsilon\Delta$ almost surely for some $0 \leq \varepsilon < 1$. Under the random-phase band construction,*

$$\mu_Q(a) \geq (1 - \varepsilon)\mu_{\mathrm{id}}(a) > 0 \qquad (a > 0). \tag{23}$$

Here μ_{id} is the identity excess of the same untruncated local tilt. If s values have positive candidate mass, its strong-tilt limit additionally satisfies $\liminf_{a\to\infty} \mu_Q(a) \geq (1 - \varepsilon)(1/2 - 2^{-s})$.

Proof. Write the candidates as v_i with masses π_i. Their colours C_i are independent for every phase because the candidates occupy distinct bands. A perturbed v_i cannot enter any other candidate's band: for $j \neq i$, $|Z - v_j| \geq 2\Delta - \varepsilon\Delta > \Delta$. Its destination therefore either retains C_i or consults a band colour independent of every colour that affected the local selection.

Define $A_i(a) = \mathbb{E}_C[\pi_{k,a}(v_i)(C_i - 1/2)]$. For fixed colours of the other candidates, let $W = \sum_{j\neq i} \pi_j e^{a(2C_j - 1)} > 0$. Averaging over C_i gives

$$A_i(a) = \frac{1}{4}\mathbb{E}_{C_{-i}}\left[\frac{\pi_i e^a}{\pi_i e^a + W} - \frac{\pi_i e^{-a}}{\pi_i e^{-a} + W}\right] > 0. \tag{24}$$

This quantity is independent of phase, and $\sum_i A_i(a) = \mu_{\mathrm{id}}(a)$. The average probability that v_i and its released value stay in the same band is $r_i = \mathbb{E}[\kappa(v_i, Z) \mid X = v_i] \geq 1 - \varepsilon$. Fresh destination colours contribute zero excess, so $\mu_Q(a) = \sum_i r_i A_i(a) \geq (1 - \varepsilon)\sum_i A_i(a)$. Finally, all s candidate bands are red with probability 2^{-s}, yielding the stated limit. □

This theorem supplies an explicit regime where accessible numerical alternatives retain positive evidence at every tilt strength, while the disjoint-address condition above removes the aligned carrier completely. It does not assume uniformly distributed marked positions inside a band. The separation condition is substantial: densely packed digit candidates and the two-stage, top-p sender require the more general analysis below. The result establishes a sufficient mechanism, and the experiments determine how closely the implemented choices realize it.

B.4 Accessible bands, correlated colours, and the units-only limit

Theorem 6 (Tilted-colour law over accessible bands)**.** *Fix a phase and a candidate law π independent of the band colours. Let p_b be the probability mass in occupied band b, let $s_{\mathrm{band}} = |\{b : p_b > 0\}|$, and let $g = \sum_b p_b C_b$. For the binary local tilt before nucleus truncation,*

$$P_a(g) = \frac{ge^{2a}}{1 + g(e^{2a} - 1)}, \qquad \frac{\partial P_a(g)}{\partial a} = 2P_a(g)[1 - P_a(g)]. \tag{25}$$

Writing $\mu(\pi, a) = \mathbb{E}_C P_a(g) - 1/2$, independent fair colours on the occupied bands give

$$0 \le \mu(\pi, a) \le \tfrac{1}{2} \tanh a, \tag{26}$$

$$\mu(\pi, a) = \frac{a}{2}\left(1 - \sum_b p_b^2\right) + O(a^2), \tag{27}$$

$$\lim_{a \to \infty} \mu(\pi, a) = \tfrac{1}{2} - 2^{-s_{\mathrm{band}}}. \tag{28}$$

For $a > 0$, the excess is positive precisely when at least two bands have positive candidate mass.

Proof. Summing the tilted mass over green and red candidates gives $ge^a/[ge^a + (1-g)e^{-a}]$, establishing Equation 25. If $E = e^{2a}$, then

$$P_a(g) - g = \frac{g(1-g)(E-1)}{1 + g(E-1)} \ge 0, \qquad \frac{\partial^2 P_a(g)}{\partial g^2} = \frac{-2E(E-1)}{[1 + g(E-1)]^3} \le 0. \tag{29}$$

Thus marginal fairness $\mathbb{E}_C g = 1/2$ and Jensen's inequality give the lower and upper bounds. At positive strength the gain is strictly positive on $0 < g < 1$. This event has positive probability exactly when two or more occupied bands have positive mass.

The expansion $P_a(g) = g + 2ag(1-g) + O(a^2)$ is uniform for $g \in [0, 1]$. Independence across occupied bands gives $\mathbb{E}_C g^2 = 1/4 + (1/4)\sum_b p_b^2$, proving Equation 27. In the strong-tilt limit, $P_a(g) \to \mathbf{1}\{g > 0\}$. All occupied bands are red with probability $2^{-s_{\mathrm{band}}}$, proving Equation 28. □

When candidate colours are individually independent, the masses p_b reduce to individual candidate probabilities, recovering the collision-entropy expression. Neighbouring units candidates in V-band often share a band. Their number of distinct digits can therefore overstate capacity. If the phase is also random and the candidate law is independent of it, averaging the theorem over phases gives the corresponding phase-averaged laws.

Theorem 7 (An arbitrary-strength ceiling for a short editable interval)**.** *Fix a key-independent nonempty set of accessible integers $\mathcal{A}$ whose span is $L = \max \mathcal{A} - \min \mathcal{A} < \Delta$. Any randomized sender that selects its output from $\mathcal{A}$, even with access to every band colour, has expected green excess at most $L/(4\Delta)$ over the uniform phase and independent fair band colours. In particular, units-only editing with a copied key-independent prefix and $\Delta \ge 10$ has excess at most $9/(4\Delta)$.*

Proof. For exactly L of the Δ possible integer phases, a boundary lies between the minimum and maximum candidates. The accessible set then touches two bands; otherwise it touches one. With one occupied band, even an optimal selection has expected colour $1/2$. With two occupied bands, it can select a green value only if at least one band is green, an event of probability $3/4$. Therefore its expected colour is bounded by

$$\left(1 - \frac{L}{\Delta}\right)\frac{1}{2} + \frac{L}{\Delta}\frac{3}{4} = \frac{1}{2} + \frac{L}{4\Delta}. \tag{30}$$

A units digit can reach at most the ten integers associated with its fixed higher prefix, whose span is nine. Restricting model support or applying nucleus truncation cannot enlarge this accessible set. □

At $\Delta = 90$, this ceiling is only 0.025 excess per value. Editing tens as well as units expands the reachable range and can expose additional band colours. The theorem gives a necessary geometric condition for capacity; it does not guarantee that the model assigns useful probability to all reachable values. The development units-only ablation has identity detection 0.297 and joint detection 0.047 at $\lambda = 3$, consistent with this restricted carrier access. These detection rates also depend on the number and dependence of the scored values, so they are not numerical estimates of the excess ceiling.

More generally, for a complete interval of accessible integers with span $L = q_0\Delta + r_0$, where $0 \le r_0 < \Delta$, the number of occupied bands is $q_0 + 1$ with probability $1 - r_0/\Delta$ and $q_0 + 2$ otherwise. Selecting a green candidate whenever one exists attains the geometric envelope

$$\mu_{\max}(L, \Delta) = \frac{1}{2} - 2^{-(q_0+1)} \left(1 - \frac{r_0}{2\Delta}\right). \tag{31}$$

This follows by averaging the probability that all accessible bands are red. For the full two-digit interval, $L = 99$ and the default $\Delta \le 90$ expose at least two bands at every phase. This avoids a purely geometric single-band obstruction when the model can use the full interval. Actual model support and conditional probabilities still determine how much of this envelope is available.

B.5 Soft completion weights and the actual sampling law

The tens weight is the geometric fraction of green completions, $w_d \in [0, 1]$. For a fixed pre-nucleus digit law π, define

$$Z(a) = \sum_d \pi_d e^{a(2w_d - 1)}, \qquad \pi_d(a) = \frac{\pi_d e^{a(2w_d-1)}}{Z(a)}, \qquad m(a) = \sum_d \pi_d(a) w_d. \tag{32}$$

Direct differentiation gives

$$\frac{d \log Z}{da} = 2m(a) - 1, \qquad m'(a) = 2 \operatorname{Var}_{\pi(a)}(w) \ge 0, \qquad \frac{d}{da} D_{\mathrm{KL}}(\pi(a) \| \pi) = 4a \operatorname{Var}_{\pi(a)}(w). \tag{33}$$

The last identity follows from $D_{\mathrm{KL}}(\pi(a)\|\pi) = a(2m(a) - 1) - \log Z(a)$. Equal weights leave the normalized digit law unchanged and are skipped by the implementation. A fractional completion fraction is a geometric score, so substituting $\sum_d \pi_d w_d$ for g in the binary formula does not yield $m(a)$.

What nucleus truncation changes. For nonconstant completion weights, the implementation masks disallowed tokens, adds the digit bias, applies temperature and nucleus truncation over the allowed vocabulary, and computes the remaining digit mass. Let $\pi(a)$ be the digit-conditioned law before nucleus truncation, $A(a)$ the surviving digit set, and $m_{\mathrm{digit}}(a)$ its mass under the full-vocabulary post-nucleus distribution. On the sampling branch, where the weights are nonconstant and $m_{\mathrm{digit}}(a) > 10^{-9}$, the actual digit law is $\widehat{\pi}(a) = \pi(a)(\cdot \mid A(a))$ and

$$\mathrm{TV}(\widehat{\pi}(a), \pi(a)) = 1 - \pi(a)(A(a)), \qquad \left|\mathbb{E}_{\widehat{\pi}(a)} w - \mathbb{E}_{\pi(a)} w\right| \le 1 - \pi(a)(A(a)). \tag{34}$$

To verify the equality, sum the absolute differences on $A(a)$ and its complement; each contributes the discarded digit probability, and total variation divides their sum by two. The expectation bound follows because $0 \le w \le 1$. The two fallback branches have a different actual law: if all completion weights are equal, or if $m_{\mathrm{digit}}(a) \le 10^{-9}$, the implementation copies the control digit d^0 and $\widehat{\pi}(a) = \delta_{d^0}$. Its total variation from a defined pre-nucleus digit law is then $1 - \pi(a)(d^0)$, so Equation 34 is restricted to the sampling branch.

On an interval of strengths with a fixed retained set, the conditional family still satisfies the variance identity in Equation 33. When the set changes, the mean can jump. For example, take digit probabilities $(0.4501, 0.45, 0.0549, 0.045)$, weights $(0, 0.8, 0.2, 1)$, and nucleus probability 0.9. At $a = 0$ only the first two digits survive. At $a = 0.01$ the third digit also enters, and the post-nucleus mean weight decreases from approximately 0.4000 to 0.3915, although the pre-nucleus mean increases. This is a constructed sampling example, not a forecast result.

The nominal top-p threshold also does not by itself bound the discarded *conditional digit* mass by 0.1: if digits carry full-vocabulary mass β, a discarded full-vocabulary mass ε can remove up

to $\min(1, \varepsilon/\beta)$ of the digit law. The sender retains the control digit when no usable digit mass survives. Likewise, a singleton unmarked nucleus does not imply zero capacity at every strength. With base digit probabilities $(0.999, 0.001)$, a positive tilt can restore the second digit before the nucleus is recomputed. A true point mass in the pre-tilt support cannot acquire an alternative through reweighting.

The two stages of VFF. For a fixed copied prefix P, the actual expected final colour is

$$\sum_{d_1} \widehat{\pi}_{k,1}(d_1 \mid Z) \sum_{d_0} \widehat{\pi}_{k,0}(d_0 \mid Z, d_1) c_k(h, P + 10d_1 + d_0). \tag{35}$$

Each $\widehat{\pi}$ in Equation 35 is the actual branch-dependent law: the conditional-nucleus distribution on a usable sampling branch, or the point mass at the corresponding control digit under either fallback. The units conditional is recomputed from the published tens digit. Consequently, this conditional and the chosen prefix can already depend on the owner colours. The fixed-law colour average in Theorem 6 cannot simply be applied twice to obtain a whole-sender power guarantee. The geometric interval weight, the model's conditional completion probability, and the final verified colour are distinct quantities. Equation 35 states how the implementation combines them, and Algorithms 1 and 2 give its executable sequence.

B.6 Perturbation retention and its empirical interpretation

The kernel in Equation 8 is exact for fixed values averaged over random keys. A separate conditional model describes retention of an already marked value at finite strength. Assume its within-band position is uniform, its conditional green excess is a constant μ across those positions and perturbations, the displacement e is independent of that position, and a destination band reached after a crossing has an independent fair colour. The transformed excess then equals μr, where

$$r = \mathbb{E}\left[\left(1 - \frac{|e|}{\Delta}\right)_+\right]. \tag{36}$$

Indeed, for fixed e, the proportion of source positions staying in their band is $(1 - |e|/\Delta)_+$. Staying preserves the conditional excess, while the assumed destination recolouring contributes zero excess. Integrating over the displacement proves the expression.

For an ideal uniform rounding error on $[-g/2, g/2]$,

$$r_{\mathrm{round}}(g) = \frac{2}{g}\int_0^{\min(g/2,\Delta)} \left(1 - \frac{u}{\Delta}\right) du = \begin{cases} 1 - \dfrac{g}{4\Delta}, & g \le 2\Delta, \\ \dfrac{\Delta}{g}, & g \ge 2\Delta. \end{cases} \tag{37}$$

For independent Gaussian error with standard deviation τ,

$$r_{\mathrm{Gauss}}(\tau) = \frac{2}{\sqrt{2\pi}\tau}\int_0^{\Delta} \left(1 - \frac{u}{\Delta}\right) e^{-u^2/(2\tau^2)} du \tag{38}$$

$$= 2\Phi(\Delta/\tau) - 1 - \frac{\tau}{\Delta}\sqrt{\frac{2}{\pi}}\left[1 - e^{-\Delta^2/(2\tau^2)}\right], \tag{39}$$

where Φ is the standard normal distribution function. All displacement and band-width quantities use serializer-quanta units. Deterministic rounding on a structured numerical distribution need not produce uniform rounding error.

The measured main-set checks use 983 units with identity excess at least 0.05. Predicted mean retention for rounding at 0.25σ and 0.5σ is 0.832 and 0.679, compared with observed excess ratios 0.706 and 0.490. Gaussian noise at 0.05σ and 0.1σ gives predictions 0.892 and 0.791, compared with 0.804 and 0.648. The respective across-unit Pearson correlations are 0.72, 0.69, 0.78, and 0.76. These observations support the ordering by displacement relative to band width. They also show that the conditional model systematically overpredicts retention, so Equation 36 should be read as an optimistic envelope rather than a calibrated predictor. One mechanism behind the optimism is boundary clustering: fractional tens weights bias sampled values toward the green side of a band

boundary, so marked values sit closer to boundaries than the uniform-position assumption allows, and small displacements then cross bands more often than predicted. Uniform marked positions, constant conditional excess, and fresh destination colours are further assumptions, so this expression is neither a universal upper bound nor a calibrated prediction of detection power. The source/background decomposition in Equation 9 identifies another term that a pure same-band retention factor omits.

B.7 Repeated bands and effective evidence

Rows can be dependent even under a fresh decoy key: two values at the same step and in the same band consult the same bit. Let a fixed release contain M scored cells $x_i = (h_i, V_i)$ and define

$$D(X) = \sum_{i,j=1}^{M} \kappa(x_i, x_j), \qquad M_{\text{eff}}(X) = \frac{M^2}{D(X)}. \tag{40}$$

Theorem 8 (Exact decoy variance and a finite-sample power certificate)**.** *For a fresh decoy independent of a fixed release,*

$$\mathbb{E}(S_k \mid X) = M/2, \qquad \mathrm{Var}(S_k \mid X) = D(X)/4. \tag{41}$$

For phase vector φ, let $n_{h,b}(\varphi_h)$ count released values in band b at step h, and set

$$Q_*(X) = \sum_h \max_{0 \le \varphi_h < \Delta} \sum_b n_{h,b}(\varphi_h)^2. \tag{42}$$

Condition on the release and owner score $S_0 = M/2 + u$ with $u > 0$. If the K decoy keys are independent of this conditioning, then

$$\Pr(\rho > t \mid X, S_0) \le \min\left\{1, \frac{K}{t} \exp\left(-\frac{2u^2}{Q_*(X)}\right)\right\}. \tag{43}$$

Proof. Conditional on phases, $S_k = \sum_{h,b} n_{h,b} C_k(h,b)$ is a weighted sum of independent fair bits. Its mean is $M/2$ and its conditional variance is $\frac{1}{4}\sum_{h,b} n_{h,b}^2$. Averaging the variance over phases introduces no variance-of-the-mean term. Each pair (i,j) contributes exactly its probability of sharing a band, which is $\kappa(x_i, x_j)$. This proves Equation 41.

For completeness, the conditional moment-generating function of $S_k - M/2$ factors into $\prod_{h,b} \cosh(\theta n_{h,b}/2)$. The bound $\cosh v \le e^{v^2/2}$ makes this at most $\exp(\theta^2 \sum_{h,b} n_{h,b}^2/8)$. Optimizing the exponential Markov bound over $\theta > 0$ gives

$$\Pr(S_k - M/2 \ge u \mid X, \varphi) \le \exp\left(-\frac{2u^2}{\sum_{h,b} n_{h,b}^2}\right) \le \exp\left(-\frac{2u^2}{Q_*(X)}\right). \tag{44}$$

This uniform phase bound survives averaging over the phase. If $J = \sum_{k=1}^{K} \mathbf{1}\{S_k \ge S_0\}$, then $\rho > t$ is the event $J \ge t$. Its expectation is at most $K \exp(-2u^2/Q_*)$, and Markov's inequality proves Equation 43. □

The exact decoy variance uses the phase-average $D(X)$; the exponential bound uses the phase maximum $Q_*(X)$. Substituting the average into that uniform tail bound is generally unjustified. Equation 43 conditions on the observed release and owner score, so it is an a-posteriori certificate for a release in hand rather than an a-priori power guarantee: the marginal law of S_0 under the implemented two-stage, nucleus-truncated sender has no closed form, and the K/t factor from Markov's inequality is loose. A tighter evaluation uses the exact one-key tail $p_X = \Pr(S_k \ge S_0 \mid X, S_0)$, which gives the conditional decision probability

$$\Pr(\rho \le t \mid X, S_0) = \sum_{j=0}^{t-1} \binom{K}{j} p_X^j (1 - p_X)^{K-j}. \tag{45}$$

This identity includes ties through the definition of p_X and makes no model assumption about how the marked release was generated.

Equation 40 has two immediate consequences. First, copying every row c times multiplies every score by c, preserves every rank, and leaves M_{eff} unchanged: M scales by c, whereas D scales by c^2. Duplicate evidence cannot improve detection. Second, under independent retention of each row with probability f, pairs of distinct rows are kept with probability f^2 and diagonal cells with probability f. Pairs from the same row but different steps have zero kernel. Consequently,

$$\mathbb{E}[M'] = fM, \qquad \mathbb{E}[D(X')] = f^2 D(X) + f(1-f)M. \tag{46}$$

An independent-cell approximation sets $D = M$ and gives the familiar $\sqrt{f}$ reduction of a mean standardized margin. Correlated occupancy changes that relation. Equation 46 is an identity for the expected numerator of the null variance, not an equality for the expectation of a ratio or a universal power curve. The exact variance is computable from released values and public Δ, allowing the evidence-count prediction to be checked without rerunning the forecasting model.

B.8 VFF DISPLACEMENT, FORECAST-SCORE STABILITY, AND FEEDBACK

Theorem 9 (Horizon-independent VFF displacement and CRPS stability)**.** *Suppose a VFF update changes only digit places in a set $\mathcal{P}$, preserving its higher copied prefix. Let $L_{\mathcal{P}} = 9\sum_{p\in\mathcal{P}} p$. Its published offset satisfies*

$$|V^{\text{pub}}_{r,h} - V^0_{r,h}| \le L_{\mathcal{P}}, \qquad |x^{\text{pub}}_{r,h} - x^0_{r,h}| \le sqL_{\mathcal{P}}. \tag{47}$$

For any finite, equally sized, row-paired ensembles with $B \ge 2$ and any finite target sequence, Equation 10 holds. Thus the default $\mathcal{P} = \{10, 1\}$ gives $L_{\mathcal{P}} = 99$ and absolute mean fair-CRPS change at most $198sq$, independently of the forecast horizon.

Proof. A digit changes by at most nine, so the triangle inequality bounds the integer offset by $9\sum_{p\in\mathcal{P}} p$. The implemented publication rule adds exactly $q(V^{\text{pub}} - V^0)$ in scaled value space; applying the affine inverse multiplies this by s.

At one forecast step, the implemented fair score can be written

$$\text{CRPS}_{\text{f}}(x, y) = \frac{1}{B}\sum_{r=1}^{B} |x_r - y| - \frac{1}{B(B-1)}\sum_{r<r'} |x_r - x_{r'}|. \tag{48}$$

For $\delta_r = x'_r - x_r$, the reverse triangle inequality bounds the change of the first term by $B^{-1}\sum_r |\delta_r|$. Each pairwise distance changes by at most $|\delta_r| + |\delta_{r'}|$, and each row appears in $B-1$ pairs. The second term therefore changes by at most the same amount. Adding and averaging over H steps gives Equation 10. The argument is deterministic and requires neither independent ensemble rows nor a distributional assumption on y. □

Applying this result requires matching the published rows to the same control rows and using the same finite-value mask in both scores. A comparison with a larger control pool must first restrict it to the paired rows. The bound is a worst-case absolute score bound; it does not imply a particular relative CRPS percentage. The editable range can be larger than one band: at $\Delta = 10$, changing the final two digits from 00 to 99 moves 99 quanta. The construction therefore has no universal displacement guarantee of Δ quanta. The observed utility benefits additionally depend on the model conditionals and the values actually selected.

A comparison with coupled feedback. To isolate propagation, consider a scalar recurrence with shared innovations, $x^0_h = F_h(x^0_{h-1}, \xi_h)$, and a coupled release $x^c_h = F_h(x^c_{h-1}, \xi_h) + e_h$. Suppose F_h is L_h-Lipschitz in its state and $|e_h| \le \varepsilon_h$. If both trajectories start at the same value, their difference obeys

$$|x^c_h - x^0_h| \le \sum_{j=1}^{h} \varepsilon_j \prod_{\ell=j+1}^{h} L_\ell. \tag{49}$$

This follows by applying $D_h \le L_h D_{h-1} + \varepsilon_h$ recursively. For constant L and ε, the bound is $\varepsilon\sum_{j=0}^{h-1} L^j$. A linear recurrence $F_h(x, \xi) = Lx + \xi$ and coherent edits $e_h = \varepsilon$ attain this state-displacement bound, showing that the accumulation term cannot be removed in this class. A VFF publication $x^{\text{pub}}_h = x^0_h + e_h$ instead has $|x^{\text{pub}}_h - x^0_h| \le \varepsilon_h$ directly.

This recurrence is a mechanism model, not an estimated Lipschitz description of a particular language model. Its implication is the possibility of accumulation through feedback and its algebraic removal by VFF. The paired development experiment reports mean relative CRPS changes of $+1.68\%$ for coupled feedback and $+0.37\%$ for VFF. The theorem explains the removed propagation path; it does not derive those percentages or imply that every coupled forecaster must worsen utility.

B.9 Worked digit-steering example

For example, take a serialized control value 1234, band width $\Delta = 10$, and an owner-green band $[1223, 1233)$. With higher prefix 12, the tens candidates 2 and 3 admit completions 1220–1229 and 1230–1239, giving green fractions 0.7 and 0.3. Equation 3 therefore increases the logit difference in favour of candidate 2 by 0.8λ. If the sender samples that digit, the new prefix is 122 and the green units candidates are 3 through 9. Recomputing the units law from this prefix can produce a release such as 1227, while the original 1234 remains in the history used for later forecasts. The example links the fractional weight of a completion interval to the conditional choice of its final digit.

B.10 Implementation-level pseudocode

Algorithm 2 details how the local conditionals in Equation 35 are obtained without replacing the control history. The control sequence includes token separators; the units forward feeds the published tens token together with the separator tokens between the two digit positions. Its position identifiers match the original positions. The attention mask exposes the original cache strictly before the tens token and the appended local tokens, while hiding the original tens token and every subsequent control token. Removing the appended cache entries after the forward keeps the next step's history unmarked.

Algorithm 2 Batched VFF cache updates for the default tens-and-units schedule

Require: Tokenized control rows with digit positions, model, n, Δ, owner prefix sums G_h, strength λ
Set $V^{\mathrm{pub}} \leftarrow V^0$; mark each position whose width equals the declared n.
for each batch of control rows **do**
  Teacher-force the full control sequences; retain logits and the cache of length L.
  for each valid cell (r, h) in the batch **do**
    Read tens position $p_{r,h}$ and the control prefix $P = 100\lfloor V^0_{r,h}/100 \rfloor$.
    Read the logits predicting that position from the teacher-forced output.
    Draw published tens using $\textsc{Steer}(\ell, h, P, 10, d^0_1)$ below.
  end for
  for each forecast step h **do**
    For every active row, append its published tens token and the intervening separator tokens.
    Restore their original position identifiers; expose cached tokens only before the original tens position.
    Run one batched local forward; read the logits predicting units from its final local token.
    for each active row r **do**
      Draw published units using $\textsc{Steer}(\ell, h, P + 10d_1, 1, d^0_0)$.
      Set $V^{\mathrm{pub}}_{r,h} = P + 10d_1 + d_0$.
    end for
    Crop the cache back to length L.
  end for
end for
Publish $A^{-1}(A(X^0) + q(V^{\mathrm{pub}} - V^0))$; retain control values at skipped cells.
function $\textsc{Steer}(\ell, h, P, p, d^0)$
  Set $w_d = [G_h(P + (d+1)p) - G_h(P + dp)]/p$ for $d = 0, \ldots, 9$.
  If all w_d are equal, return d^0; otherwise add $\lambda(2w_d - 1)$ to digit logits.
  Apply the allowed-token mask, temperature 0.7, and top-p 0.9 over allowed tokens.
  If surviving digit mass is at most 10^{-9}, return d^0; otherwise renormalize it and sample.
end function

For a one-digit declared value or a units-only ablation, the single marked position uses its teacher-forced conditional and needs no local-prefix forward. The sender's diagnostic compares selected masked-cache digit logits with a direct forward on the explicitly edited prefix. This checks conditional equivalence; it does not treat a key-selected prefix as independent of the key. The reported KL diagnostic and the KL-budgeted ablation use the digit law before nucleus truncation, matching Equation 33.

Connections to the empirical results. The empirical results illustrate distinct consequences of the analysis. Row reordering preserves every score exactly, consistent with the equal 0.996 identity, sort, and shuffle detection rates. Restricting access to the units interval must reduce attainable colour diversity, consistent with the units-only ablation. Increasing strength in a fixed local law must approach a limit determined by occupied bands, consistent with main-set identity rates 0.944, 0.986, 0.996, 0.999, 0.999, and 0.999 for $\lambda = 3, \ldots, 8$. These observations are compatible with saturation, while top-p changes prevent treating the local formula as an exact whole-sender fit. The control-NLL bins also caution against equating NLL with capacity: mean identity excess is 0.389 below 0.1 nats and 0.309 at or above 1.5 nats, with Spearman correlation -0.19. Finally, the occupancy formula predicts that exact row duplication leaves ranks unchanged and that the fresh-key variance equals $D/4$.

C Generation-time comparator protocols

All comparator releases use the same numerical forecast interface, control-derived transformation amplitudes, and nine release operations. KGW (Kirchenbauer et al., 2023) uses green fraction $\gamma = 0.5$ and bias $\delta = 2$. EXP and ITS (Kuditipudi et al., 2023) use numerical-alphabet sampling constructions and aligned key-coordinate statistics. SynthID-Text (Dathathri et al., 2024), Permute-and-Flip (Zhao et al., 2025), DiPmark (Wu et al., 2024), and Unigram (Zhao et al., 2023) use their numerical-alphabet adapters and associated scoring rules. The detection decision uses the smallest owner rank among each method's declared detector readings: coordinates for KGW/EXP/ITS, n-gram lengths for SynthID-Text, green fractions for DiPmark, and scoring statistics for Permute-and-Flip. Unigram has one reading. These declared readings are aggregated without an additional multiplicity correction.

For each token arm, the detection denominator includes units with at least one valid detector evaluation. An unavailable transformation result within an included unit counts as nondetection, and joint survival requires all nine detection decisions to pass. Units with no valid evaluation for an arm are excluded. The main evaluation set contains 984 units, and the identity-evaluation counts for KGW, ITS, and Permute-and-Flip lie between 981 and 983. Wilson intervals use available identity evaluations, and paired comparisons use matched units. On finance/M4, DiPmark detection covers all 199 units and Permute-and-Flip covers 195: its generator enforces an internal nucleus-mass invariant (its assertion P-L4) and refuses the remaining units, with 199 units for the corresponding V-band comparison.

Table 7: Token-comparator identity/joint detection on additional evaluation sets. Rates use the aligned EXP/ITS detectors and prespecified aggregated readings for other methods. A blank denotes an unreported comparison. Finance/M4 PF has partial coverage (195 of 199 units).

Method	Finance/M4	ETT	LSF4	Monash/Darts
KGW	0.744 / 0.010	0.631 / 0.000	0.940 / 0.000	0.856 / 0.010
EXP	0.869 / 0.005	0.857 / 0.000	0.988 / 0.000	0.904 / 0.010
ITS	0.759 / 0.005	0.631 / 0.000	0.905 / 0.000	0.875 / 0.000
SynthID-Text	0.834 / 0.000			
PF ($n = 195$)	0.841 / 0.010			
DiPmark	0.809 / 0.000			
Unigram	0.744 / 0.030			

D Transfer across 26 generator checkpoints

D.1 Checkpoint identities and release references

Table 8 maps every evaluated checkpoint group to its technical report or release documentation. Qwen3-4B refers specifically to the Instruct-2507 release, the 1B and 3B Llama checkpoints belong to Llama 3.2, and the Mistral group comprises three distinct releases. The model-size abbreviations in Table 9 use these exact versions.

Table 8: Sources for all 26 transfer checkpoints. Sizes within a row share the stated release family and reference.

Release	Evaluated sizes / variant	Reference
Qwen2.5-Instruct	0.5B, 1.5B, 3B, 7B, 14B, 32B	Qwen et al. (2024)
Qwen3	0.6B, 1.7B, 8B, 14B, 32B	Yang et al. (2025)
Qwen3-Instruct-2507	4B	Qwen Team (2025)
Qwen3.5	2B, 4B, 9B, 27B	Qwen Team (2026)
Gemma 3 IT	1B, 4B, 12B, 27B	Gemma Team (2025)
Llama 3.2 Instruct	1B, 3B	Meta (2024)
Llama 3.1 Instruct	8B	Grattafiori et al. (2024)
Mistral 7B	v0.1 base	Jiang et al. (2023)
Ministral 8B	Instruct-2410	Mistral AI (2024)
Mistral Small 24B	Instruct-2501	Mistral AI (2025)

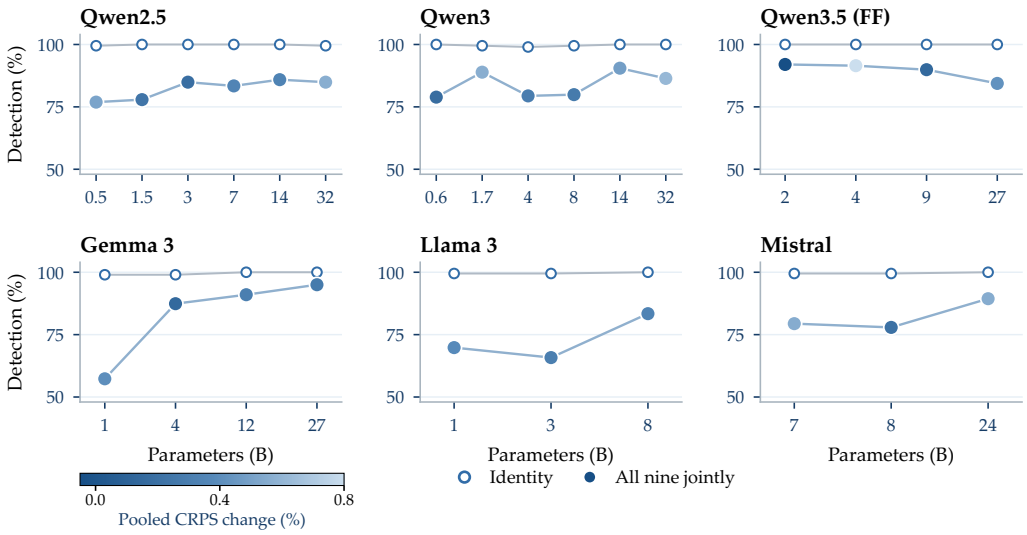

Figure 5: **Transfer across 26 checkpoints.** Six small multiples show identity and joint detection on the 199-unit finance/M4 set. Point colours encode pooled CRPS change, and lines connect the evaluated checkpoints within each family. Qwen3.5 uses independent in-loop FF generation; the other 22 checkpoints use VFF. The band rule, strength, and verifier are shared across the study.

The separate numerical-format diagnostic uses gpt-oss-20b (OpenAI, 2025); it is outside the 26 compatible-checkpoint transfer table. The Qwen3.5 reference follows the authors' supplied citation, while the Qwen3-2507 and Mistral model cards identify the specific releases used here.

The confirmation experiment evaluates 26 compatible checkpoints on the same 199-unit finance/M4 set with 128 forecast rows, $\lambda = 5$, and each model's own control ensemble. The band-width rule, verifier, and control-scale transformations are fixed across checkpoints. Table 10 gives complete detection results, Table 9 gives utility with uncertainty, and Table 11 reports additional main-set runs.

Twenty-two checkpoints use VFF. The four Qwen3.5 checkpoints use in-loop feedback-free (FF) generation because their hybrid cache does not support the masked VFF update. As in the digit-level FF ablation (Figure 7a), model logits condition on unmarked prefixes. In this implementation, a separate autoregressive pass samples the unmarked fed trajectory and its published digit substitutions, so the marked outputs are independently sampled relative to the reference control rows. This sampling difference is part of the reported sender configuration, and it has two consequences. First, cost: in-loop FF uses 0.75–1.0 additional generations against VFF's 0.22–0.23 marked pass (Appendix F). Second, uncertainty: the row-paired stability bound of Proposition 3 does not constrain independently sampled ensembles, so the Qwen3.5 utility estimates in Table 9 include independent Monte Carlo sampling variance, visible in their wider intervals. The Qwen2.5 models and base Mistral-7B use the

Table 9: **V-band across all 26 checkpoints, with 199 finance/M4 units each.** I_4: identity at rank 4; Joint: all nine transformations at rank 40. Detection and pooled CRPS changes are percentages; brackets give 95% cluster-bootstrap intervals. Higher detection and lower CRPS are better. The four Qwen3.5 rows use the independent in-loop FF sender (Section 5.3); their utility estimates include independent-sampling variance, and the row-paired bound of Proposition 3 does not apply to them. Release references appear in Table 8.

Checkpoint	$I_4 \uparrow$	Joint↑	CRPS↓ [95% CI]	Checkpoint	$I_4 \uparrow$	Joint↑	CRPS↓ [95% CI]
Qwen2.5-0.5B	99.0	76.9	+0.51 [+0.10, +0.92]	Qwen3-0.6B	97.0	78.9	+0.17 [-0.28, +0.58]
Qwen2.5-1.5B	100.0	77.9	+0.23 [+0.00, +0.51]	Qwen3-1.7B	98.5	88.9	+0.56 [+0.28, +0.88]
Qwen2.5-3B	100.0	84.9	+0.18 [-0.07, +0.48]	Qwen3-4B	99.0	79.4	+0.29 [+0.08, +0.50]
Qwen2.5-7B	99.5	83.4	+0.34 [+0.14, +0.57]	Qwen3-8B	98.0	79.9	+0.27 [+0.08, +0.51]
Qwen2.5-14B	100.0	85.9	+0.33 [+0.12, +0.58]	Qwen3-14B	98.5	90.5	+0.48 [+0.26, +0.74]
Qwen2.5-32B	99.5	84.9	+0.56 [+0.36, +0.81]	Qwen3-32B	100.0	86.4	+0.61 [+0.37, +0.83]
Gemma-3-1B	97.0	57.3	+0.40 [-0.13, +0.89]	Qwen3.5-2B	100.0	92.0	-0.05 [-0.69, +0.61]
Gemma-3-4B	99.0	87.4	+0.14 [-0.06, +0.33]	Qwen3.5-4B	100.0	91.5	+0.80 [+0.28, +1.32]
Gemma-3-12B	100.0	91.0	+0.30 [+0.04, +0.60]	Qwen3.5-9B	100.0	89.9	+0.14 [-1.10, +1.27]
Gemma-3-27B	100.0	95.0	+0.27 [-0.01, +0.55]	Qwen3.5-27B	100.0	84.4	+0.42 [-0.89, +1.50]
Llama-3.2-1B	99.5	69.8	+0.38 [+0.08, +0.72]	Mistral-7B-v0.1	99.5	79.4	+0.55 [+0.22, +0.93]
Llama-3.2-3B	96.5	65.8	+0.30 [-0.07, +0.68]	Ministral-8B	98.5	77.9	+0.20 [-0.30, +0.67]
Llama-3.1-8B	100.0	83.4	+0.33 [+0.04, +0.71]	Mistral-Small-24B	100.0	89.4	+0.55 [+0.27, +0.86]

Transformers 4.46.3/PyTorch 2.5.1 environment. Gemma-3, Qwen3, Llama-3, and the additional Mistral models use Transformers 5.16.1; Qwen3.5 uses PyTorch 2.6.

Table 10: All 26 checkpoint detection results on finance/M4 ($n = 199$ per checkpoint). Rates are proportions. I_{40} and I_4 are identity detection; $Q_{.5}$ is rounding at 0.5σ; Joint requires detection after all nine registered operations, as in Table 1. FF denotes the independent feedback-free sender, and all other rows use VFF.

Checkpoint	Sender	I_{40}	I_4	$Q_{.5}$	Joint
Qwen2.5-0.5B-Instruct	VFF	0.995	0.990	0.774	0.769
Qwen2.5-1.5B-Instruct	VFF	1.000	1.000	0.779	0.779
Qwen2.5-3B-Instruct	VFF	1.000	1.000	0.849	0.849
Qwen2.5-7B-Instruct	VFF	1.000	0.995	0.834	0.834
Qwen2.5-14B-Instruct	VFF	1.000	1.000	0.859	0.859
Qwen2.5-32B-Instruct	VFF	0.995	0.995	0.849	0.849
Qwen3-0.6B	VFF	1.000	0.970	0.789	0.789
Qwen3-1.7B	VFF	0.995	0.985	0.910	0.889
Qwen3-4B-Instruct-2507	VFF	0.990	0.990	0.804	0.794
Qwen3-8B	VFF	0.995	0.980	0.799	0.799
Qwen3-14B	VFF	1.000	0.985	0.905	0.905
Qwen3-32B	VFF	1.000	1.000	0.864	0.864
Qwen3.5-2B	FF	1.000	1.000	0.920	0.920
Qwen3.5-4B	FF	1.000	1.000	0.915	0.915
Qwen3.5-9B	FF	1.000	1.000	0.899	0.899
Qwen3.5-27B	FF	1.000	1.000	0.844	0.844
gemma-3-1b-it	VFF	0.990	0.970	0.573	0.573
gemma-3-4b-it	VFF	0.990	0.990	0.874	0.874
gemma-3-12b-it	VFF	1.000	1.000	0.910	0.910
gemma-3-27b-it	VFF	1.000	1.000	0.950	0.950
Llama-3.2-1B-Instruct	VFF	0.995	0.995	0.709	0.698
Llama-3.2-3B-Instruct	VFF	0.995	0.965	0.663	0.658
Llama-3.1-8B-Instruct	VFF	1.000	1.000	0.834	0.834
Mistral-7B-v0.1	VFF	0.995	0.995	0.804	0.794
Ministral-8B-Instruct-2410	VFF	0.995	0.985	0.779	0.779
Mistral-Small-24B-Instruct-2501	VFF	1.000	1.000	0.894	0.894

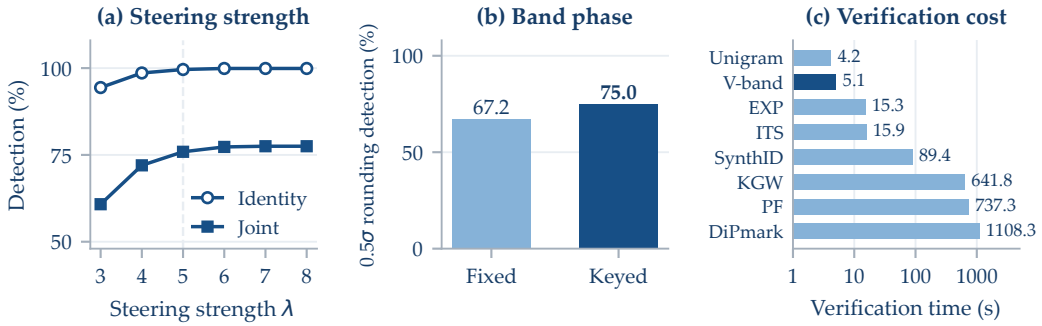

Figure 6: **Steering, band geometry, and computation.** (a) Main-set detection across the strength sweep, with the default $\lambda = 5$ marked. (b) Fixed and keyed phases on dev-64 under 0.5σ rounding. (c) Measured verification time on one pinned CPU core, with a logarithmic axis, for V-band and every token implementation timed on the shared block of 64 rows, 128 steps, and 4,096 tested keys.

The compatible checkpoints span identity detection of 0.990–1.000 and joint survival of 0.573–0.950. Increasing model size changes joint survival nonmonotonically in several families, even when identity is nearly constant. Band widths below the 90-quantum cap accompany several models with high reuse survival, including Gemma-3-12B/27B and Qwen3-14B. The checkpoint comparison thus measures variation in the sampled numerical distributions as well as model scale.

Table 11: Additional generator results on the 984-unit main set. Detection rates are percentages; CRPS is pooled relative change. All three rows use VFF and 64 forecast trajectories.

Generator	Identity	Joint	CRPS (%)
Qwen2.5-3B	99.6	75.9	+0.26
Gemma-3-4B	97.6	73.5	+0.17
Qwen3-8B	97.5	72.1	+0.36

E STEERING STRENGTH AND LOCAL DESIGN

The main strength sweep evaluates $\lambda \in \{3, 4, 5, 6, 7, 8\}$ under the same band-width rule. Table 12 gives identity and joint detection, pooled and median CRPS changes, and interval-coverage changes. The default $\lambda = 5$ is near the measured detection knee. The empirical choice uses pooled CRPS as the primary utility statistic. The local capacity analysis in Theorem 6 explains why increasing tilt eventually gives diminishing changes in green evidence, while the band geometry controls survival through numerical transformations. The sender comparison requests 128 forecast rows on dev-64 at

Table 12: Main-set strength sweep: reuse detection and forecast utility summaries. Detection rates are proportions. All CRPS changes are percentages and coverage is in percentage points, with 95% intervals. These are the bootstrap intervals from the strength-sensitivity analysis.

λ	Identity	Joint	Pooled CRPS [CI]	Median [CI]	Coverage [CI]
3	0.944	0.608	+0.19 [+0.09, +0.29]	+0.17 [+0.12, +0.22]	+0.14 [-0.04, +0.33]
4	0.986	0.720	+0.27 [+0.14, +0.41]	+0.28 [+0.21, +0.37]	+0.21 [-0.04, +0.49]
5	0.996	0.759	+0.26 [+0.09, +0.43]	+0.39 [+0.31, +0.50]	+0.22 [-0.08, +0.55]
6	0.999	0.773	+0.29 [+0.08, +0.50]	+0.48 [+0.38, +0.56]	+0.15 [-0.16, +0.48]
7	0.999	0.775	+0.30 [+0.07, +0.54]	+0.49 [+0.39, +0.61]	+0.13 [-0.19, +0.46]
8	0.999	0.775	+0.31 [+0.08, +0.54]	+0.50 [+0.42, +0.61]	+0.06 [-0.27, +0.40]

$\lambda = 1.5$. Coupled, digit-level FF, and VFF give mean relative CRPS changes of +1.68%, +0.38%, and +0.37%, respectively (Figure 7a). Appendix I.3 reports actual row counts and paired uncertainty. FF computes model logits from unmarked digit prefixes, while VFF recomputes the units law from

published tens; both evaluate watermark weights at the published prefix. The four Qwen3.5 transfer checkpoints use independent in-loop FF for their hybrid caches; the other 22 use VFF on unmarked control rows. Keyed phase raises dev-64 detection after 0.5σ rounding from 67.2% to 75.0%, with identity fixed at 96.9%.

F EMBEDDING AND VERIFICATION EFFICIENCY

The measured VFF pass uses 0.22–0.23 times an unmarked generation on the named 3B–7B models in Table 13. Because VFF first constructs the unmarked control ensemble, its end-to-end total is 1.22–1.23 unmarked generations; the token watermarks embed within a single generation at totals of 1.05–1.15, and in-loop FF uses 0.75–1.0 additional generations, for totals of 1.75–2.0, in the reported comparisons. Table 13 states both the added and the total cost for every method so the two accounting conventions cannot be conflated. These ratios are measured implementation costs and can change with batching, cache support, and checkpoint size.

Table 13: Embedding cost on one H800, relative to one unmarked generation of the same ensemble. "Added" is the compute beyond that unmarked generation (for VFF, the separate marked pass; for token watermarks, the in-generation overhead). "Total" is end-to-end marked-release compute, including VFF's control ensemble.

Method	Added	Total
Qwen2.5-3B VFF	0.23	1.23
Qwen2.5-7B VFF	0.23	1.23
Mistral-7B VFF	0.23	1.23
Gemma-3-4B VFF	0.22	1.22
In-loop FF (reported range)	0.75–1.0	1.75–2.0
KGW	0.10	1.10
EXP	0.05	1.05
ITS	0.05	1.05
SynthID-Text	0.14–0.15	1.14–1.15
Permute-and-Flip	0.09–0.10	1.09–1.10
DiPmark	0.12	1.12
Unigram	0.07	1.07

Table 14 measures CPU verification over the same 4,096 candidate keys on one pinned core. The first block is M4 quarterly with 64 rows and 128 steps, containing 8,192 values. The second is finance/M4 monthly with 128 rows and 24 steps, containing 3,072 values. The 5.1-second V-band identity decision on the first block includes approximately 1.9 seconds of key-table construction. The six cells missing at submission (DiPmark and Unigram on the main block; KGW and the aligned EXP/ITS on finance/M4) were measured in one additional sitting after generating the missing single-unit dumps, on the same host class and pinned-core protocol; re-timing every pre-existing cell in that sitting reproduced it within 7%. Value-space comparator embedding is a post-hoc operation with reported millisecond cost; no comparable per-arm timing breakdown is available.

The timing advantage comes from precomputing keyed band tables and gathering their colours in batches. Token detectors incur different hashing and scoring costs under their numerical adapters. The table compares the measured implementations on the stated unit and CPU core, with each timing column using a common block and 4,096 candidate keys.

G RELATED WORK

Numerical carriers and output editing. WGTD and Ngo et al. already place keyed evidence in numerical intervals, while TabularMark uses a partitioned perturbation domain (He et al., 2024; Ngo et al., 2026; Zheng et al., 2024). These precedents establish the numerical-carrier idea. WGTD's original insertion algorithm takes a table and a binning rule and moves values into selected intervals; it does not require changing generator weights. HTW likewise modifies forecast values through a keyed heads/tails mapping (van Schaik, 2024). Such output editing can operate independently of the generator, and we adapt WGTD, Ngo et al., and HTW as measured arms in Appendix H.

Table 14: Measured verification times in seconds on one pinned CPU core. Each column identifies one fixed block, and every detector is timed on its own arm's dump of that block. Full-suite rows state the number of variants; these differ across detector families.

Detector/task	Main: 64×128	Finance/M4: 128×24
V-band identity	5.1	1.0
V-band all 53 variants	66.5	31.0
EXP aligned, 13 variants	15.3	8.6
ITS aligned, 13 variants	15.9	7.5
Unigram identity	4.2	4.5
Unigram all reported variants	6.9	4.5
SynthID-Text identity	89.4	36.3
SynthID-Text all reported variants	826.6	305.9
KGW identity	641.8	373.0
Permute-and-Flip identity	737.3	623.6
DiPmark identity	1108.3	670.6

TabularMark differs on the verifier side: it perturbs selected key cells within a keyed green/red partition of a bounded deviation range, and its one-proportion z-test compares the suspected data against the owner-retained original records after key-cell matching. That detector requires the originals and per-cell correspondence rather than released values and public metadata alone, and no reference implementation accompanies the paper, so we cite it as numerical-carrier precedent instead of adapting it as a measured arm. V-band uses conditional digit probabilities to choose watermark-bearing values, with explicit isolation of feedback across forecast steps. This additional access supplies a model-based preference among nearby values.

Encoder-guided post-hoc insertion. Waltz targets LLM-based time-series forecasters with a time-series encoding interface (Yuan et al., 2025). It constructs a watermark book from rarely aligned token embeddings and uses projected gradient descent to optimize bounded output perturbations. The optimized variables are the perturbations, not the forecasting model's parameters. Its reported experiments watermark the patch-encoder forecasters TEMPO and UniTime by optimizing against their token-embedding spaces; digit-serialized decoding from a frozen LLM exposes no patch-embedding interface of this kind, and no implementation of the optimization has been released. Our Waltz-labelled comparator is therefore a bounded-perturbation proxy, not the original encoder-guided algorithm; its measurements bound what an unguided perturbation budget achieves and do not establish a performance comparison with the full Waltz method.

Generation-time diffusion watermarking. TabWak embeds during tabular diffusion sampling and verifies through latent reconstruction; its released implementation uses the TabSyn VAE and diffusion backbone (Zhu et al., 2025). TimeWak targets multivariate time-series diffusion models, embeds temporally chained seeds in the initial noise, and recovers them using diffusion inversion (Soi et al., 2025). Its BDIA-DDIM inversion uses the model's noise estimator, and its inversion-error theorem assumes Lipschitz continuity of that estimator. A frozen digit-serialized LLM offers neither a diffusion sampler to seed nor a noise estimator to invert, so these methods cannot be instantiated in our setting; they are the diffusion counterparts of generation-time value-space watermarking rather than runnable comparators here.

Generation-time watermarking with a learned codebook. LVQMark biases red/green token choices in the autoregressive transformer of L-VQVAE (Kim et al., 2026). Its carrier is a learned codebook for multivariate time series, and detection re-encodes the released signal with a robust encoder. Both stages require training the L-VQVAE, so its relation to a frozen digit-vocabulary LLM is architectural rather than something a shared benchmark can measure. This recent work establishes another generation-time interface beyond diffusion. V-band uses the existing digit vocabulary of a frozen LLM and defines its carrier directly on numerical intervals, so verification requires no learned re-encoding model.

Scope of the present contribution. The distinctive contribution is conditional-probability insertion into numerical bands with feedback isolation for probabilistic forecast ensembles from frozen

autoregressive LLMs. VFF can teacher-force unmarked control trajectories and resample local digits before release, so generation-time access can be implemented as a separate forward pass. Verification scores values with the declared scaler and serializer, using no generator or learned encoder. The measured advantage concerns the tradeoff between numerical-reuse detection and forecasting quality at the reported operating points. Both generation-side and output-only methods remain relevant under their respective access conditions.

Language models for numerical forecasting. Language models provide several routes from historical observations to forecasts. LLMTIME serializes observations as digit strings and treats forecasting as next-token prediction, including a mapping from token probabilities to continuous predictive densities (Gruver et al., 2023). PromptCast expresses the forecasting task through textual prompts (Xue & Salim, 2022), while pretrained-transformer adaptation, Time-LLM, and LLM4TS learn interfaces between numerical series and language-model representations (Zhou et al., 2023; Jin et al., 2024; Chang et al., 2023). A complementary line pretrains models directly on time series: Chronos uses scaled, quantized value tokens, TimesFM uses a patched decoder, and Lag-Llama models probabilistic forecasts with lag covariates (Ansari et al., 2024; Das et al., 2023; Rasul et al., 2023). PatchTST further demonstrates the utility of subseries patches as forecasting tokens (Nie et al., 2023). These representation choices concern how a model learns and generates a forecast. V-band addresses the subsequent provenance question for digit-serialized ensembles: how can a provider embed evidence that remains detectable in the released numerical values?

Watermarks embedded during generation. Keyed changes to language-model sampling can leave statistically detectable evidence in generated text. KGW promotes a context-dependent green list, while Unigram-Watermark uses a fixed vocabulary partition to improve robustness to text editing (Kirchenbauer et al., 2023; Zhao et al., 2023). Distribution preservation motivates complementary constructions. Kuditipudi et al. couple a keyed random sequence to inverse-transform or exponential-minimum sampling and detect the resulting dependence through alignment (Kuditipudi et al., 2023). Unbiased Watermark and DiPmark develop distribution-preserving reweighting, and Permute-and-Flip pairs a watermark with its alternative decoder (Hu et al., 2024; Wu et al., 2024; Zhao et al., 2025). SynthID-Text introduces Tournament sampling and supports integration with speculative sampling at production scale (Dathathri et al., 2024). Cryptographic work additionally formalizes undetectability to observers who lack the secret key (Christ et al., 2024). V-band retains keyed sampling as its insertion mechanism and changes the object that carries the evidence. Digit biases are computed from the green mass of numerical intervals, and verification scores values at their forecast steps. This construction makes row order irrelevant and connects tolerance to numerical perturbations with the width of the intervals.

Verification after modification. Watermark robustness depends on the transformation and the surviving evidence. Studies of paraphrasing and text mixing show that retained fragments can support detection when enough tokens remain (Kirchenbauer et al., 2024), while recursive paraphrasing, signature inference, and watermark stealing expose distinct removal and spoofing threats (Sadasivan et al., 2023; Jovanović et al., 2024). Statistical tests for learned spoofing artifacts further distinguish detecting a mark from establishing how it was produced (Gloaguen et al., 2024). Our evaluation of numerical reuse measures both detection and the quality of the modified forecast ensemble. Fair CRPS accounts for finite ensemble size, and interval coverage complements the scalar score with a calibration diagnostic (Ferro, 2014; Gneiting & Raftery, 2007; Gneiting et al., 2007). The resulting comparisons relate the evidence retained by a watermark to the forecasting cost of the operation that modifies it.

H Supplementary comparator results

H.1 Post-hoc value-space comparators

The value-space comparators edit the same unmarked ensembles used by V-band. HTW (van Schaik, 2024) rewrites selected digits toward green digits, with rewrite rate r and a pooled digit-colour statistic. The Waltz-labelled arm (Yuan et al., 2025) applies keyed bounded perturbations with budget ϵ in scaled numerical units. WGTD (He et al., 2024) edits paired value bins with a position-keyed green rule, while the adaptation of Ngo et al. (2026) hashes numerical bins. Their bin width δ is

expressed in scaled units and rounded to the serializer quantum. The digit-based and bin-based arms use their respective statistics, and the bin arms attempt edits at all eligible values. These results evaluate the stated forecasting adaptations. The Waltz arm in particular is a bounded-noise proxy: it applies unguided uniform perturbations, whereas the published Waltz method optimizes its perturbations with projected gradient descent against a patch-encoder forecaster's token embeddings and has no released implementation for digit-serialized forecasts (Appendix G). Its rows therefore measure what a given perturbation budget buys, not the published method's operating frontier.

Tables 15 and 16 present the main-set value-space operating points. Several obtain perfect identity detection, with different numerical reuse survival and utility costs. The bounded-noise proxy at $\epsilon = .10$ reaches 76.8% joint survival at +1.82% pooled CRPS, WGTD at $\delta = .05$ reaches 89.7% at +2.26%, and HTW at $r = .5$ reaches 87.0% at +23.6%. V-band at $\lambda = 5$ reaches 75.9% at +0.26%. The comparison concerns the measured operating points and their predictive costs.

Table 15: Main-set detection of the post-hoc value-space comparators at all reported strengths, with the default V-band point for reference. Rates are proportions. Q_a and G_a use the same control-standard-deviation perturbation amplitudes across arms. Joint requires all nine registered decisions. The Waltz-proxy rows are unguided uniform bounded noise, not the encoder-guided Waltz optimization.

Method	Identity ↑	$Q_{.25}$ ↑	$Q_{.5}$ ↑	$G_{.1}$ ↑	Drop ↑	Crop ↑	Joint ↑
V-band $\lambda = 5$	0.996	0.946	0.761	0.941	0.995	0.991	0.759
HTW $r = .1$	0.821	0.190	0.098	0.272	0.819	0.728	0.052
HTW $r = .5$	1.000	0.977	0.871	0.999	1.000	1.000	0.870
Waltz proxy $\epsilon = .02$	1.000	0.440	0.179	0.564	1.000	1.000	0.170
Waltz proxy $\epsilon = .05$	1.000	0.801	0.495	0.922	1.000	1.000	0.487
Waltz proxy $\epsilon = .10$	1.000	0.923	0.773	0.980	1.000	1.000	0.768
WGTD $\delta = .02$	1.000	0.848	0.480	0.976	1.000	1.000	0.480
WGTD $\delta = .03$	1.000	0.963	0.636	0.993	1.000	1.000	0.636
WGTD $\delta = .05$	1.000	0.986	0.898	0.996	1.000	1.000	0.897
Ngo et al. $\delta = .02$	1.000	0.845	0.466	0.943	1.000	1.000	0.462
Ngo et al. $\delta = .03$	1.000	0.956	0.646	0.983	1.000	1.000	0.644

Table 16: Identity intervals, the stricter rank threshold, and forecast utility for the main-set value-space comparison. Intervals are Wilson 95% intervals; I_4 is identity at rank ≤ 4. CRPS columns are percentage changes, and coverage is in percentage points. The Waltz-proxy rows are unguided uniform bounded noise (Appendix G).

Method	Identity ↑ [95% CI]	I_4 ↑	Pooled ↓	Median ↓	Coverage Δ
V-band $\lambda = 5$	0.996 [0.990, 0.998]	0.988	+0.26	+0.39	+0.22
HTW $r = .1$	0.821 [0.796, 0.844]	0.678	+0.25	-0.23	+3.48
HTW $r = .5$	1.000	0.999	+23.6	+3.26	+7.97
Waltz proxy $\epsilon = .02$	1.000	1.000	+0.11	+0.07	-0.13
Waltz proxy $\epsilon = .05$	1.000	1.000	+0.59	+0.44	-0.55
Waltz proxy $\epsilon = .10$	1.000	1.000	+1.82	+1.62	-1.14
WGTD $\delta = .02$	1.000	1.000	+0.41	+0.31	-0.16
WGTD $\delta = .03$	1.000	1.000	+0.89	+0.81	-1.03
WGTD $\delta = .05$	1.000	1.000	+2.26	+1.93	-1.04
Ngo et al. $\delta = .02$	1.000	1.000	+0.33	+0.13	+0.28
Ngo et al. $\delta = .03$	1.000	1.000	+0.57	+0.39	-0.50

H.2 Paired identity comparisons and diagnostic subsets

Table 17 supplies identity intervals, median utility changes, and paired counts for the token configurations. The declared best-reading aggregation and available-case denominators are specified in Appendix C. Because alternative readings are aggregated without a further multiplicity correction, these comparator rates do not inherit the single-reading bound in Theorem 1.

Table 17: Additional main-set statistics for generation-time comparators. Identity rates are proportions with Wilson 95% intervals, median CRPS changes are percentages, and the last two columns count matched units detected by only one method. Each paired comparison uses V-band at $\lambda = 5$.

Comparator	Identity ↑ [95% CI]	Median CRPS ↓	V-band only	Other only
KGW	0.785 [0.759, 0.810]	+0.64	209	1
EXP (aligned)	0.884 [0.863, 0.903]	+0.00	111	1
ITS (aligned)	0.788 [0.763, 0.814]	+0.00	206	1
SynthID-Text	0.859 [0.836, 0.879]	-0.53	136	1
Permute-and-Flip	0.853 [0.830, 0.874]	+0.48	142	1

The corresponding reported exact two-sided McNemar p-values range from 2.6×10^{-61} for KGW to 4.4×10^{-32} for EXP. The HTW $r = .1$ comparison has 173 V-band-only units and one HTW-only unit. Stronger value-space configurations with perfect identity have four comparator-only units and no V-band-only units, giving $p = 0.13$. Their difference from V-band is therefore assessed through reuse survival and utility as well as identity.

Table 18: Main-set diagnostic subset with control NLL above 0.3 nats. The V-band subset contains 829 units, or 84.2% of the main set; its pooled CRPS change is +0.33%. Token rates use available records and their declared readings.

Method	Identity ↑	Joint ↑
V-band $\lambda = 5$	0.999	0.753
KGW	0.844	0.004
EXP (aligned)	0.930	0.002
ITS (aligned)	0.888	0.001
SynthID-Text	0.918	0.000
Permute-and-Flip	0.917	0.001

The control-NLL subset is always reported alongside the full set. On ETT, V-band detects 80/84 identity releases, with Wilson interval [0.884,0.981], and EXP detects 72/84, with interval [0.767,0.916]. Their discordant counts are 9/1 in favour of V-band. On the 49-unit diagnostic subset, V-band detects 48/49 and EXP 49/49. On the LSF4 diagnostic subset, both detect 77/78. EXP rounding-0.5σ detection is 0.429 on ETT and 0.583 on LSF4, while its row-removal detection is 0.036 and 0.060. These outcomes distinguish identity capacity from robustness under the same release operations.

H.3 Value-space operating points on additional sets

Table 19 retains the additional reported value-space comparisons. ETT WGTD reaches 90.5% joint survival with a pooled CRPS reduction of 0.7% and a coverage increase of 2.0 percentage points. The main-set operating-point ordering consequently does not extend to every dataset. Each set should be read using both its detection and its own forecast-quality measurements.

Table 19: Additional reported value-space operating points. CRPS is the pooled percentage change, and joint is a proportion. The Waltz-proxy arm is unguided uniform bounded noise.

Set	Comparator	Joint ↑	CRPS ↓
Finance/M4	WGTD $\delta = .03$	0.759	+2.62
Finance/M4	Waltz proxy $\epsilon = .05$	0.548	+1.64
ETT	HTW $r = .5$	0.750	+2.4
ETT	WGTD $\delta = .05$	0.905	-0.7
LSF4	WGTD $\delta = .05$	0.881	+2.1

H.4 Edit-distance detection for EXP and ITS

The fixed-origin edit-distance configuration scores the released numerical stream against the keyed sequence using the upstream Levenshtein kernels, with one whole-stream alignment and a fixed key

origin. It does not include the full robust detector's search over text/key subsequences and key origins. Each permutation test uses 100 null evaluations with fresh keys and vocabulary permutations. A release is detected when no null evaluation reaches the owner statistic, corresponding to $1/101 \approx 0.990\%$. The insertion/deletion penalty is $c_{\text{indel}} \in \{0, 1\}$, with 0 the upstream default; we write c_{indel} rather than the upstream γ because γ denotes the KGW green-list fraction elsewhere in this paper.

Table 20 reports the completed measurements on all 104 Monash/Darts units for every configuration. At $c_{\text{indel}} = 1$, EXP identity is 0.894 and ITS identity is 0.856, matching their aligned detectors on the same units (0.904 and 0.856). The edit-distance statistic recovers part of the row-operation loss for EXP, raising row shuffle from 0.115 to 0.202 and row removal from 0.077 to 0.115, but none of it for ITS, whose row operations stay at or below 0.048 under both scorings. Under value edits it performs worse than alignment: rounding at 0.5σ falls from 0.346 to 0.019 for EXP and from 0.115 to 0.029 for ITS. ITS with $c_{\text{indel}} = 0$ has a degenerate near-zero edit cost for both owner and null alignments on this ten-digit alphabet, producing zero detections even on identity; that configuration does not measure useful discrimination. The aligned detectors' collapse under numerical edits is therefore not an artefact of alignment: the upstream robust statistic inherits it.

Table 20: Fixed-origin edit-distance comparisons on Monash/Darts, complete on the same 104 units per cell. Rates are proportions. The last column counts aligned-only / edit-only discordant units at $c_{\text{indel}} = 1$. † marks the degenerate ITS configuration.

Arm	Operation	Aligned ↑	$c_{\text{indel}} = 0$ ↑	$c_{\text{indel}} = 1$ ↑	Discordant
EXP	Identity	0.904	0.837	0.894	2 / 1
EXP	Crop 30%	0.904	0.827	0.846	6 / 0
EXP	Drop 30%	0.077	0.087	0.115	5 / 9
EXP	Shuffle	0.115	0.144	0.202	4 / 13
EXP	Round 0.5σ	0.346	0.010	0.019	35 / 1
ITS	Identity	0.856	0.000†	0.856	4 / 4
ITS	Crop 30%	0.875	0.000†	0.750	13 / 0
ITS	Drop 30%	0.038	0.000†	0.038	1 / 1
ITS	Shuffle	0.048	0.000†	0.038	1 / 0
ITS	Round 0.5σ	0.115	0.000†	0.029	10 / 1

Median runtime per unit/transformation is 61–90 seconds for EXP and 21–42 seconds for ITS, including 101 score evaluations. These edit-distance timings use the associated dynamic-programming implementation and differ from the aligned-detector timing in Appendix F.

I Component ablations and utility heterogeneity

I.1 Operating-point selection and aggregation

Table 12 gives the complete main-set detection and pooled-utility sweep. Table 21 adds the mean of unit-level CRPS ratios, the mean after excluding small control scores, and the 5%-trimmed mean. Twenty-three main-set units have control CRPS below 10^{-3}. These units receive little weight in the pooled ratio but can contribute large relative changes to an unweighted mean. At $\lambda = 5$, excluding them reduces the mean from +7.41% to +2.55%, while the median is +0.39% and the pooled change is +0.26%. The residual difference shows that small denominators explain part, but not all, of the heterogeneity.

The original development criterion selected the largest strength whose CRPS interval stayed below +2% and whose coverage interval stayed within ± 2 percentage points. Sensitivity to near-zero control scores motivated replacing the mean of unit-level CRPS ratios with the pooled ratio as the primary utility statistic. On the main set, the pooled, median, and 5%-trimmed summaries admit every tested strength from 3 to 8 under this criterion. The untrimmed mean admits none, and the mean excluding near-zero controls admits only $\lambda = 3$. The reported $\lambda = 5$ is an empirical point near the detection knee, rather than the largest strength admitted by the pooled criterion.

On finance/M4, increasing λ from 1.5 to 5 raises identity detection from 0.874 to 1.000 and joint survival from 0.638 to 0.849, while pooled CRPS moves from +0.02% to +0.18%. All five utility

Table 21: Sensitivity of mean relative CRPS to unit aggregation on the main set. All entries are percentage changes with 95% intervals. The excluded-control column removes 23 units with control CRPS below 10^{-3}.

λ	Mean ↓ [CI]	Controls $\geq 10^{-3}$ ↓ [CI]	5%-trimmed ↓ [CI]
3	+1.10 [+0.47, +2.11]	+0.67 [+0.43, +1.02]	+0.31 [+0.25, +0.38]
4	+4.24 [+0.87, +10.98]	+1.41 [+0.71, +2.44]	+0.49 [+0.38, +0.60]
5	+7.41 [+1.37, +18.21]	+2.55 [+1.05, +4.94]	+0.64 [+0.52, +0.78]
6	+10.70 [+2.38, +25.56]	+3.85 [+1.40, +7.18]	+0.72 [+0.58, +0.89]
7	+15.08 [+3.18, +35.03]	+5.39 [+1.56, +11.13]	+0.76 [+0.62, +0.93]
8	+18.07 [+3.92, +40.84]	+6.54 [+1.69, +14.07]	+0.78 [+0.64, +0.95]

summaries satisfy the stated criterion through $\lambda = 5$ on this set. On ETT, the untrimmed mean at $\lambda = 5$ is +1.09% [95% CI: +0.28,+2.33]. On LSF4, the corresponding coverage interval reaches -2.03 percentage points. The shared default exposes these differences rather than adapting the strength to each evaluation set.

I.2 Separating temporal feedback and local conditioning

The sender comparison fixes dev-64, the unmarked controls, a requested 128 forecast rows, $\lambda = 1.5$, the spread-based band rule, and keyed phases. Coupled generation feeds marked values into later predictions. FF evaluates base model logits at unmarked inter-value history and local digit prefixes; both FF and VFF compute watermark weights at the published prefix. VFF additionally recomputes the local units-digit model distribution after publishing the tens digit. Table 22 reports detection and mean per-unit CRPS changes with 2,000-draw intervals. Appendix I.3 gives ensemble sizes and the paired comparison using 10,000 source-series bootstrap draws.

Table 22: Sender ablation on dev-64 at $\lambda = 1.5$. Detection rates are proportions. CRPS is the mean per-unit relative change in percent, with its reported 95% interval from 2,000 bootstrap draws; Table 23 recomputes these intervals with 10,000 draws, so the bounds differ slightly. Coverage is in percentage points. This development statistic differs from the pooled main-set statistic.

Sender	Identity ↑	$Q_{.5}$ ↑	Mean CRPS ↓ [95% CI]	Coverage Δ
Coupled	0.922	0.703	+1.68 [-1.58, +6.15]	+2.29
FF	0.828	0.594	+0.38 [+0.13, +0.59]	+0.50
VFF	0.828	0.594	+0.37 [+0.07, +0.61]	+0.59

Coupled generation increases detection at this strength while also increasing forecast dispersion: its interval-width ratio is 1.077, compared with 1.000 for FF. In 40 development units with $H \geq 64$, the coupled-minus-FF owner-green-rate gap grows across horizon quarters from 0.020 to 0.059, 0.086, and 0.099. This progression is consistent with edits affecting later model contexts. FF and VFF have the same identity and rounding detection in the controlled comparison. Their distinguishing feature is the conditioning law within a published value, together with the batched local implementation used by VFF.

I.3 Controlled sender comparison

The transformation panels in Figure 1a,b use Qwen2.5-3B-Instruct on all 984 main-set evaluation units. An unavailable identity or transformation result counts as nondetection. Each comparator uses the most favourable of its prespecified detector readings, aggregated without an additional multiplicity correction. Detection intervals use 2,000 paired bootstrap draws over 581 source-series clusters, with the same resampled clusters for every method and transformation. The available-case comparisons in other tables retain their stated denominators.

These comparisons characterize the evaluated token-watermark and detector configurations, not an impossibility result for token watermarking. KGW and SynthID retain their identity decisions under row shuffling, whereas the aligned EXP and ITS detectors are sensitive to row operations. Table 20 reports an alternative detector configuration. V-band's row-order invariance retains original

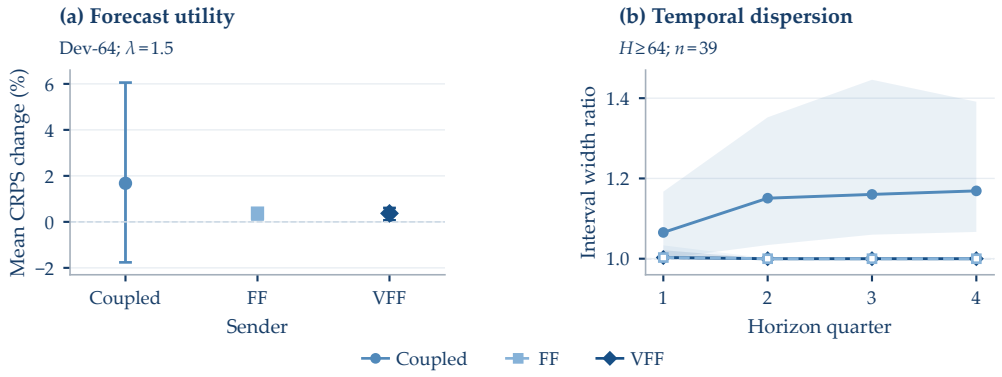

Figure 7: **Sender-conditioning ablation.** (a) Mean per-unit relative CRPS change for coupled, FF, and VFF generation on dev-64 at $\lambda = 1.5$. (b) Median 10–90% interval-width ratio to control by horizon quarter, on 39 units with $H \geq 64$ and nonzero quarter widths. The senders share controls, keys, and band widths and request $B = 128$. Error bars and shading give 95% source-series bootstrap intervals. The protocol below specifies ensemble sizes, aggregation, and the excluded zero-width unit.

forecast-step indices and does not assert invariance to arbitrary time permutation; its numerical tolerance also depends on whether edits cross band boundaries.

FF and VFF are compared on the same 64 development windows with shared unmarked control trajectories, observed contexts, keys, nonces, band widths, digit widths, and sampling settings. Both use the unmarked history for subsequent forecast steps. Ensembles contain 128 valid trajectories in every arm for 63 windows. In the remaining window, the control, FF, and VFF ensembles contain 120 trajectories and the coupled ensemble contains 114. All windows contribute to the utility comparison, and fair CRPS uses each ensemble's actual size.

Each sender samples its own marked trajectories, and comparisons are paired by evaluation window. FF and VFF have identical identity detection decisions (53/64), but only two units have exactly identical releases. Averaged over units, 31.34% of released numerical entries differ. In particular, FF already evaluates the watermark's band weights at the published prefix. VFF differs by additionally evaluating the underlying units-digit model logits under that prefix. It recomputes the units distribution after publishing the tens digit, using a batched update that shares the control-history cache, while retaining unmarked history across forecast steps. Similar observed utility or detection therefore does not establish equality of their sampling laws.

Table 23 distinguishes the mean of per-unit CRPS ratios from the pooled ratio of CRPS sums. Bootstrap draws resample the 42 (dataset, source series) clusters with replacement, retaining all windows in each selected cluster and pairing the draw across senders. We use 10,000 percentile draws with seed 20260913. The mean gives equal weight to each window's relative change, whereas the pooled statistic weights these changes by the unmarked CRPS. Reporting both statistics distinguishes average within-window relative changes from changes in the total forecasting score.

Table 23: Dev-64 sender utility with paired source-series bootstrap 95% intervals. Relative changes are percentages.

Sender	Mean per-unit CRPS change	Pooled CRPS change
Coupled	+1.68 [−1.76, +6.06]	+8.87 [−5.32, +20.16]
FF	+0.38 [+0.10, +0.58]	+0.02 [−0.04, +0.06]
VFF	+0.37 [+0.07, +0.60]	+0.03 [−0.03, +0.07]

Paired differences in the mean per-unit relative CRPS statistic are coupled minus FF: +1.30 [−2.11, +5.75]; coupled minus VFF: +1.31 [−2.11, +5.78]; and FF minus VFF: +0.01 [−0.05, +0.07] percentage points. All include zero. These results do not establish a significant CRPS improvement

from feedback isolation or an accuracy advantage from VFF's local-prefix consistency. Coupled identity detection is higher at this strength (59/64 versus 53/64 for either isolated sender).

Figure 7b restricts to $H \geq 64$ for a horizon-quarter diagnostic and excludes one of 40 such units because its control interval has zero width throughout a quarter. The remaining 39 units comprise 27 source-series clusters and 3,712 forecast steps. The coupled median interval-width ratio to control increases from 1.065 in the first quarter to 1.169 in the last, while FF and VFF remain approximately 1.000. Quartiles refer to forecast position within each unit, not to separate horizon datasets. These ratios describe changes in dispersion and do not establish improved calibration or a significant monotone temporal trend.

I.4 Digit schedule, phase, and width

Table 24 changes the schedule or phase on dev-64 at $\lambda = 3$. Units-only steering provides few opportunities to change band colour. Editing all digit places provides wider moves and higher joint survival, but the median per-unit CRPS change reaches +9.31%. Keyed phase increases survival under rounding and band offsets without changing identity detection.

Table 24: Schedule and phase ablations on dev-64 at $\lambda = 3$. Detection rates are proportions. Offset adds half a band width. Utility columns give pooled and median per-unit relative CRPS changes in percent, computed from each arm's frozen scored summary.

Configuration	Identity ↑	$Q_{.5}$ ↑	Offset ↑	Joint ↑	Pooled ↓	Median ↓
Keyed, band-capped	0.969	0.750	0.422	0.750	+0.13	+0.19
Fixed phase	0.969	0.672	0.375	0.672	+0.15	+0.43
Units only	0.297	0.078	0.031	0.047	-0.00	+0.00
All digit places	0.953	0.906	0.812	0.891	+1.38	+9.31

The width rule also matters at a fixed strength. On dev-64 with VFF and $\lambda = 2$, fixed $\Delta = 30$ quanta gives 92.2% identity and 51.6% rounding-0.5σ detection. The control-spread rule with ratio 0.40 gives 90.6% and 64.1%, respectively. At $\lambda = 3$, ratios 0.25 and 0.40 both give 96.9% identity, while rounding survival is 67.2% and 75.0%. This width comparison is separate from the phase ablation in Table 24.

Table 25: Band-cap and coarse-digit ablations at $\lambda = 4$. All CRPS values are mean per-unit relative changes in percent. The coarse multiplier applies to the newly reached hundreds place. Detection rates are proportions.

Set	Cap / coarse multiplier	Identity ↑	$Q_{.5}$ ↑	Joint ↑	CRPS ↓
LSF4	90 / default	0.976	0.643	0.643	+0.36
LSF4	300 / 1	0.988	0.798	0.798	+1.78
LSF4	300 / 0.5	0.976	0.714	0.702	+0.44
LSF4	300 / 0.25	0.964	0.679	0.667	+0.27
ETT	90 / default	0.881	0.655	0.655	+0.29
ETT	300 / 1	0.905	0.750	0.750	+1.42
ETT	300 / 0.5	0.905	0.738	0.738	+0.51
Finance/M4	90 / default	1.000	0.834	0.834	+0.89
Finance/M4	300 / 0.5	0.990	0.859	0.859	+1.74

Expanding the cap can help the high-spread LSF4 units, while the hundreds-digit edit adds forecast cost (Table 25). At $\lambda = 5$, the 300-quantum cap with half-strength hundreds-digit steering reaches joint survival of 0.859 on dev-64, 0.889 on finance/M4, and 0.810 on both ETT and LSF4. The corresponding mean per-unit CRPS changes are +3.23% (dev-64), +2.17% (finance/M4), +1.49% (ETT), and +0.78% (LSF4). Thus the larger cap changes the measured reuse survival–utility tradeoff differently across domains.

Two further schedules test how the edit budget is allocated. A tapered schedule applies decaying tilt to coarser places. At $\lambda = 4$, its mean per-unit CRPS increases are +13.36% on dev-64, +3.80% on ETT, and +4.53% on LSF4. A KL-budgeted tilt on ETT with $\kappa = 0.5$ and $\lambda_{\max} = 8$ gives 58.3%

identity detection on the full set and 79.6% on the diagnostic subset, at +0.12% mean per-unit CRPS. These ablations support restricting both the numerical edit scale and the allocation of local tilt.

I.5 Evidence quantity and common perturbation scale

The final main-set $\lambda = 5$ run gives 99.4% identity detection with 32 retained rows and 99.6% with all 64. At $\lambda = 3$, the corresponding rates are 93.6% and 94.4%. Finance/M4 at $\lambda = 5$ retains 100% identity detection with 32, 64, and 128 rows. These are within-release evidence reductions, so they do not measure the cost of generating a new ensemble at each row count. The same scorer preserves original forecast-step indices after cropping and makes no row-order contribution to the score.

Using the released ensemble's spread to set a perturbation amplitude can expose methods with wider marked ensembles to larger absolute perturbations. The common control-derived scale avoids this dependence on the watermark arm. A finance/M4 check using a 192-row unmarked pool gives the same $\lambda = 5$ identity and joint rates, 1.000 and 0.849, as the control-scale evaluation. This check establishes agreement at that operating point; the registered comparisons use the control scale throughout.

J Empirical tests of the local mechanism

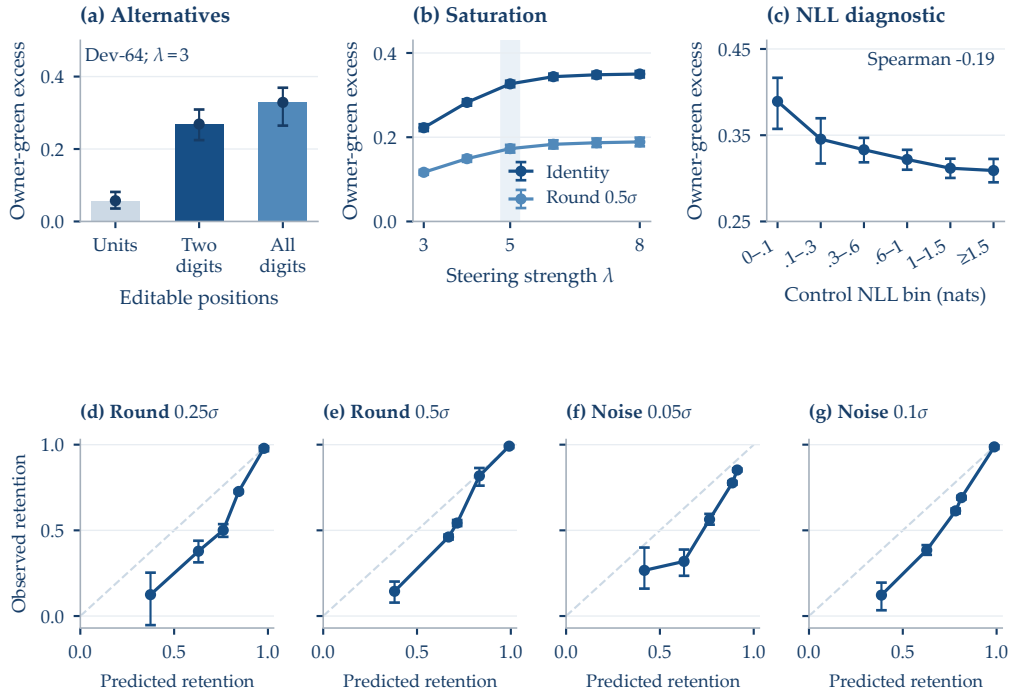

Figure 8: **Additional mechanism diagnostics.** Editable-digit interventions on dev-64 and the main-set strength sweep measure owner-green excess directly. The control-NLL diagnostic shows that mean excess does not increase with sampled-digit NLL. Panels (d–g) separately compare predicted and observed retention for each transformation, using five bins of predicted retention on 983 units with identity excess at least 0.05. Dashed lines denote equality. Intervals use 2,000 bootstrap draws over dataset/source series.

These diagnostics quantify how numerical alternatives and steering strength create watermark evidence. For the main strength sweep, mean identity excess at $\lambda = 3, 4, 5, 6, 7, 8$ is $0.223, 0.283, 0.327, 0.344, 0.348, 0.350$. On the 64 shared development units at $\lambda = 3$, units-only, two-digit, and all-digit schedules yield mean identity excess 0.057 $[0.036, 0.081]$, 0.269

$[0.225, 0.309]$, and 0.329 $[0.265, 0.369]$, respectively. These measurements intervene on accessible numerical moves; they do not reconstruct the full probability support of each model conditional.

J.1 Control NLL and evidence creation

The control-only NLL diagnostic averages the unmarked model's negative log probability at the edited digit places. The 0.3-nat threshold was fixed on dev-64 and does not filter the primary evaluation. Table 26 shows the measured green excess on all main-set units. Three of the four identity misses are below 0.3 nats, and one is above it. Mean excess decreases across the reported bins, with a reported Spearman correlation of -0.19 between NLL and excess. NLL therefore provides a diagnostic of the sampled digit law, not a monotone predictor of watermark strength.

The local capacity theorem concerns supported alternatives and their keyed colours. A sampled-digit NLL is a different quantity, and nucleus truncation can change the effective support. The strong-tilt limit also explains why a larger NLL need not yield a larger gain once the available green alternatives already carry nearly all tilted mass. The bin results are consistent with this saturation mechanism without identifying the complete support of each conditional law.

Table 26: Main-set mechanism diagnostic at $\lambda = 5$: green excess and identity detection by control-NLL bin. NLL is measured in nats.

NLL interval	n	Owner green excess ↑	Identity ↑
$[0, 0.1)$	66	0.389	0.955
$[0.1, 0.3)$	89	0.345	1.000
$[0.3, 0.6)$	214	0.333	1.000
$[0.6, 1.0)$	235	0.322	1.000
$[1.0, 1.5)$	252	0.312	0.996
$[1.5, \infty)$	128	0.309	1.000

J.2 Numerical perturbations and retained evidence

The retention diagnostic uses the 983 main-set units whose identity owner-green excess is at least 0.05. For each transformed release, the observed retention is its green excess divided by its identity excess. The prediction evaluates Equation 36 using the same unit's band width and control stepwise standard deviations, then averages across forecast steps. Table 27 compares their means and between-unit correlations.

Table 27: Idealized and observed retention of owner green excess on 983 main-set units with identity excess at least 0.05. The observed quantity divides transformed excess by identity excess. Correlations are Pearson correlations across units.

Transformation	Predicted	Observed	Correlation
Round 0.25σ	0.832	0.706	0.72
Round 0.5σ	0.679	0.490	0.69
Gaussian 0.05σ	0.892	0.804	0.78
Gaussian 0.1σ	0.791	0.648	0.76

The conditional model predicts the ordering of the four transformations, and Pearson correlations are 0.69–0.78 across units. Pooling the four transformations and binning predicted retention into $[0, .5)$, $[.5, .7)$, $[.7, .8)$, $[.8, .9)$, and $[.9, 1]$ gives observed mean retention of 0.14, 0.45, 0.58, 0.73, and 0.88. The monotone binned pattern supports the relevance of displacement relative to band width.

Every mean prediction exceeds the observed value, so Equation 36 is best read as an optimistic envelope rather than a calibrated predictor. Uniform within-band positions and constant within-band excess are idealizations of digitwise steering: fractional completion weights bias sampled values toward band boundaries, where small displacements cross into fresh bands more often than a uniform position predicts. The retention approximation further assumes that displacement is independent of within-band position. The observed optimism does not establish a universal upper bound or a guarantee against adaptive removal. Retention here measures the ratio of green excesses, not detection

probability; the diagnostic retains changes in the valid scored cells after transformation. Forecast utility also requires its own measurement: editing tens and units while copying higher digits can move a serialized integer by as much as 99 quanta, so Δ alone is not a pointwise displacement bound.

J.3 Why cellwise z-scores are not calibrated

A numerical ensemble can contain many cells that reuse the same keyed evidence. If M valid values contribute to the owner score S_0, treating their green indicators as independent fair Bernoulli trials gives

$$z_{\mathrm{ind}} = \frac{S_0 - M/2}{\sqrt{M/4}}. \tag{50}$$

Under the null assumptions of Theorem 1, the release is independent of the evaluated key functions. Equation 41 then gives $\mathrm{Var}(S_k \mid X) = D(X)/4$ for a fixed release and an independent ideal key. Thus

$$\mathrm{Var}(z_{\mathrm{ind}} \mid X) = \frac{D(X)}{M}, \tag{51}$$

where $D(X)$ sums the shared-band contributions defined in Equation 40. Values at the same forecast step can consult the same band bit, so counting them as independent trials understates the null variance. Conditional on the phase, a band occupied m times contributes $m^2/4$ to the score variance, rather than $m/4$. More forecast rows therefore need not supply proportionally more independent evidence.

Two fixed-width calibration settings at $\lambda = 0$ give z standard deviations of 3.422 and 4.546, exceeding the standard-normal value of one (Table 28). At threshold 2.326, the observed exceedance fractions are 41.27% and 58.29%. The summaries do not specify whether thresholding used z or $|z|$; either normal benchmark, approximately 1% or 2%, is far below the observed rates.

Table 28: Null diagnostics for fixed-width V-band at $\lambda = 0$. The z exceedance fractions use threshold 2.326; the stored summaries do not record whether z or $|z|$ was thresholded, and either normal reference (about 1% one-sided or 2% two-sided) is far exceeded. Rank decisions compare 4,096 keys, and the last column is a percentage. These settings precede the current spread-based band rule.

Setting	Units	Δ (quanta)	z SD	z exceedance (%)	Rank ≤ 40 (%)
Development	63	20	3.422	41.27	0.000
Confirmation	199	30	4.546	58.29	0.503

The decoy-key test ranks complete scores on the same release, retaining their dependence across cells. Its finite-sample null bound follows from key exchangeability (Theorem 1) without a normal approximation. The historical rank fractions describe the earlier samples; the current evaluation is reported separately below.

J.4 Null observations across evaluation sets

Table 29 reports the unmarked-control observations for the fixed owner key and eight fresh keys. The nominal guarantee conditions on the release and averages over exchangeable keys. It does not impose a false-positive ceiling on each dataset for every fixed key. Fresh-key evaluations share controls, so their total count is a descriptive aggregate over key–unit pairs, not an assertion that all pairs are independent Bernoulli trials.

K Generator diagnostics and additional transfer experiments

Table 10 and Table 9 report detection and utility for the 26 finance/M4 checkpoints. Table 30 adds the control-NLL subset share, the 90th-percentile band width, and the fresh-key null observations. The default cap is 90 quanta. Several high-survival models have a 90th-percentile width below this cap, including Gemma-3-12B/27B and Qwen3-14B. This relation connects model-dependent forecast spread with the displacement-to-band-width diagnostic, while the nonmonotone size curves show that parameter count alone does not order reuse survival.

Table 29: Unmarked-control null observations at rank ≤ 40. Each fresh-key column uses eight keys in the owner position on the same controls. The nominal single-reading bound is $40/4096$.

Evaluation set	Units	Fixed owner positives	Fresh-key positives
Main	984	18/984	83/7872
Finance/M4	199	1/199	22/1592
ETT	84	0/84	8/672
LSF4	84	0/84	8/672
Monash/Darts	104	1/104	6/832

Table 30: Numerical diagnostics for the 26 finance/M4 checkpoints. Gate denotes the control-NLL subset share; $\Delta_{.9}$ is the 90th percentile band width in integer quanta. Null is the fresh-key rate at rank ≤ 40, from 1,592 key–unit evaluations per checkpoint.

Checkpoint	Gate (%)	$\Delta_{.9}$	Null (%)
Qwen2.5-0.5B-Instruct	79.9	90	1.01
Qwen2.5-1.5B-Instruct	84.4	90	1.01
Qwen2.5-3B-Instruct	92.5	90	1.38
Qwen2.5-7B-Instruct	91.5	90	0.88
Qwen2.5-14B-Instruct	93.5	86	1.01
Qwen2.5-32B-Instruct	92.0	86	1.13
Qwen3-0.6B	63.3	90	0.75
Qwen3-1.7B	80.4	82	1.07
Qwen3-4B-Instruct-2507	84.9	90	0.75
Qwen3-8B	91.0	90	0.75
Qwen3-14B	92.5	75	1.13
Qwen3-32B	93.0	85	0.75
Qwen3.5-2B	89.4	75	0.69
Qwen3.5-4B	91.5	68	0.88
Qwen3.5-9B	93.0	75	1.01
Qwen3.5-27B	89.9	88	0.94
gemma-3-1b-it	53.8	90	0.88
gemma-3-4b-it	84.4	81	1.07
gemma-3-12b-it	89.9	69	1.26
gemma-3-27b-it	88.9	72	1.32
Llama-3.2-1B-Instruct	79.4	90	1.38
Llama-3.2-3B-Instruct	82.9	90	1.01
Llama-3.1-8B-Instruct	95.5	90	0.75
Mistral-7B-v0.1	86.9	90	0.88
Ministral-8B-Instruct-2410	82.4	90	0.44
Mistral-Small-24B-Instruct-2501	94.5	75	0.94

K.1 Transfer beyond finance/M4

Table 31 reports additional transfer experiments on ETT and LSF4. Qwen2.5-1.5B was also evaluated on these two sets, complementing the checkpoints with complete finance/M4 reporting. Each model uses its own control ensemble and the same band rule at $\lambda = 5$. Qwen3.5-9B uses independent in-loop FF; the other rows use VFF.

On LSF4, Qwen3.5-9B gives +2.84% pooled CRPS [95% CI: +1.71,+4.43] at 85.7% joint survival. This sender/domain combination has a larger utility change than the finance/M4 range in the main text. For the additional main-set runs, Gemma-3-4B and Qwen3-8B have control-NLL subset shares of approximately 71% and 78%, gated identity rates of 0.990 and 0.996, and fresh-key null rates of 0.978% and 0.940% across 7,872 evaluations. The full-set rates in Table 11 retain the low-NLL units.

K.2 Numerical-format compatibility

The 26-checkpoint study includes models that sustain the declared numerical serialization. A separate gpt-oss-20b (OpenAI, 2025) format test produces 114/128 valid rows at $H = 24$, but only 1/128 at horizons of at least 64. Long digit runs replace the required value separators. These observations

Table 31: Additional cross-generator detection on ETT and LSF4, each with 84 units and 128 sampled rows. Rates are proportions at rank ≤ 40. Main-set generator results are given separately in Table 11.

Checkpoint	ETT identity ↑	ETT joint ↑	LSF4 identity ↑	LSF4 joint ↑
Qwen2.5-3B	0.952	0.750	0.988	0.738
Qwen2.5-1.5B	0.976	0.643	0.988	0.583
Qwen2.5-7B	0.952	0.738	0.976	0.667
Mistral-7B	0.964	0.798	0.988	0.690
Gemma-3-4B	0.857	0.667	0.952	0.655
Qwen3-8B	0.905	0.726	0.964	0.667
Qwen3.5-9B (FF)	0.952	0.774	0.988	0.857

describe an upstream format failure and are not reported as watermark-detection outcomes. They delimit the compatible-generator claim.

L Public scale recovery

L.1 Public scale search and its null observations

Table 32: Recovery under public decimal unit changes on the 984-unit main set. Plain-full counts nonparsable releases as undetected; plain-parsable reports its indicated subset. Scale search evaluates five candidates and uses rank ≤ 8. Quality scores are undefined until the numerical unit is restored.

Release	Plain-full ↑	Plain-parsable (n) ↑	Scale search ↑	Raw disp.
Identity	0.996	0.996 (984)	0.989	0
$\times 10$	0.013	0.049 (263)	0.989	150
$\times 0.1$	0.002	0.007 (290)	0.989	15

Table 33: Main-set null observations for fixed-unit and scale-aware verification. Fresh-key counts evaluate eight keys in the owner position on the same unmarked controls.

Null evaluation	Plain rank ≤ 40	Five-scale minimum rank ≤ 8
Fixed owner	18/984 (1.83%)	6/984 (0.61%)
Eight fresh keys	83/7872 (1.05%)	29/7872 (0.37%)

The scale search returns the declared original scale for 979/984 identity units, the factor-ten correction for 979/984 multiplied units, and the factor-one-tenth correction for 981/984 divided units. These choices and detection are separate outcomes. Five complete candidate evaluations require approximately five times a fixed-unit verification; this is an operation-count estimate, not a separate runtime measurement.

The five candidate rescalings are $10^{-2}, 10^{-1}, 1, 10, 10^{2}$. Each candidate uses the original public serializer metadata after applying its correction to the observed release. The minimum-rank decision uses threshold 8, so the union bound in Equation 16 gives the same nominal $40/4096$ ceiling. The candidate corrections can be dependent. Scale recovery changes the tested family of public transformations, and consequently incurs both the threshold adjustment and five evaluations.

Fixed-owner and fresh-key null counts in Table 33 are reported separately for the same reason as in Table 29. The 98.9% scale-aware detection rate is an empirical power measurement under decimal unit changes. This result concerns recovery from changes of numerical units under the stated five-factor search.

The recovery protocol is scoped to decimal place shifts, whose corrections re-quantize the release onto the original integer grid exactly. General continuous affine transformations, such as non-decimal unit conversions or z-score renormalizations, break the integer grid alignment behind Equation 2; covering them would require a continuous parameter search whose multiplicity the union bound in Equation 16 can no longer absorb at a useful threshold without a much larger decoy pool. Scale-aware

verification under arbitrary affine reuse therefore remains open, and the reported robustness claims do not extend to it.

M Distributional and paired views of the experiments

M.1 Numerical reuse across datasets and detection thresholds

Figure 9 resolves the aggregate main-set result into dataset-level responses and full empirical rank distributions. All 984 units and all nine registered operations contribute. Sorting and shuffling reproduce the identity ranks exactly; the heatmap displays identity once and includes both operations in the joint column. The worst-operation rank for a unit is the maximum of its nine ranks, so its empirical cumulative distribution gives joint survival at every rank threshold. Coarse rounding accounts for most of the gap between identity and joint detection, with particularly low retention on Traffic. The Traffic loss is a direct consequence of the band-width cap: the 90-quantum ceiling in Equation 1 binds on 62% of Traffic units (the highest share of any dataset), so the 0.5σ rounding grid there reaches a median of 125 quanta, 1.4 times the band width, and 17% of Traffic units sit in the regime where the grid is at least twice the band width and the triangular kernel in Equation 8 retains at most Δ/g of the evidence. The cap is what bounds physical displacement to two digit places (Proposition 3), so this is a designed tradeoff: bounded distortion on every series against rounding vulnerability on series whose volatility exceeds the capped band scale.

M.2 Reuse survival and forecast quality

Figure 10 compares all six main-set V-band strengths, $\lambda \in \{3, \ldots, 8\}$, with all thirteen evaluated post-hoc value-space configurations. These comprise HTW adaptations at $r \in \{0.1, 0.5, 1.0\}$ (van Schaik, 2024), unguided bounded-noise proxies standing in for Waltz at $\epsilon \in \{0.01, 0.02, 0.05, 0.10\}$ (Yuan et al., 2025), WGTD adaptations at $\delta \in \{0.01, 0.02, 0.03, 0.05\}$ (He et al., 2024), and the Ngo et al. adaptation at $\delta \in \{0.02, 0.03\}$ (Ngo et al., 2026). Every point uses the same 984 main units, the nine registered transformations, and pooled CRPS relative to its unmarked control.

The low-cost region separates the V-band strength sweep from the post-hoc configurations. For example, V-band at $\lambda = 5$ gives 75.9% joint survival at +0.26% pooled CRPS, while the bounded-noise Waltz proxy at $\epsilon = 0.10$ gives 76.8% at +1.82%. Increasing the HTW rewrite rate to 1.0 reaches 98.2% joint survival at +103.49% pooled CRPS. The full-range panel retains these high-cost configurations, while the second panel resolves costs below +3%.

M.3 Paired sender effects

The paired dev-64 experiment compares senders under the same context, control ensemble, keys, and local strength. Figure 11 shows the full unit-level coupled/VFF comparison and the three paired mean differences. The scatter exposes variation hidden by one aggregate CRPS number; the interval plot retains the pairing and source-series dependence. All paired intervals cross zero, consistent with the uncertainty reported in Appendix I.3.

N Controlled experiment on forecast feedback

We test how a localized watermark edit affects subsequent forecasts when it enters the autoregressive history. The experiment uses Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on the same 112 windows: eight 64-step windows from each of the 14 main-set datasets, selected by a fixed identifier-hash ordering. Each window supplies 128 observed context values. Both models generate 64 trajectories per window in two independent groups of 32.

One-digit intervention. We sample the first numerical value and retain its sign and higher-order digit prefix. Under this fixed prefix, we draw an unmarked units digit and a marked units digit using the same uniform variable. The marked conditional adds the KGW bias to the five green digits in the ten-digit vocabulary, with $\gamma = 0.5$ and $\delta = 2$ for the primary experiment. A second arm uses $\delta = 4$. KGW green lists are conditioned on the preceding tens digit. Values with no preceding tens

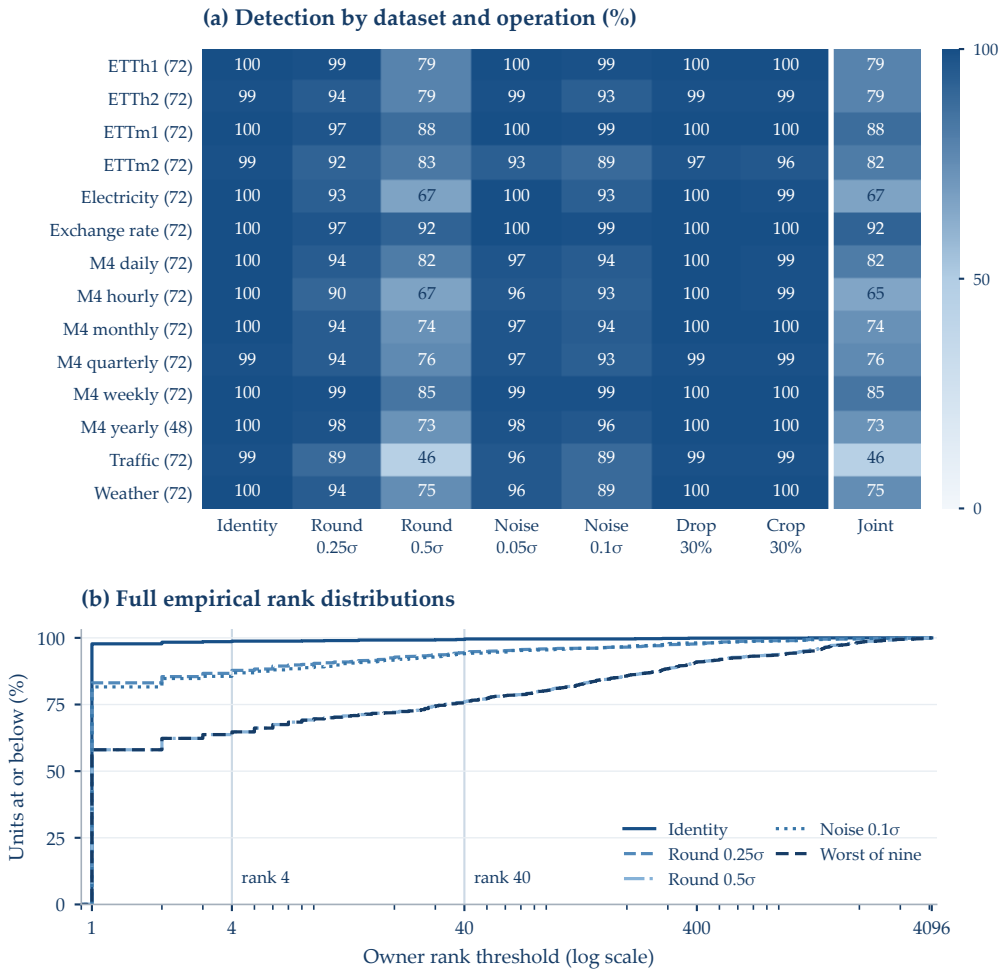

Figure 9: **Detection heterogeneity under numerical reuse.** (a) Detection percentages at rank ≤ 40, with dataset sample counts in parentheses. Cell annotations are rounded to whole percentages; colour uses the unrounded values. (b) Empirical cumulative distributions of owner ranks over all 984 units. The vertical lines mark ranks 4 and 40. The worst-of-nine curve uses the largest rank from the nine separately transformed releases for each unit. Dataset sources appear in Appendix A.2.

digit receive no edit and remain in the analysis. The two first values therefore differ by at most nine serializer quanta, or $9sq$ in the original units with $q = 10^{-3}$ and the context-derived scale s.

Paired continuations. One continuation conditions on the original first value, and the other conditions on its edited counterpart. Both release the same edited first value. Every later value is sampled without watermark steering. Uniform draws are indexed by window, replicate, row, forecast step, and token slot within the value, and are shared across the two continuations. Changes in digit count therefore do not shift the random draws assigned to subsequent forecast steps. Both arms use the same numerical grammar, temperature 0.7, top-p 0.9, and eight-digit cap. A zero-edit control generated with the same sampling randomness yields identical continuations in all 224 window–model comparisons. Trajectories with an unchanged initial value also reproduce their entire baseline continuation.

Effect size. Let X^0 denote the unmarked continuation and X^{fb} the continuation that reads the edited first value. For window i, σ_i is the standard deviation of the observed context. The normalized

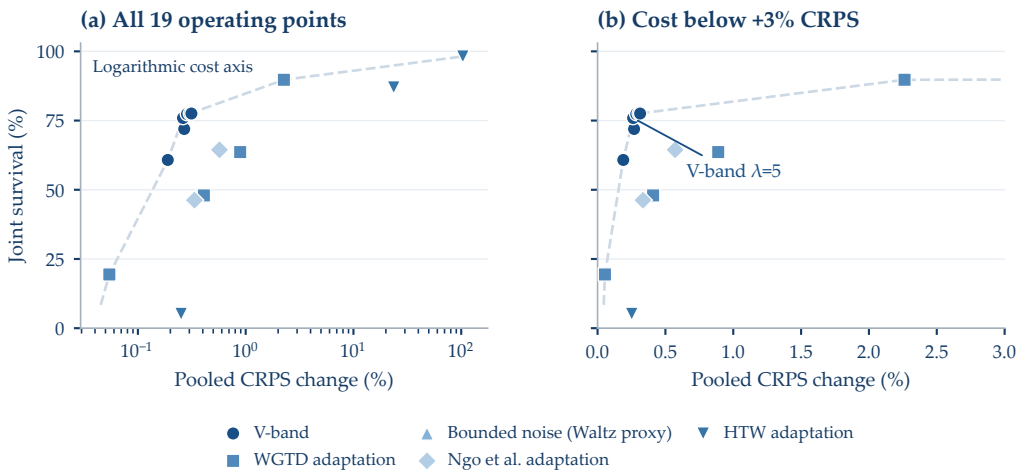

Figure 10: **Joint survival versus forecasting cost.** Each point is one measured configuration on 984 units. Higher survival and lower cost are preferred. (a) All 19 configurations on a logarithmic cost axis. (b) The same measurements with a linear axis restricted to costs below +3%. Dashed segments connect nondominated measured points as a visual guide; intermediate configurations have not been evaluated. The Waltz-proxy points are unguided uniform bounded noise, not the encoder-guided Waltz optimization; comparator adaptations are described in Appendix H.

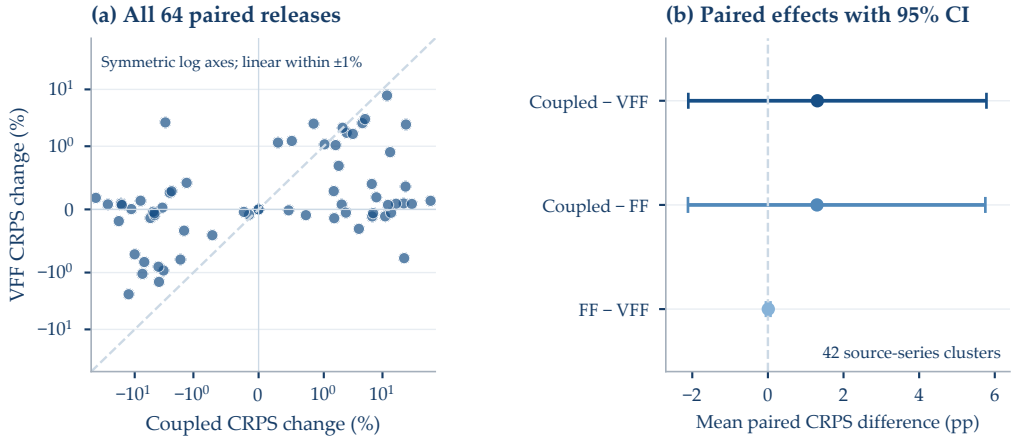

Figure 11: **Forecast quality in paired sender comparisons.** (a) Each point is one of the 64 development units at $\lambda = 1.5$. Both coordinates are relative CRPS changes from the same control; the dashed diagonal marks equality. Symmetric logarithmic axes retain negative and positive changes and are linear within $\pm 1\%$. (b) Mean paired differences in percentage points with 95% intervals from 10,000 bootstrap draws over 42 source-series clusters. Positive differences favour the second named sender.

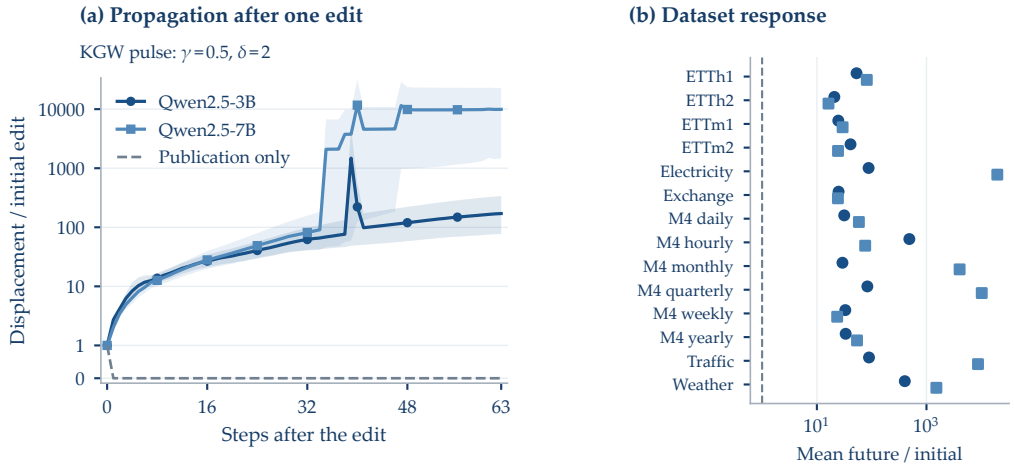

Figure 12: **Response to a single KGW edit.** (a) Paired displacement at each forecast lag divided by mean initial displacement, with 95% source-series bootstrap intervals. The vertical scale is linear up to one and logarithmic above it. Both models use the same 112 windows and primary $\delta = 2$ pulse. (b) Mean future-to-initial displacement ratio for each dataset. The vertical line marks a ratio of one.

paired displacement at lag h is

$$D_h = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B\sigma_i}\sum_{r=1}^{B}\left|X^{\mathrm{fb}}_{i,r,h} - X^{0}_{i,r,h}\right|, \qquad A = \frac{1}{D_0(H-1)}\sum_{h=1}^{H-1} D_h, \tag{52}$$

where $h = 0$ is the edited value, $H = 64$, and $B = 64$. The publication-only release uses the edited value at $h = 0$ and the baseline continuation thereafter. Its subsequent displacement is zero. The reported amplification A includes all prespecified windows and trajectories, including those whose initial edit is zero. A window is amplified when its mean future displacement exceeds its mean initial edit. A later value exceeds the edit bound when its displacement is greater than $9s_iq$. We also compare the sorted ensembles using empirical Wasserstein-1 distance and evaluate fair CRPS against the verified future targets. The CRPS comparison between feedback and publication-only releases holds the first published value fixed.

Table 34 reports aggregate effects, and Table 35 gives each dataset's response.

Table 34: Single-digit feedback intervention on 112 windows per model. Amplification is mean future paired displacement divided by mean initial displacement. W1 / initial uses the future ensemble Wasserstein-1 distance with the same denominator. Changed values are counted over forecast lags $h = 1$ to 63 after the edited step $h = 0$ (steps 2–64 in 1-based indexing). CRPS compares feedback with publication-only releases, which publish the same first value. Brackets give 95% source-series bootstrap intervals.

Model	δ	Amplification	W1 / initial	Changed (%)	CRPS change (%)
3B	2	97.46 [40.78, 184.48]	59.66 [17.34, 135.22]	41.4	−0.75 [−2.77, +1.12]
3B	4	86.19 [37.77, 162.11]	51.52 [15.40, 116.87]	44.5	−0.85 [−2.82, +0.99]
7B	2	3531.40 [548.48, 7689.99]	3431.76 [463.97, 7577.15]	36.9	+2.23 [−1.18, +6.05]
7B	4	3716.47 [858.15, 7589.50]	3631.78 [783.80, 7495.93]	40.6	+1.79 [−1.27, +5.36]

Intervals use 10,000 paired bootstrap draws over 92 source-series clusters, with seed 20260915. Every selected cluster contributes all its windows and the same draw is used for each arm. The generator uses float32 inference with TF32 disabled on NVIDIA H800 GPUs, PyTorch 2.5.1, and Transformers 4.46.3.

Table 35: Dataset-level feedback response at the primary KGW bias $\delta = 2$. Each dataset has eight windows per model and 64 trajectories per window. Amplification is the ratio of mean future displacement to mean initial displacement. Amplified windows have greater mean future displacement than initial displacement.

Dataset	Amplification		Amplified windows	
	3B	7B	3B	7B
ETTh1	52.9	81.4	7/8	5/8
ETTh2	20.8	16.3	8/8	8/8
ETTm1	24.6	29.5	7/8	5/8
ETTm2	41.1	24.3	7/8	6/8
Electricity	87.8	19625.9	8/8	8/8
Exchange rate	24.9	24.2	8/8	8/8
M4 daily	31.5	58.4	8/8	8/8
M4 hourly	486.8	76.1	6/8	5/8
M4 monthly	29.3	4026.1	8/8	8/8
M4 quarterly	83.4	10205.1	8/8	8/8
M4 weekly	33.0	23.5	8/8	8/8
M4 yearly	33.5	54.0	8/8	8/8
Traffic	89.3	8719.7	8/8	8/8
Weather	402.3	1516.7	7/8	6/8

What the intervention establishes. The measured effect is trajectory-level. Feeding the edit produces future-to-initial displacement amplification of $97.5\times$ and $3531.4\times$ and moves 37–44% of later values beyond the initial nine-quanta bound, while withholding it reproduces the baseline continuation exactly. In contrast, the fair-CRPS difference between the feedback and publication-only releases has 95% intervals that include zero for both models and both pulse strengths (Table 34), and the paired dev-64 sender comparison likewise finds no significant CRPS gap between coupled and isolated senders (Appendix I.3). Autoregressive feedback thus demonstrably changes which trajectories are released and widens their dispersion, without a CRPS degradation detectable at these sample sizes. VFF's contribution is accordingly deterministic rather than statistical: it confines published offsets to the insertion budget and keeps dispersion at the unmarked level (Proposition 3), instead of repairing a measured accuracy loss.