TOKAUDIT: AUDITING WHAT EXTRA CAPACITY BUYS IN TIME-SERIES TOKENIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tokenizer capacity is usually treated as a scalar budget, but reconstruction error does not reveal what extra capacity preserves. We introduce TokAudit, a measurement framework for auditing marginal capacity increases by separating reconstruction gains into latent-structure and observation-noise components, benchmarking their allocation against a matched reference, and measuring downstream utility independently. Across the tested tokenizer constructions, most of what extra capacity buys in reconstruction is observation noise rather than latent structure, while learned codecs are usually less noise-oriented than the matched reference. More noise-oriented allocation is also associated with smaller downstream utility gains, but this relationship is population-dependent. TokAudit therefore separates capacity growth, reconstruction composition, and task utility, reframing tokenizer scaling as an auditable allocation problem.

1 INTRODUCTION

In many time-series modeling pipelines, tokenizer capacity is fixed before downstream model training and is therefore costly to revisit. Capacity—how many discrete levels the representation may use—must balance two risks: too little can discard useful structure, while too much can preserve observation-specific detail that does not recur. Common diagnostics such as reconstruction error, vocabulary utilization, token occupancy, and downstream performance describe different aspects of this trade-off, but they do not reveal what an improvement in reconstruction is made of.

A lower reconstruction error can arise because a tokenizer recovers more of the latent state, or because it reproduces more of the observation noise. Utilization and occupancy indicate how much of the available code space is used, but not what the resulting representation preserves. Improved state fidelity need not translate into greater utility for a particular downstream task. These diagnostics therefore leave the composition of a marginal trained capacity gain unmeasured: how much capacity is added, what the resulting reconstruction gain preserves, and how downstream utility changes are distinct quantities.

We introduce TokAudit, a measurement framework that takes the *marginal tokenizer capacity transition* as its unit of analysis and makes the *marginal reconstruction gain* of a trained capacity increase an explicit object of measurement. For one tokenizer family T trained at capacities $C_1 < C_2$, TokAudit asks what the resulting reconstruction improvement consists of. In controlled environments that expose the latent state and permit observation-noise resampling, it decomposes the marginal reconstruction gain into state- and noise-related components, compares their allocation with a same-occupancy second-moment-matched Gaussian reference, and measures downstream utility independently (Figure 1). Complementing configuration-to-performance studies such as Roger et al. (2025), TokAudit measures the composition of reconstruction gains alongside independently measured downstream utility.

This distinction matters empirically. Across the tested tokenizer constructions, noise-dominated marginal reconstruction gains are the stable pattern: they persist across learned quantizers and tokenizer controls. In contrast, below-reference behavior and allocation–utility relations are population-dependent. Reconstruction improvement, its state–noise composition, and downstream utility therefore need not move together.

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.

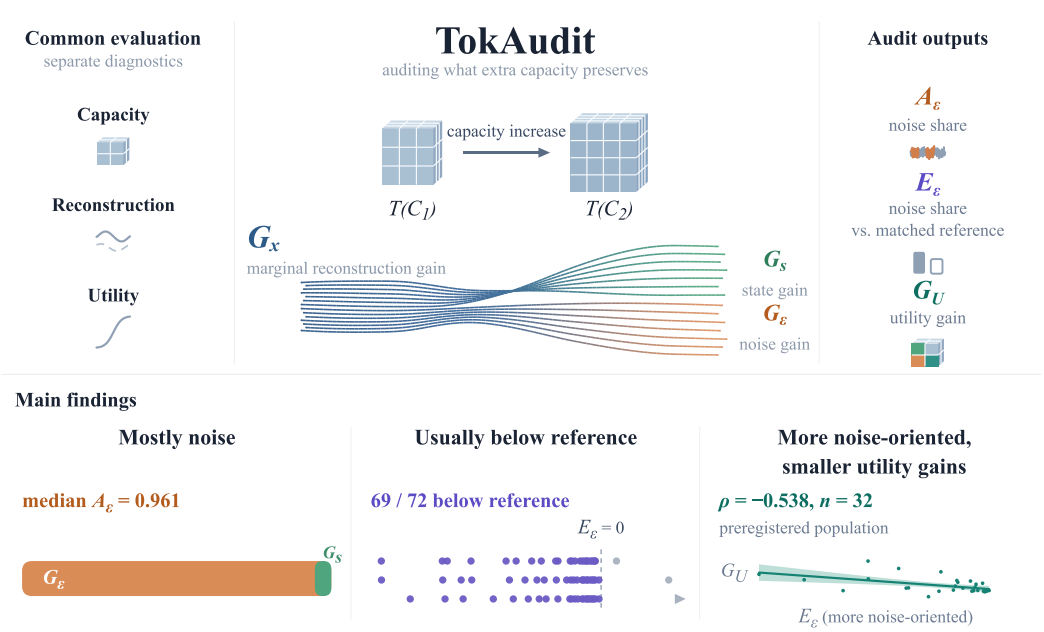


Figure 1: What extra tokenizer capacity preserves. TokAudit separates marginal reconstruction gains into state and noise, compares their allocation with a matched reference, and measures downstream utility independently. Across the studied populations, extra capacity is predominantly spent on noise; learned allocations are usually less noise-oriented than the reference, and more noise-oriented allocation is associated with smaller utility gains in the preregistered population.

Our contributions are threefold. First, we define the trained capacity transition $T(C_1) \to T(C_2)$ as the unit of analysis, separating how much capacity is added from the composition of its marginal reconstruction gain and the accompanying change in downstream utility. Second, we provide a reproducible measurement procedure combining a bias-corrected finite-replica estimator, a same-occupancy Gaussian reference, and an independent measurement of downstream utility. Third, we show that noise-dominated marginal reconstruction gains persist across the tested tokenizer constructions and controls, while reference-relative behavior and allocation–utility relations remain population-dependent.

2 Related Work

Discrete tokenization for time series. Our population spans three representative quantizer mechanisms: learned-codebook vector quantization (van den Oord et al., 2017), finite scalar quantization (Mentzer et al., 2024) and binary spherical quantization (Zhao et al., 2025). The third is instantiated as residual binary quantization (RBQ), a two-stage residual construction with a binary-spherical quantizer at each residual stage. Time-series foundation models tokenize by scalar binning, task-independent VQ-VAE tokens, localized frequency representations, hierarchical discrete targets, or byte-pair encoding (Ansari et al., 2024; Talukder et al., 2024; Masserano et al., 2025; Feng et al., 2025; Götz et al., 2026). We study trained capacity transitions across these mechanisms through a common measurement framework.

What is measured about a tokenizer. The closest empirical studies are Roger et al. (2025), who examine scaling, quantization, vocabulary size and pretrained versus random initialization, analyze ideal-predictor discretization error and relate token utilization to forecasting performance, and Götz et al. (2026), who compare vocabulary sizes within a forecasting pipeline. Both relate tokenizer configurations to downstream performance. TokAudit takes the trained capacity transition between two configurations of the same tokenizer as its unit of analysis, asking what the marginal reconstruction gain of that transition is made of, while measuring the accompanying change in downstream utility separately.

Rate–distortion, predictive information, and benchmarks. Predictive rate–distortion asks how compression trades off against predictive information (Hahn & Futrell, 2019), and the information bottleneck frames a related trade-off as a constraint (Tishby et al., 2000). Remote-source, or indirect, rate–distortion asks how well a compressed noisy observation can represent a hidden source (Wolf & Ziv, 1970; Kipnis et al., 2015). TokAudit addresses a complementary empirical question: once a codec has been trained to reconstruct the observation, where did the reconstruction gain from additional capacity actually go? Forecasting benchmarks standardize datasets, splits, metrics and evaluation protocols (Godahewa et al., 2021; Zhou et al., 2021; Wu et al., 2021); TokAudit standardizes the quantities used to characterize a trained capacity transition.

3 TokAudit: Measurement Specification

3.1 The audit transition

TokAudit takes the *audit transition* as its unit of analysis: a tuple $\tau = (T, C_1, C_2, \mathcal{E})$ comprising a tokenizer mechanism T, two capacities $C_1 < C_2$ at which T has been trained using the same recipe, and an environment $\mathcal{E}$. The capacity index is the nominal per-stage code width in bits. The ledger records the nominal step and the realized occupancy increment separately, since the two need not coincide. Training and snapshot-selection details are given in section A.2.

3.2 Conditions for measurement

TokAudit measures each transition in an environment $\mathcal{E}$ with $x_t = s_t + \varepsilon_t$, where the latent state is known and the observation noise can be resampled. These properties make the state–noise decomposition possible: access to the state allows the conditional-mean state distortion to be estimated rather than bounded, while resampling the observation noise lets the same state be replayed under fresh noise. We instantiate $\mathcal{E}$ with a regime-switching vector autoregression under additive Gaussian observation noise, evaluated across four experimental settings spanning a predictability-by-observation-noise square. Full generator, regime, and panel specifications are given in Appendices A.1–A.2.

3.3 The ledger

The ledger records three aspects of a transition: the reconstruction gain and its state–noise decomposition; the fraction of that gain assigned to noise relative to the matched reference; and downstream utility.

For a transition τ, the ledger is

$$\mathcal{L}(\tau) = \big(\Delta R_{\mathrm{nom}}, \Delta H, G_x, G_s, G_\varepsilon, A_\varepsilon, A_\varepsilon^{\mathrm{ref}}, E_\varepsilon, G_U\big). \tag{1}$$

ΔR_{nom} is the nominal per-token step in bits, twice the per-stage width increment for a two-stage codec, and ΔH is the realized joint-token occupancy increment in the same unit. Write f_C for the complete trained codec at capacity C — encoder, quantizers and decoder — and let all squared-error distortions be reported as per-channel means in the environment's standardized coordinates. Then

$$D_x(C) = \mathbb{E}\,\|s + \varepsilon - f_C(s+\varepsilon)\|^2, \qquad D_s(C) = \mathbb{E}\,\|s - \mathbb{E}_\varepsilon f_C(s+\varepsilon)\|^2, \tag{2}$$

with $D_\varepsilon(C) = D_x(C) - D_s(C)$. The transition gains are $G_x = D_x(C_1) - D_x(C_2)$, and likewise for G_s and G_ε, so $G_x = G_s + G_\varepsilon$ exactly.

D_s is the conditional-mean state distortion — the error of the codec's noise-averaged reconstruction rather than of any single reconstruction. A vanishing state account is therefore not a denoised reconstruction: the identity codec has $D_s = D_\varepsilon = D_x = 0$ while reproducing the observation noise exactly. G_s measures a fall in conditional-mean state distortion, not a gain in usable state information. D_s and D_ε are estimated with a bias-corrected finite-replica estimator over resampled observation noise, with corrected estimates retained rather than clipped (section B.1).

The marginal noise share is $A_\varepsilon = G_\varepsilon / G_x$. It is a ratio of two reconstruction gains, not a share of ΔR_{nom} or ΔH. It is defined where $G_x > 0$ and is not clipped to $[0, 1]$: $A_\varepsilon > 1$ exactly when $G_s < 0$, a measurement outcome rather than an error.

To interpret A_ε, the ledger places it against a direct observation-reconstruction second-moment-matched Gaussian reference at the same realized occupancies and computes the same transition-level noise gain share under that reference. Write $H(C)$ for the codec's realized joint-token occupancy at nominal capacity C; the reference is evaluated at $H(C)$, not at C:

$$A_\varepsilon^{\mathrm{ref}}(\tau) \;=\; \frac{D_\varepsilon^{\mathrm{ref}}(H(C_1)) - D_\varepsilon^{\mathrm{ref}}(H(C_2))}{D_x^{\mathrm{ref}}(H(C_1)) - D_x^{\mathrm{ref}}(H(C_2))}, \qquad E_\varepsilon = A_\varepsilon - A_\varepsilon^{\mathrm{ref}}. \tag{3}$$

$A_\varepsilon^{\mathrm{ref}}$ (Eq. (3)) reports what fraction of its reconstruction gain a second-moment-matched Gaussian object would allocate to noise over the same two occupancies. It is an interpretive baseline, not an optimal codec, and "below the reference" is an observation about this population rather than a distance to an optimum (Appendix B.4). State-information capture uses a distinct indirect, remote-source reference (section F.2). Appendix B.2 summarizes the ledger quantities and their interpretation.

3.4 Utility as an independent account

G_U is the increment in downstream utility between the two capacities. In the primary analysis, utility is the per-step cross-sectional Spearman rank correlation (RankIC) of a ridge readout that predicts the next-step state from a four-step latent history; we report the first state channel, and section 6 varies the reader family. Because task utility depends on what a reader can extract, G_U is a *readout quantity*, reported separately from the reconstruction decomposition. Implementation and account details are given in Appendices A.2 and B.3.

4 Estimator Stability and Calibration

Replica estimates converge with budget: 9 of the 18 calibration transitions meet the operational tolerance at 32 replicas, 14 at 128, and all 18 do at 512 relative to the 2048-replica endpoint. The measured noise account recovers the known observation-noise ordering and changes monotonically with added training noise; Gaussian and distortion-matched controls preserve the reference-relative ordering. A noiseless analytic case yields a vanishing noise account, while a stricter finite-replica accuracy criterion is not met at the finest rung: 5 of the 48 component comparisons deviate from the closed-form value by more than one Monte Carlo standard error (Appendix C; corrections in Appendix H.2).

5 What Extra Capacity Buys

Most of what extra capacity buys in reconstruction is observation noise, not latent structure. We measure this allocation, compare it with a matched reference, and relate it to downstream utility. Table 1 lists the populations the principal analyses draw on, with full grids and replication schemes in Table 4; the specification status of every study is defined in Appendix H.1.

population	analysis	trans.
calibration	measurement validation and core allocation	18
extended	allocation characterisation	54
prognostic	allocation–utility endpoint	32
cross-generator	dynamics robustness	24
frozen re-read	reader robustness	56
operational	observational-data construction	4
external	real-data stress test	9
multi-site	multi-site external endpoint	72
matched contrast	allocation discrimination	2

Table 1: Populations and controlled contrasts used in the principal analyses. Each row gives the population and the analysis it supports; counts are audit transitions. Full grids and replication schemes are in Table 4; specification status is reported in Table 11.

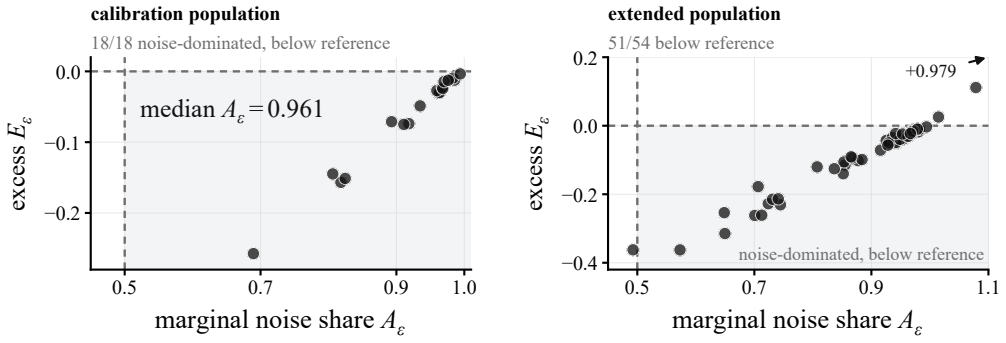


Figure 2: Extra capacity is mostly spent on noise, but usually less so than the matched reference. Measured marginal noise share A_ε against the reference-relative excess $E_\varepsilon = A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$, for the calibration (left) and extended (right) populations. Dashed guides mark one half of the gain (vertical) and the matched reference (horizontal); the single off-axis transition in the extended population is retained and labelled.

5.1 Most reconstruction gain goes to noise

For the typical capacity increase, more than ninety-five per cent of the reconstruction gain is observation noise rather than latent-state recovery (calibration-population median marginal noise share 0.961). Noise dominance strengthens with observation noise and remains in nearly every low-noise transition (Table 8). Reconstruction error alone cannot reveal this composition.

5.2 Learned codecs are noise-dominated and usually below the reference

Noise dominance is an absolute statement about composition; comparing A_ε with its matched reference gives a relative statement about allocation. The learned codecs are noise-dominated, but less so than the matched reference: all 18 calibration transitions and 51 of 54 extended transitions lie below it (Figure 2; Table 7). This comparison concerns the composition of reconstruction gains, not how much observation noise could in principle be removed.

5.3 Mechanisms use capacity differently

The three quantizers use added capacity differently. Over their shared capacity range, VQ-VAE adds the least realized occupancy (0.44–0.48 bits against 1.73–1.76 for RBQ and 1.90–2.03 for FSQ) and has the smallest utility increment in the two high-predictability regimes, where that increment is negative; section D.3 prints the regime-by-regime contrast. Less added occupancy does not imply more useful capacity. Targeted controls weaken several implementation-specific explanations for these differences (section D.3).

5.4 The negative allocation–utility relation is population-dependent

In an independent preregistered population, more reference-relative noise allocation is associated with smaller downstream utility gains (Spearman $\rho = -0.538$; Figure 3). A post-hoc conditioning on ladder position and realized occupancy leaves the association essentially unchanged in this population (section E.1), but not in the cross-generator population of section 6. We therefore treat the relation as population-dependent rather than a general property of tokenizer scaling.

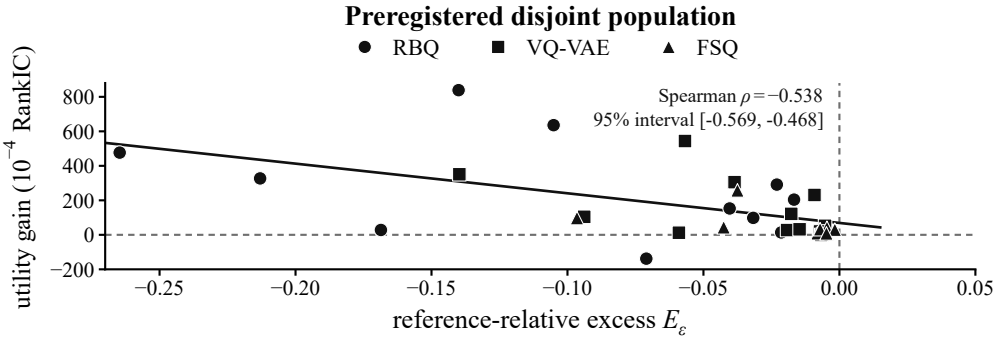


Figure 3: More reference-relative noise allocation is associated with smaller utility gains in the preregistered population ($\rho = -0.538$, 95% interval [-0.569, -0.468], $n = 32$). The line is an OLS visual guide; inference uses Spearman correlation.

6 Robustness and Limits across Controls, Dynamics, and Readers

6.1 Noise dominance survives tokenizer controls

A training-free scalar-binning control shows the same strong noise-dominated reconstruction pattern. It does not, however, consistently fall below the matched reference or reproduce the negative allocation–utility association (Figure 5, left; section E.4).

Adding a limited causal temporal receptive field leaves the allocation essentially unchanged: the median remains strongly noise-oriented and all transitions remain below the reference, while the difference from the pointwise encoder is inconclusive (Figure 5, right). The pattern survives this limited temporal extension. This control does not establish the same result for patch-based, convolutional, or frequency-domain tokenizers (section E.4).

6.2 The audit separates allocation at matched reconstruction and occupancy diagnostics

A paired training comparison asks whether the audit adds discriminating power rather than a better predictor. Two conditions shared one mechanism, one capacity span, one budget and one set of initializations; the second added a small term that pulls two independent observation-noise views of the same state toward one reconstruction, and the latent state was never a training label. On the development panels the two satisfy the matching bands the study predeclared — practical bands and a deterministic decision rule, not a formal equivalence test: reconstruction gains differ by 0.7% and realized occupancy increments by 0.0006 bits, with the two capacities' reconstruction-error levels within 2.6-4.6%. Their allocations are not the same. On the test panels the measured marginal noise share falls from 0.772 to 0.734, a paired change of -0.038 (95% interval $[-0.048, -0.027]$), while the utility gain moves by -0.00003 rank information coefficient. Within this design the audit therefore reads a difference that reconstruction and occupancy do not show (section E.6). One generator, one mechanism, one capacity span and one intervention bound the claim.

6.3 The within-stratum relation persists across dynamics and readers

Across nonlinear and seasonal dynamics, the allocation–utility relation remains negative within strata (Figure 4), and every transition lies below its matched reference (section E.2); the first panel's ridge row is the pooled preregistered ρ, the other rows are stratified predeclared direction tests. The same within-stratum relation persists under nonlinear readers, which also preserve the capacity ordering observed with the linear reader (section E.3).

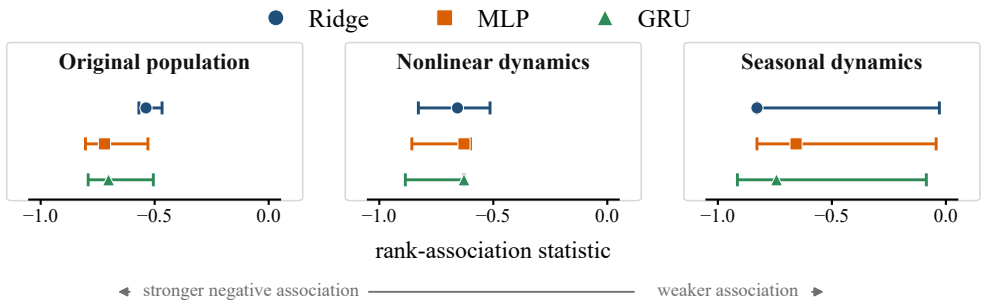


Figure 4: Allocation–utility associations remain negative across the tested dynamics and readout families. Points and bars show the rank-association estimate and interval for each setting; because the estimand varies by setting, only the direction is directly comparable. Appendices E.1–E.3 give the conditioning, dynamics, and reader analyses.

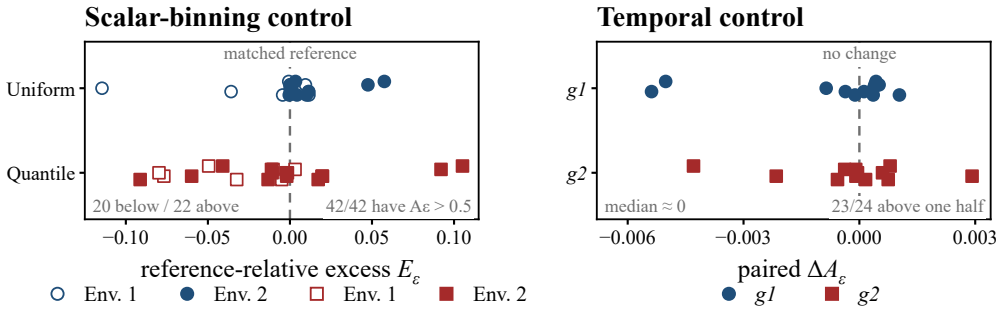


Figure 5: Noise dominance survives both tokenizer controls. Scalar binning (left) does not consistently fall below the matched reference, while the causal temporal encoder (right) leaves measured allocation nearly unchanged. Dashed lines mark the matched reference and zero paired change, respectively; full endpoints are reported in Appendix E.4.

6.4 The pooled association does not survive conditioning

The pooled cross-generator association disappears after conditioning on ladder position and realized occupancy. The cross-generator evidence therefore supports a consistent within-stratum pattern, not a generator-invariant prognostic relation (section E.1).

7 External Stress Tests on Real Dynamics

On the Beijing multi-site air-quality series, we inject known Gaussian observation noise while retaining the observed multivariate dynamics, with twelve stations as the cross-section and train-only global normalization. The prespecified stratified endpoint reproduces the negative association ($Q = -0.501$; 95% interval $[-0.574, -0.138]$). This extends the result from synthetic generators to real multivariate dynamics under identifiable injected noise (section G.1).

A separate stress test injects known corruption into the Monash Saugeen series; its preregistered endpoint is inconclusive, and differences in training budget and provenance prevent a like-for-like replication claim (section G.2).

8 Discussion and Conclusion

TokAudit makes the marginal reconstruction gain of a trained capacity increase an explicit object of measurement. Across the tested tokenizer constructions—three learned quantizer mechanisms, a training-free scalar-binning ladder and a limited causal temporal encoder—most of that gain is assigned to observation-specific noise rather than latent-state recovery. The headline allocation is not confined to learned quantizers or a strictly pointwise encoder.

The finer patterns are conditional. Below-reference behavior and the negative allocation–utility association do not persist uniformly across populations. We therefore treat them as empirical properties of the tested populations rather than universal consequences of tokenizer scaling.

TokAudit separates how much capacity is added, what the reconstruction gain preserves, and how downstream utility changes. The decomposition applies when latent state and observation noise are identifiable or explicitly specified. Within that scope, TokAudit provides a reproducible measurement of what additional tokenizer capacity preserves. Together, these quantities expose a dimension of tokenizer scaling that reconstruction error, code utilization, and downstream utility alone do not reveal.

AI Use Statement

Generative AI tools were used extensively in this work. We used them to assist with research methodology and experimental design, implementation and validation of experimental pipelines, analysis and interpretation of experimental results, literature search, and manuscript drafting and revision. In particular, the initial validation scripts and substantial parts of the experiment authoring, monitoring, analysis, and reporting pipeline were developed with AI assistance; the frozen-bundle and generator machinery described in section I was also built with such assistance. Generative AI was additionally used to locate relevant literature and to revise the presentation of the manuscript. All experimental outputs, implementations, reported numerical results, citations, and scientific conclusions were reviewed and checked by the authors. The authors take responsibility for the final content of this work, including all AI-assisted text, code, analyses, and artifacts.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Shibo Feng, Peilin Zhao, Liu Liu, Pengcheng Wu, and Zhiqi Shen. HDT: Hierarchical discrete transformer for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 746–754, 2025. doi: 10.1609/aaai.v39i1.32057.

Rakshitha Godahewa, Christoph Bergmeir, Geoff I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

Leon Götz, Marcel Kollovieh, Stephan Günnemann, and Leo Schwinn. Byte pair encoding for efficient time series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2505.14411.

Michael Hahn and Richard Futrell. Estimating predictive rate-distortion curves via neural variational inference. *Entropy*, 21(7):640, 2019. doi: 10.3390/e21070640.

Alon Kipnis, Stefano Rini, and Andrea J. Goldsmith. Indirect rate-distortion function of a binary i.i.d source. arXiv:1505.04875, 2015.

Luca Masserano, Abdul Fatir Ansari, Boran Han, Xiyuan Zhang, Christos Faloutsos, Michael W. Mahoney, Andrew Gordon Wilson, Youngsuk Park, Syama Sundar Rangapuram, Danielle C.

Maddix, and Bernie Wang. Enhancing foundation models for time series forecasting via wavelet-based tokenization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43248–43275. PMLR, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Alexis Roger, Gwen Legate, Kashif Rasul, Yuriy Nevmyvaka, and Irina Rish. Small vocabularies, big gains: Pretraining and tokenization in time series models. *arXiv preprint arXiv:2511.11622*, 2025. Presented at the AAAI 2026 AI4TS Workshop.

Sabera J. Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: Tokenized time series embeddings for general time series analysis. *Transactions on Machine Learning Research*, 2024. `https://openreview.net/forum?id=QlTLkH6xRC`.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jack K. Wolf and Jacob Ziv. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Transactions on Information Theory*, 16(4):406–411, 1970.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations*, 2025.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

NOTATION AT A GLANCE

symbol	meaning
$\tau = (T, C_1, C_2, \mathcal{E})$	audit transition
$x_t = s_t + \varepsilon_t$	observation, latent state, observation noise
f_C	trained codec at capacity C
$H(C)$	realized joint-token occupancy
D_x, D_s, D_ε	observation, state and noise-related distortions
G_x, G_s, G_ε	corresponding marginal reconstruction gains
A_ε	measured marginal noise share G_ε / G_x
$A_\varepsilon^{\mathrm{ref}}$	same-occupancy matched-reference noise share
E_ε	$A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$
G_U	downstream utility gain
ρ, τ_b	Spearman and Kendall rank correlations (the latter distinct from the transition τ)
Q, Q_g	study-specific stratified association endpoints
m, d	number of antithetic pairs and channel dimension

Table 5 gives the fuller interpretation of the ledger entries; local symbols are defined where they are used.

GUIDE TO THE APPENDIX

The appendix follows the measurement workflow, then presents the supporting studies. Study status and corrections are collected at the end; conditions needed to interpret a result remain with that result.

main-text topic	supporting material
Experimental settings and training	Sections A.1–A.3; Table 4
Measurement and uncertainty	Sections B.1–B.5
Estimator validation	Section C
Noise allocation and mechanism differences	Tables 7–8; Section D.3
Conditioning, dynamics and readers	Sections E.1–E.3
Tokenizer and matched-training controls	Sections E.4–E.6
Utility analyses and selection	Sections F.1–F.4
External studies	Sections G.1–G.2
Study status, corrections and reproduction	Sections H.1–I

A Experimental setup and populations

A.1 Regime definitions

The core generator is a regime-switching vector autoregression with additive Gaussian observation noise, a latent state known by construction and resampleable noise:

$$s_{i,t} = s_{i,t-1}A_{r(t)}^{\top} + \lambda\,\ell_i f_t + \xi_{i,t}, \qquad \xi_{i,t} \sim \mathcal{N}(0, \sigma_\xi^2 I), \tag{4}$$

$$f_t = \rho_f f_{t-1} + 0.6\,\eta_t, \qquad \eta_t \sim \mathcal{N}(0,1), \tag{5}$$

$$x_{i,t} = s_{i,t} + \varepsilon_{i,t}, \qquad \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I), \tag{6}$$

with $s_{i,t}, x_{i,t}, \varepsilon_{i,t}, \xi_{i,t} \in \mathbb{R}^4$ and $\ell_i \in \mathbb{R}^4$. The index i runs over entities within a panel; all vectors are row vectors. Each panel has its own scalar factor path f_t, shared across that panel's entities. The regime index is $r(t) = \lfloor t/64 \rfloor \bmod 2$, and the two transition matrices A_0 and A_1 are diagonal, each scaled by that setting's state scale ς. Observation noise is scaled so that the raw noise energy equals the signal energy divided by the target signal-to-noise ratio, so predictability and observation-noise level vary independently through (ς, ρ_f) and the target SNR.

The four regimes complete a predictability-by-observation-noise square. The generator's state scale ς and factor persistence ρ_f set predictability, and the target SNR sets the observation-noise level. Table 2 reports the realized noise fraction alongside the nominal one, because the two differ. Each regime trains on 8 panels. The internal tag `lowSNR` denotes high observation noise: at $\mathrm{SNR} = 0.1$, roughly ninety per cent of the observation is noise. The table uses observation-noise levels throughout.

Regime	ς	ρ_f	target SNR	realized f_ε
low predictability, high obs. noise	0.65	0.55	0.1	0.9068
high predictability, high obs. noise	1.10	0.90	0.1	0.9013
high predictability, low obs. noise	1.10	0.90	1.0	0.5749
low predictability, low obs. noise	0.65	0.55	1.0	—

Table 2: Regime definitions, with the generator's remaining constants: the state matrices are $\varsigma\,\mathrm{diag}(0.85, 0.70, 0.55, 0.40)$ and $\varsigma\,\mathrm{diag}(0.55, 0.40, 0.25, 0.10)$; loadings are $\ell_i \sim \mathrm{Unif}(0.5, 1.5)^4$; the factor enters at $\lambda = 0.2$ with innovation $0.6\,\eta_t$; the state innovation is $\sigma_\xi = 0.25$; and the burn-in is 128 steps. The fourth cell completes the square and is used only by the preregistered population of section 5.4. Its realized fraction was not measured in the synthetic core and is shown as a dash.

A.2 The synthetic core: codec, budget and optimizer

The three learned quantizer families share the architecture, training budget and readout described here. The regime definitions are in section A.1, and Table 4 gives the capacity grids and replication schemes.

Codec. The environment emits one 4-dimensional observation per time step, and the codec is applied per time step rather than to a window. The three mechanisms share one MLP skeleton and differ only in the two residual quantizers between its halves:

$$z = E(x), \qquad z_1 = q_1(z), \qquad z_2 = q_2(z - z_1), \qquad \hat{x} = D(z_1 + z_2), \tag{7}$$

$$E = \mathrm{LN} \circ \mathrm{Lin}(192 \to 64) \circ \mathrm{GELU} \circ \mathrm{LN} \circ \mathrm{Lin}(4 \to 192), \tag{8}$$

$$D = \mathrm{Lin}(192 \to 4) \circ \mathrm{GELU} \circ \mathrm{LN} \circ \mathrm{Lin}(64 \to 192). \tag{9}$$

The 2 residual stages quantize the 64-dimensional latent in sequence, and the decoder reconstructs from the sum of the two stage codes. A capacity is a per-stage width w applied to both stages, so a transition's nominal per-token step is $2\Delta w$ bits. Every quantizer uses a straight-through estimator. Fitted codecs range from 27,972 to 158,532 parameters.

The three quantizers. *Residual binary quantization* (RBQ) applies the binary-spherical quantizer of Zhao et al. (2025) as the stage operator of the two residual quantizers in (7): it normalizes the latent, projects it through a learned bias-free hyperplane to w signs, and decodes through a learned map back to 64 dimensions. Its 2^w codes are implicit, and every code is reachable. Its objective adds to the reconstruction error an entropy regularizer on the batch-mean bit marginals, with weight 0.01. *Codebook VQ-VAE* gives each stage an explicit learned codebook of 2^w vectors in $\mathbb{R}^{64}$ and nearest-neighbour quantizes. Its objective adds a unit-weight codebook term and an encoder-side commitment term of weight 0.25. *Finite scalar quantization* projects the latent to a small product grid through a learned linear map, rounds coordinate-wise on a *fixed* grid, and projects back. The level factorization is fixed by width: 4×4 at $w = 4$, $8 \times 4 \times 4$ at $w = 7$ and $8 \times 8 \times 4 \times 4$ at $w = 10$, each giving exactly 2^w codes. Its objective is the reconstruction error alone. Table 3 collects the per-family objective. The binary implementation stores a commitment coefficient of 0.194 recorded in the protocol, but never applies it in the forward path or objective. Its fitted objective is reconstruction plus entropy. The VQ-VAE commitment coefficient 0.25 is active.

mechanism	code structure	term added to the reconstruction error
residual binary quantization	implicit 2^w, signed hyperplanes	entropy, weight 0.01
codebook VQ-VAE	explicit learned $2^w \times 64$	codebook + $0.25\times$ commitment
finite scalar quantization	fixed grid, level product $= 2^w$	none

Table 3: The three mechanisms of section 5 share the codec skeleton and update budget; the table lists their implemented objectives.

Budget, optimizer and snapshot. Each fit takes exactly 1,600 optimization steps at minibatch size 8,192, using Adam at learning rate 10^{-4} with the library's default moments, no weight decay, the gradient norm clipped to 1.0, and a cosine schedule that anneals the rate to zero over the 1,600 steps. The training split is 8 panels of 64 entities over 256 usable steps, which is 131,072 observation rows, so one epoch is 16 minibatches and the budget is exactly 100 epochs, with batches drawn by a per-epoch seeded permutation. Every fit in the population, for every mechanism, regime, width and initialization, shares this budget and optimizer. The controlled differences are mechanism and capacity. The codec is evaluated every 200 steps, at 8 checkpoints, on the held-out validation panels. It is scored by the mean squared error with which it reconstructs those panels' *observations*; the latent state is not used for selection. We retain the checkpoint with the lowest target-free validation reconstruction error; the validation split is used only for this choice.

Reader. The fitted codec's frozen latent is read by a ridge probe whose features are the last 4 steps of the summed 64-dimensional stage latents, concatenated into $h_t = [\, z_{t-3}; z_{t-2}; z_{t-1}; z_t \,] \in \mathbb{R}^{256}$, with $z_t = z_{1,t} + z_{2,t} \in \mathbb{R}^{64}$; the label is the known state at the next step. The primary score is the first state channel's per-step cross-sectional Spearman rank correlation (the rank-information coefficient); the four-channel average is a secondary readout. The penalty is chosen on the inner training panels only (6 of them) from a 9-point logarithmic grid $10^{-4}, \ldots, 10^{4}$, by the mean per-step cross-sectional rank information coefficient on those panels. The probe is then refit with that penalty on all training panels.

Population and splits. A cell is one (mechanism, regime, width) triple trained at 5 initializations seeded 42, 43, 44, 45, 46, and a transition is a pair of widths. Each regime trains on 8 panels and holds out 8 validation panels. All fits are scored on 32 test panels generated once, after the capacity choices are locked.

population	codecs	trans.	capacity grid, replication	specification
calibration	135	18	3 mech., 4/7/10 bits; 3 regimes, 5 inits	post-hoc
extended	315	54	RBQ/VQ 2–10, FSQ 4/7/10; 3 regimes, 5 inits	characterisation
prognostic	220	32	RBQ/VQ 2/4/7/10, FSQ 4/7/10; 4 regimes, 5 inits	preregistered
cross-generator	96	24	2 mech., 2/4/7/10 bits; 2 gen., 2 noise, 3 inits	predeclared locally
frozen re-read	316	56	as above, ridge/MLP/GRU readers; fresh panels	predeclared locally
operational	30	4	2 mech., 4/7/10 bits; ETTm1, 5 inits	post-hoc
external	49	9	VQ, K=32–256; 3 corruptions + clean, 3 inits	preregistered†
multi-site	48	72	RBQ/VQ, widths 2/4/7/10; 3 cells (clean + 2 noise), 2 inits, seasons as strata	predeclared locally
matched contrast	20	2	RBQ, widths 4/7; 2 conditions, 5 paired inits	predeclared locally

Table 4: Population engineering detail. Codec and transition counts, capacity grids and replication schemes for the principal populations, with the status terminology of Appendix H.1. The calibration and frozen-re-read rows re-read existing codecs rather than fit new ones. †That study's predeclared endpoint is inconclusive.

A.3 The prognostic endpoint

The prognostic population is the four-regime square of Table 2: RBQ and VQ-VAE at per-stage widths $\{2, 4, 7, 10\}$ and FSQ at $\{4, 7, 10\}$, 5 initializations at seeds 47–51, all disjoint from every earlier population in this project. Training and validation panels are generated under new seeds; the test panels are generated once under a further new seed and are read by no earlier experiment. No capacity is selected: every rung enters the endpoint. The codec skeleton, update budget, snapshot rule and reader recipe are those of section A.2; the endpoint is the one stated in section 5.4.

The main text reports the primary endpoint and its corrected interval. Four descriptive secondaries were declared alongside it: the state gain correlates 0.815 with the utility gain, the total reconstruction gain 0.799, the occupancy gain 0.562, and the state-information efficiency of the reference population -0.468. All 32 transitions lie below their reference, with excesses from -0.265 to -0.002 and no exception in either direction. The cross-generator study is reported separately in section E.2.

A.4 Population grids and replication

Table 4 records the codec counts, capacity grids and replication schemes behind the main-text evidence map. The claims those populations support are mapped in Table 1; their specification status is defined in Appendix H.1.

B Measurement and uncertainty

B.1 Estimator and its analytic controls

Estimator form. The estimator computes, per state, the pair means $Y_j = \frac{1}{2}(f(s+e_j) + f(s-e_j))$ for independent noise draws e_j, and their average $\bar{Y}$. Its population identity is

$$\mathbb{E}_\varepsilon\big[\|s - \bar{Y}\|^2 \mid s\big] = \|s - \mu(s)\|^2 + \frac{\operatorname{tr}\operatorname{Cov}(Y \mid s)}{m\,d}, \qquad \mu(s) = \mathbb{E}_\varepsilon[f(s+\varepsilon) \mid s]\,, \tag{10}$$

where m is the number of independent pairs (the replica count is $2m$) and d the channel count ($d = 4$ in these studies). Every squared norm here is a per-channel average, $\|v\|^2 = d^{-1}\sum_{k=1}^{d} v_k^2$, so the bias term is the channel-averaged covariance trace over m pairs: with tr read as the usual sum over channels, $\operatorname{tr}\operatorname{Cov}(Y)$ carries the channel sum and must be divided by $m\,d$. The sample variance of the pair means with $m-1$ degrees of freedom, averaged over channels, estimates that term, and subtracting it gives the bias-removed state account. The same term is added to the noise account, so the two still sum to the observation-reconstruction error D_x. The estimator is therefore

$$\widehat{b}(s) = \frac{1}{m(m-1)}\sum_{j=1}^{m}\|Y_j - \bar{Y}\|^2 = \frac{1}{m(m-1)d}\sum_{j=1}^{m}\sum_{k=1}^{d}(Y_{jk} - \bar{Y}_k)^2, \tag{11}$$

The right-hand expression expands the per-channel norm explicitly. We compute the correction per state, then average over sampled states, retaining negative corrected estimates. Independent, identically distributed draws and a fixed codec are assumptions; the estimator cannot verify them from its inputs.

The two accounts. Write $\mu_C(s) = \mathbb{E}[f_C(x) \mid s]$ for the codec's noise-averaged reconstruction. Because μ_C is the conditional mean, the single-sample state-reconstruction error splits exactly,

$$\mathbb{E}\|s - f_C(x)\|^2 = D_s(C) + \mathbb{E}\|f_C(x) - \mu_C(s)\|^2, \tag{12}$$

where the left side takes its error against the *state* rather than against the observation, so it is neither D_s nor D_x. D_s scores the averaged reconstruction, and the second term is the codec's own conditional spread around it. Under $\mathbb{E}[\varepsilon \mid s] = 0$ the same split reads on the residual $r_C = x - f_C(x)$ as

$$D_s(C) = \mathbb{E}\big\|\mathbb{E}[r_C \mid s]\big\|^2, \qquad D_\varepsilon(C) = \mathbb{E}\big\|r_C - \mathbb{E}[r_C \mid s]\big\|^2, \tag{13}$$

a conditional mean and a conditional variance.

For the identity codec $f_C(x) = x$, all three distortion accounts vanish, while its single-sample error against the state equals the observation-noise energy. Thus G_s measures a reduction in conditional-mean state distortion; it is neither usable state information nor a noise mutual information, bit rate, or count of encoded noise bits. Equation (13) requires conditionally zero-mean noise. The antithetic estimator additionally requires symmetric noise, independent pairs and a fixed codec.

Closed-form cases. The estimator is exercised against five closed-form cases: an identity codec (all components zero), a constant codec in arbitrary channel dimension, a non-diagonal affine codec whose conditional mean and zero bias are exact, a square map whose Monte-Carlo bias matches the population formula $2\sigma^4/m$, and a discrete sign map with exact enumeration whose corrected state account is negative, which demonstrates retention rather than clipping. The share helper is exercised separately: zero, negative, sub-threshold and non-finite denominators return an invalid marker rather than being silently conditioned away, and valid ratios are not clipped to $[0, 1]$.

B.2 The ledger's entries

The ledger has three layers. The first is the reconstruction gain and its two components, $G_x = G_s + G_\varepsilon$. The second is the share of that gain assigned to noise, $A_\varepsilon = G_\varepsilon / G_x$, and its reference-relative excess $E_\varepsilon = A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$. The third is an independently measured utility gain G_U. Table 5 defines all nine entries of Eq. (1), including the nominal capacity step and realized occupancy increment. The latter is the entropy increment of the paired two-stage token $z = z_1 V + z_2$ and determines the two rates used by the reference.

entry	what it is, and how it may be read
$\Delta R_{\mathrm{nom}},\ \Delta H$	nominal per-token step and realized joint-token occupancy increment, in bits.
$G_x,\ G_s,\ G_\varepsilon$	reconstruction gain, split into a fall in conditional-mean state distortion (G_s) and a fall in the noise-related residual component (G_ε); $G_x = G_s + G_\varepsilon$ by construction, so a zero residual is a consistency check, not evidence.
A_ε	marginal noise share G_ε / G_x, a ratio of two reconstruction gains and not a share of ΔR_{nom} or ΔH; defined where $G_x > 0$, above one exactly when $G_s < 0$ and below zero exactly when $G_\varepsilon < 0$.
$A_\varepsilon^{\mathrm{ref}}$	the same-occupancy second-moment-matched reference put through the same gain arithmetic as the measurement, Eq. (3); a Gaussian proxy rather than an optimal codec.
E_ε	excess $A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$: both are noise gain shares over the same transition, so the comparison is direct; it is not a distance to an optimum.
G_U	future-utility gain, measured on its own readout and derived from none of the entries above.

Table 5: The ledger's entries and how each may be read. A_ε and $A_\varepsilon^{\mathrm{ref}}$ are the same kind of quantity — noise gain shares over the same transition — computed from the measurement and from the reference respectively, so every reference comparison in this paper is direct.

B.3 The accounts, and what each one measures

Table 6 distinguishes the accounts by target, readout, metric and time index. State and noise distortions measure same-time reconstruction under the codec; utility measures one-step-ahead

prediction under a specified readout. The two oracle arms in Table 9 agree to within 10^{-4} in every regime, consistent with the state carrying the future information and the observation adding none given it. The observation-only linear readout is a baseline, so the gap to the oracle arms is not a recovery rate.

account	target	read-out	loss / metric	index	in
conditional-mean state distortion D_s	s_t	codec f_C	squared error, ch-avg.	t	(2)
noise-related residual component D_ε	residual around its conditional mean	codec f_C	conditional variance, ch-avg.	t	(2)
utility gain G_U	s_{t+1}	ridge, 4-step latent history	per-step cross-sectional rank IC (first channel)	$t+1$	section 3
linear-readout baseline	s_{t+1}	linear on obs. or state	per-step cross-sectional rank IC (first channel)	$t+1$	section D.2
state-information capture	s_t	ridge on the latent	state-error drop / base	t	section F.4
closure probes	s_t, $\mathbb{E}[s_{t+1} \mid H_t]$	ridge on the latent	R^2	t, $t{+}1$	section F.2

Table 6: The accounts the main text reports, named by what each one predicts, what reads it, how it is scored and at which time index. They are not interchangeable: the state account scores the codec's noise-averaged reconstruction, the noise account scores the residual around that conditional mean — the utility account is a one-step-ahead rank score under a fixed linear reader, the selection rule's capture is a normalized drop in same-time ridge state error against an indirect Gaussian reference rather than a score, and the closure study's two probes are scored by R^2 on their own populations. The interface study's arms are a further pair of read-outs, described in section F.1.

B.4 The matched reference: construction and scope

The reference evaluates the same transition-level gain share as the measured ledger, using a Gaussian test channel at the codec's realized occupancies. Its noise account includes test-channel randomness. It provides an allocation baseline, not an optimal allocation of physical observation noise; the noiseless control in section C makes this distinction explicit.

Second moments and water-filling. Let the standardized observation and signal satisfy $x = s + e$ elementwise and be flattened to N rows of dimension d. The reference forms the empirical second-moment matrices

$$\hat{\Sigma}_x = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^\top, \qquad \hat{\Sigma}_s = \frac{1}{N}\sum_{n=1}^{N} s_n s_n^\top, \tag{14}$$

symmetrizes each as $(\hat{\Sigma} + \hat{\Sigma}^\top)/2$, and defines the noise second moment by exact subtraction, $\hat{\Sigma}_e = \hat{\Sigma}_x - \hat{\Sigma}_s$. These are uncentered second moments of the training-standardized arrays: the construction does not subtract a sample mean, so it matches the second-moment matrix rather than the centered sample covariance, and the reference carries that name. Taking the signal second moment as measured and the noise as the residual makes the identity below hold by construction and folds the finite-sample cross term into the noise account. The construction requires $\lambda_{\min}(\hat{\Sigma}_e) \geq 0$ and raises otherwise.

Diagonalize $\hat{\Sigma}_x = U\,\mathrm{diag}(\lambda)U^\top$, floor the eigenvalues at 10^{-12}, and for a level θ define

$$R(\theta) = \sum_{i:\,\lambda_i > \theta} \tfrac{1}{2}\log_2 \frac{\lambda_i}{\theta}, \tag{15}$$

zero when no direction is retained. The level saturating the requested rate is found by bisecting *geometrically* — the trial point is $\sqrt{\theta_{\mathrm{lo}}\theta_{\mathrm{hi}}}$, starting from $\lambda_{\max}10^{-14}$ — and stops when $\theta_{\mathrm{hi}}/\theta_{\mathrm{lo}} - 1 < 10^{-14}$, so the operative tolerance is a relative one on θ.

Shrinkage and the three distortions. At level θ the test channel's complement is diagonal *in the eigenbasis*: $\widetilde{W} = \mathrm{diag}(w)$ with $w_i = \max(0, 1 - \theta/\lambda_i)$, whose original-coordinate form is $W = U\widetilde{W}U^\top$. The traces below are therefore taken in the eigenbasis, against the rotated second-moment matrices $\tilde{\Sigma}_s = U^\top\hat{\Sigma}_s U$ and $\tilde{\Sigma}_e = U^\top\hat{\Sigma}_e U$. All factors in each trace use this eigenbasis.

With the coding noise on retained directions $\nu_i = (\lambda_i - \theta)\theta/\lambda_i$ and zero on the rest, the reference's per-channel-mean distortions at the same $1/d$ normalization the measured accounts use are

$$D_x^{\mathrm{ref}} = \frac{1}{d}\sum_i \min(\lambda_i, \theta), \qquad D_s^{\mathrm{ref}} = \frac{1}{d}\,\mathrm{tr}\big[(I - \widetilde{W})\tilde{\Sigma}_s(I - \widetilde{W})\big], \tag{16}$$

$$D_\varepsilon^{\mathrm{ref}} = \frac{1}{d}\Big[\mathrm{tr}\big[(I - \widetilde{W})\tilde{\Sigma}_e(I - \widetilde{W})\big] + \sum_i \nu_i\Big], \tag{17}$$

so $D_x^{\mathrm{ref}} = D_s^{\mathrm{ref}} + D_\varepsilon^{\mathrm{ref}}$ identically, to 10^{-9} on the frozen runs. The reference quantities carry the ref superscript throughout to keep them apart from the measured accounts of the same name. A two-dimensional case with a non-diagonal $\hat{\Sigma}_x$ and a non-isotropic $\hat{\Sigma}_s$ has closed-form answers, and the frozen implementation reproduces them to machine precision and is invariant to rotating the problem; section I records the check.

The rate handed to the reference is the model's own realized joint-token entropy at that configuration, not a nominal codebook size, and the reference is evaluated panel by panel over the frozen test panels. For a transition the reference gains and their share are

$$G_s^{\mathrm{ref}} = D_s^{\mathrm{ref}}(H(C_1)) - D_s^{\mathrm{ref}}(H(C_2)), \qquad G_\varepsilon^{\mathrm{ref}} = D_\varepsilon^{\mathrm{ref}}(H(C_1)) - D_\varepsilon^{\mathrm{ref}}(H(C_2)), \tag{18}$$

$$G_x^{\mathrm{ref}} = G_s^{\mathrm{ref}} + G_\varepsilon^{\mathrm{ref}}, \qquad A_\varepsilon^{\mathrm{ref}} = G_\varepsilon^{\mathrm{ref}}/G_x^{\mathrm{ref}} \;\; (G_x^{\mathrm{ref}} > 0), \tag{19}$$

which is Eq. (3); the measured share is formed the same way from the calibrated accounts and is undefined when G_x is not positive.

Scalar proxy for Saugeen. This construction is estimable only when the observation second-moment matrix rests on enough effective samples. On the Saugeen series of section G.2 it is not: the window second-moment matrix has on the order of 12 effective samples and the measured residual has a minimum eigenvalue of -1.50 against a mean eigenvalue of 1.0. The reference is therefore replaced there by a per-time-point scalar variance-matched construction at the model's realized occupancy per time point, and that replacement is declared a proxy. The measured accounts are unchanged by the substitution; only the reference share is. This substitution is specific to the measured second moments of that series.

B.5 Aggregation and uncertainty

Aggregation order and bias removal. Every per-configuration distortion is the mean over that configuration's initializations of the seed's own distortion, and each seed's distortion is already a channel average and an average over its sampled states. A transition gain is the difference of the two configuration means, $\overline{D}(C_1) - \overline{D}(C_2)$; because both configurations share the same initialization set this equals the mean over initializations of the per-seed paired difference. An allocation share is formed *once* from those aggregated gains — a ratio of means, $A_\varepsilon = \overline{G}_\varepsilon/\overline{G}_x$ and $A_\varepsilon^{\mathrm{ref}} = \overline{G}_\varepsilon^{\mathrm{ref}}/\overline{G}_x^{\mathrm{ref}}$, so $E_\varepsilon = A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$ — and is never an average of per-seed shares. The reference is put through the same arithmetic, so both sides of Eq. (3) are the same estimand.

The state account removes the finite-pair Monte-Carlo bias in expectation under the estimator's stated assumptions — independent, identically distributed symmetric pairs and a fixed codec, as stated with the estimator in section B.1; it is not an exact removal in every finite realization. Negative corrected values are retained rather than clipped.

Reported intervals. The prognostic endpoint's interval $[-0.569, -0.468]$ resamples the 5 initializations with replacement, one draw shared by all transitions inside a replicate, and recomputes the transition means, differences, ratios and excesses before taking the Spearman correlation, so the resampled functional matches the point estimate; $10{,}000$ replicates and $2.5/97.5$ percentiles are used, and a replicate is dropped only if it has fewer than three usable transitions or a constant vector. The superseded [-0.584, -0.474] averaged per-seed ratios instead and is quoted only where flagged. The replica-convergence criterion applies the operational tolerance $\max(10^{-5}, 0.10 \times |G_s|)$ to the bootstrap interval of the difference between a coarse rung's corrected state gain and the finest rung's, as an equivalence band. The 17 of 18 calibration transitions counted as resolved use a family-wise adjusted per-transition interval ($\alpha = 0.05/18$ over the 18 transitions): a transition counts only if its corrected state gain is positive and the lower end of that interval is above zero.

Exact seed-support enumeration. The $n = 5$ seed bootstrap behind the prognostic interval was enumerated exactly. Every draw of five labels from five seeds is a multiset, so the replicate distribution is supported on $\binom{9}{5} = 126$ atoms with multinomial probabilities $5!/(\prod_i n_i!) \,/\, 5^5$. Each atom was evaluated once through the same functional as the corrected analysis (seed means, gains, ratio of means, reference share, excess, Spearman over 32 transitions). The identity multiset reproduces the prognostic point estimate exactly. Replaying the original RNG stream maps all 10000 stored Monte-Carlo draws to their atoms with zero missing atoms. All 126 atoms are valid and take 103 distinct correlation values.

enumeration quantity	value
identity-multiset correlation	-0.5381231671554252
maximum deviation from stored draws	0.000
exact step interval	$[-0.5692815249266863, -0.4677419354838709]$
smoothed linear interval	$[-0.5692815249266863, -0.4695381231671553]$
support	$[-0.5879765395894428, -0.4266862170087977]$
identity-atom weight	0.0384
frequency total-variation distance	0.0365
weighted probability of $\rho < 0$	1.0000

The exact step interval equals the corrected Monte-Carlo interval. Reading the 126 atoms as equally likely would widen the interval to $[-0.5762463343108505, -0.4475806451612903]$, 27% wider than the exact interval. Multinomial weights are required for the bootstrap distribution. This is a post-hoc supplementary analysis of a preregistered endpoint: it removes Monte-Carlo resampling error from the reported percentiles, but it does not enlarge the population and does not change the point estimate, decision table or verdict.

The corrected matched-pair evaluation's intervals come from the shared hierarchical resampler over initialization, panel and sampled state. Before an interval is emitted the runner forms the mean of the replicate values and *asserts* that it carries the point estimate's sign; a disagreement raises and aborts the run. The check therefore neither filters nor down-weights draws: the percentile interval is computed from the full replicate array. The prognostic runner records the same diagnostic without asserting on it.

C ESTIMATOR VALIDATION

Figure 6 collects the replica-convergence and dose-response summaries discussed in this subsection.

Known observation-noise ordering. The measured noise account recovers the ordering of the known noise fractions. The per-case means are 0.961–0.998 and 0.943–0.988 in the two high-observation-noise regimes, against 0.766–0.928 in the low-observation-noise regime. The predeclared criterion passes 9 of 9 cases, a one-sided rank test of the pooled high-noise regimes against the low-noise one is significant in 9 of 9 cases with and without a Bonferroni divisor of nine, and no state account exceeds the signal energy (0 violations in 135 models).

Training-noise dose. The dose study re-trains one mechanism with an encoder-input corruption $c\,\eta$ added to the observation, where $c = \sqrt{k-1}$ and η is drawn once and scaled so that its mean square equals the observation noise's; $k = 1$ and $k = 2$ reuse frozen checkpoints from an earlier arm and only $k = 4$ is newly trained, after a check that the re-implemented recipe reproduces the frozen $k = 1$ arm's recorded validation error. The dose study predeclared no direction and makes no utility claim; the degraded reconstruction at high dose is expected and is reported as such. The per-record noise shares behind the three curve means are recorded per capacity and initialization — the state share of the $7\to10$ increment runs 0.088, 0.215 and 0.502 across the three doses — and the rank-scoring convention used for the utility term is cross-checked against an independent implementation with agreement to machine precision.

Gaussian and distortion-matched controls. The Gaussian control source is constructed to be Gaussian with the same covariance structure as the benchmark, and the check is a normality test on the whitened observations: the maximum absolute marginal skewness is 0.0053 and the maximum

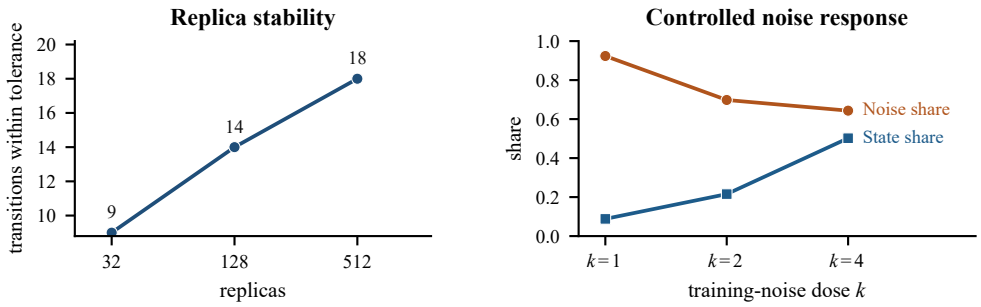


Figure 6: Finite-replica convergence and response to training-time noise. These descriptive checks do not replace the failed predeclared analytic criterion discussed below; the dose study predeclared no direction and makes no utility claim.

absolute excess kurtosis is 0.0112, both far inside the band that would signal a departure. The ordering reported in the main text is reproduced on this source — every one of its 12 transitions falls below the reference, with excesses from -0.103 to -0.007. The ordering therefore also occurs without a regime mixture. Moving each reference point to the model's own measured distortion instead of its own rate preserves the ordering over the 135 frozen models, whose reference shares then span 0.887–0.998. Like the main construction, both controls compare a deterministic codec with a stochastic reference. The closed-form quantizer below isolates that distinction.

Replica detail. The ladder is read in two layers. The 32- and 128-replica rungs are assessed against the 512-replica reference, and the 512-replica rung is then assessed against the 2048-replica endpoint. Against 512, 9 and 14 of the 18 transitions fall within the operational tolerance; against 2048, all 18 do. With the endpoint count, 17 of the 18 transitions have a positive corrected state gain with an adjusted interval excluding zero, and 1 does not. The bias the estimator removes at the coarse rung spans 3.21×10^{-4}–1.76×10^{-3}, against 4.81×10^{-6}–2.75×10^{-5} at the endpoint. It is small in absolute terms, but of the same order as the gains themselves, which is why the endpoint count is the one reported. The across-stream spread that underpins the endpoint prediction is a descriptive spread over the three declared streams, not a variance decomposition and not an error bound.

The closed-form construction is a scalar Gaussian source with unit variance, a deterministic uniform quantizer and a decoder equal to the quantizer, so the conditional mean and both accounts are available through the standard normal CDF. Rate is the entropy of the quantized output, computed from the corrected bin probabilities, and the second-moment-matched reference is evaluated at that same realized rate. The grid is three noise settings, four step sizes and two codec offsets. The offset control was not in the protocol and is recorded as a deviation, because a zero-offset uniform quantizer has a nearly unbiased conditional mean that would have left the state account untested. The reported rates and reference quantities use the corrected output-bin probabilities; Table 12 records the superseded values. The histogram defect is described in section H.2.

The noiseless criterion (K1) passes: at zero observation noise the corrected noise account is below 10^{-30} in the noiseless cells, while the reference's *level* share q^{ref} assigns 0.9908–0.9999 of that account to noise. q^{ref} is its noise account over its own error account at one configuration, which is the analytic study's own quantity and not the ledger's Eq. (3). Where observation noise dominates the two agree to a median ratio of 1.000.

The second predeclared criterion, the finite-replica accuracy criterion (K2), failed. It required two conditions: the corrected error against the closed form had to fall as the replica count rose, and at the finest rung it had to lie within one Monte Carlo standard error. Judged component by component, 5 of the 48 comparisons — 24 configurations times two components, not 48 independent configurations — exceed their own standard error, and 3 exceed it when both components are judged against the larger of the two. Among the violations, the error runs from 1.83×10^{-9} to 2.58×10^{-4} and is 1.01–1.54

of the component's own standard error. The consequence the protocol declared was applied: the calibration claim of that study is withdrawn.

The study's *gain* share over adjacent step sizes has the same qualitative behavior as its level share. In the noisy settings the deterministic codec's noise gain share is 1.000 against a reference gain share of 0.999, and in the noiseless cell it is 0.000 against 0.988. The two agree where observation noise dominates and separate exactly where it vanishes, which is the same boundary K1 draws.

D ALLOCATION RESULTS AND MECHANISM CONTROLS

D.1 ALLOCATION ACROSS CAPACITIES AND REGIMES

Table 7 reports the extended population's 54 transitions, each aggregated over the 5 initializations of its cell. Columns are the nominal per-token step, the realized occupancy increment, the reconstruction gain, its state and noise components, the marginal noise share, and the same-occupancy reference gain share. A transition whose reconstruction gain is not positive has no defined share and is marked as such rather than dropped. The excess over the reference is the difference of the last two columns.

Table 8 groups the same shares by observation-noise level, showing both the stronger noise dominance in high-noise cells and the exceptions in low-noise cells. Calibration and extended populations appear in separate blocks.

D.2 LINEAR-READOUT BASELINES

Table 9 compares linear readouts of observations, states, and their combination on the same four-lag history used for utility. The two state-informed arms give nearly identical scores. No tokenized arm in this population exceeds the observation-only baseline, although a nonlinear latent representation could allow a linear readout to do so. These unquantized arms are linear-readout baselines, not Bayes ceilings.

D.3 MECHANISM CONTROLS

Five descriptive control arms examine whether codebook updates, parameter count or loss weights account for the family differences.

Before those controls, the display below prints the contrast itself from the same frozen rows as Table 7. Each cell sums the 7→10-bit span's seed-mean increments: three one-bit steps for RBQ and VQ-VAE, one three-bit step for FSQ. ΔH is the realized joint-token occupancy increment in bits; G_U is the same rank-information-coefficient gain as in Table 7. VQ-VAE adds the least occupancy in every regime, and its utility increment is the smallest and negative in the two high-predictability regimes.

regime	family	ΔH (bits)	G_U
high pred., high obs. noise	RBQ	1.73	$+0.0154$
	VQ-VAE	0.44	-0.0014
	FSQ	1.99	$+0.0032$
high pred., low obs. noise	RBQ	1.76	$+0.0379$
	VQ-VAE	0.47	-0.0012
	FSQ	1.90	$+0.0089$
low pred., high obs. noise	RBQ	1.73	$+0.0061$
	VQ-VAE	0.48	$+0.0005$
	FSQ	2.03	$+0.0004$

Codebook schedule and parameter count. Replacing the codebook update with an exponential moving average (decay 0.99) and restarting a code that has gone unused for ten steps raises first-stage support by a factor of 9.4 and cuts reconstruction error to 0.329 of the published value, adding 5.04 bits of realized occupancy. The 7→10 utility increment does not move despite the gains in code use and reconstruction: the plateau survives this codebook-schedule change. One schedule intervention does not rule out an interaction with the optimizer, the budget or the training path, and the claim is

regime	family	$C_1 \to C_2$	ΔR_{nom}	ΔH	G_x	G_s	G_ε	A_ε	$A_\varepsilon^{\mathrm{ref}}$
highpred_highSNR	bsq	2→3	2	1.27	0.1569	0.0796	0.0773	0.493	0.855
highpred_highSNR	bsq	3→4	2	1.18	0.0926	0.0326	0.0601	0.649	0.902
highpred_highSNR	bsq	4→5	2	1.02	0.0750	0.0320	0.0430	0.573	0.935
highpred_highSNR	bsq	5→6	2	0.60	0.0349	0.0097	0.0253	0.724	0.951
highpred_highSNR	bsq	6→7	2	0.86	0.0408	0.0122	0.0286	0.701	0.963
highpred_highSNR	bsq	7→8	2	0.42	0.0076	0.0011	0.0065	0.855	0.969
highpred_highSNR	bsq	8→9	2	0.57	0.0253	0.0073	0.0180	0.713	0.974
highpred_highSNR	bsq	9→10	2	0.77	0.0209	0.0026	0.0183	0.877	0.979
highpred_highSNR	fsq	4→7	6	3.80	0.1054	0.0269	0.0785	0.745	0.975
highpred_highSNR	fsq	7→10	6	1.90	0.0184	0.0027	0.0157	0.852	0.992
highpred_highSNR	vqvae	2→3	2	1.54	0.1395	0.0410	0.0986	0.706	0.884
highpred_highSNR	vqvae	3→4	2	1.04	0.0609	0.0117	0.0492	0.807	0.927
highpred_highSNR	vqvae	4→5	2	0.56	0.0310	0.0083	0.0227	0.731	0.946
highpred_highSNR	vqvae	5→6	2	0.44	0.0206	0.0053	0.0153	0.741	0.954
highpred_highSNR	vqvae	6→7	2	0.17	0.0059	0.0009	0.0051	0.853	0.959
highpred_highSNR	vqvae	7→8	2	0.28	0.0111	0.0018	0.0093	0.837	0.962
highpred_highSNR	vqvae	8→9	2	0.12	0.0042	0.0015	0.0027	0.650	0.964
highpred_highSNR	vqvae	9→10	2	0.06	0.0017	-0.0001	0.0018	1.079	0.967
highpred_lowSNR	bsq	2→3	2	1.29	0.1291	0.0171	0.1120	0.867	0.961
highpred_lowSNR	bsq	3→4	2	1.15	0.1020	0.0073	0.0948	0.929	0.974
highpred_lowSNR	bsq	4→5	2	1.00	0.0683	0.0079	0.0604	0.884	0.983
highpred_lowSNR	bsq	5→6	2	0.61	0.0276	0.0023	0.0253	0.916	0.987
highpred_lowSNR	bsq	6→7	2	0.94	0.0401	0.0025	0.0377	0.938	0.990
highpred_lowSNR	bsq	7→8	2	0.48	0.0094	0.0002	0.0091	0.974	0.992
highpred_lowSNR	bsq	8→9	2	0.53	0.0221	0.0013	0.0208	0.942	0.993
highpred_lowSNR	bsq	9→10	2	0.72	0.0200	0.0004	0.0196	0.980	0.995
highpred_lowSNR	fsq	4→7	6	3.83	0.0884	0.0040	0.0844	0.955	0.993
highpred_lowSNR	fsq	7→10	6	1.99	0.0166	0.0003	0.0162	0.980	0.998
highpred_lowSNR	vqvae	2→3	2	1.58	0.1387	0.0104	0.1283	0.925	0.968
highpred_lowSNR	vqvae	3→4	2	0.96	0.0584	0.0030	0.0554	0.949	0.980
highpred_lowSNR	vqvae	4→5	2	0.66	0.0313	0.0018	0.0295	0.942	0.985
highpred_lowSNR	vqvae	5→6	2	0.40	0.0168	0.0005	0.0162	0.968	0.988
highpred_lowSNR	vqvae	6→7	2	0.22	0.0089	0.0002	0.0087	0.979	0.989
highpred_lowSNR	vqvae	7→8	2	0.24	0.0086	0.0003	0.0082	0.963	0.990
highpred_lowSNR	vqvae	8→9	2	0.16	0.0051	0.0003	0.0048	0.942	0.991
highpred_lowSNR	vqvae	9→10	2	0.04	0.0002	-0.0002	0.0005	1.969	0.990
lowpred_lowSNR	bsq	2→3	2	1.28	0.1243	0.0167	0.1075	0.865	0.956
lowpred_lowSNR	bsq	3→4	2	1.15	0.1023	0.0066	0.0958	0.936	0.970
lowpred_lowSNR	bsq	4→5	2	0.99	0.0667	0.0047	0.0620	0.930	0.980
lowpred_lowSNR	bsq	5→6	2	0.63	0.0291	0.0021	0.0270	0.928	0.985
lowpred_lowSNR	bsq	6→7	2	0.93	0.0378	0.0019	0.0359	0.949	0.989
lowpred_lowSNR	bsq	7→8	2	0.49	0.0095	0.0004	0.0092	0.962	0.991
lowpred_lowSNR	bsq	8→9	2	0.54	0.0219	0.0007	0.0213	0.970	0.992
lowpred_lowSNR	bsq	9→10	2	0.70	0.0197	0.0007	0.0190	0.963	0.994
lowpred_lowSNR	fsq	4→7	6	3.80	0.0858	0.0015	0.0843	0.982	0.992
lowpred_lowSNR	fsq	7→10	6	2.03	0.0166	0.0001	0.0165	0.994	0.998
lowpred_lowSNR	vqvae	2→3	2	1.56	0.1348	0.0079	0.1269	0.941	0.964
lowpred_lowSNR	vqvae	3→4	2	0.95	0.0566	0.0026	0.0540	0.953	0.977
lowpred_lowSNR	vqvae	4→5	2	0.71	0.0335	0.0010	0.0325	0.970	0.983
lowpred_lowSNR	vqvae	5→6	2	0.33	0.0133	0.0004	0.0129	0.971	0.986
lowpred_lowSNR	vqvae	6→7	2	0.25	0.0100	0.0003	0.0097	0.970	0.987
lowpred_lowSNR	vqvae	7→8	2	0.25	0.0087	0.0002	0.0085	0.979	0.988
lowpred_lowSNR	vqvae	8→9	2	0.17	0.0062	0.0002	0.0060	0.967	0.989
lowpred_lowSNR	vqvae	9→10	2	0.06	0.0007	-0.0000	0.0007	1.015	0.989

Table 7: The reconstruction-side ledger entries for the 54 transitions of the extended population, aggregated over 5 initializations per cell. ΔR_{nom} is the nominal per-token step in bits and ΔH the realized joint-token occupancy increment, also in bits. A share is left blank where the reconstruction gain is not positive, because the ratio is then undefined; such transitions are retained in the population and are reported rather than dropped.

limited to the schedule it tested. At the largest width the repaired first stage saturates at 1024 active codes against 46 for the published baseline.

Widening the binary mechanism to match the codebook model's parameter count (166340 parameters at the largest width) moves utility by 6.7% and changes reconstruction by a factor of 0.985, without reordering the accounts. Parameter count is a real but modest contributor; the preserved ordering is a qualitative result.

Over the extended population the pooled rank correlations with the utility gain are 0.601 for the state gain, 0.497 for total reconstruction gain, 0.440 for occupancy gain and -0.470 for the reference-relative excess. They are descriptive: the ordering does not reproduce inside every mechanism-by-regime

population	observation-noise level	n	min	median	max
calibration	high (low pred.)	6	0.935	0.973	0.994
calibration	high (high pred.)	6	0.911	0.963	0.986
calibration	low	6	0.690	0.821	0.918
extended	high (low pred.)	18	0.865	0.965	1.015
extended	high (high pred.)	18	0.867	0.946	1.969
extended	low	18	0.493	0.736	1.079

Table 8: The marginal noise share A_ε grouped by population and observation-noise level. The two populations are summarised separately and never pooled: the calibration population has 18 transitions and the extended ladder 54. Each cell is the median of that group's transition shares, and a transition's share is itself already a ratio of seed-mean gains (section B.5), so no per-seed share is averaged here. The extended ladder's three mechanisms do not contribute equally many transitions (24, 24 and 6), so its overall median is not a mechanism-equal-weight statistic. Transitions below one half and above one are retained in the min/max columns rather than trimmed.

regime	raw observation only	latent state only	observation and state
low pred., high obs. noise	0.1159	0.4361	0.4360
high pred., high obs. noise	0.2367	0.7925	0.7925
high pred., low obs. noise	0.4967	0.7925	0.7925

Table 9: Per-step cross-sectional rank information coefficient (first channel) of three linear readers on the same four-lag history, one regime per row. The two oracle arms agree to within 10^{-4} in every regime, so once the state is available the observation adds nothing about the future. The observation-only arm is a *linear-readout baseline*, not a Bayes ceiling, and the gap between it and the oracle arms is therefore not a recovery rate.

stratum, where 5, 5 and 5 of 9 strata are positive for the three gains and 2 is for the excess. The pooled ordering is therefore insufficient to rank the accounts across strata.

Hyperparameters. Three arms move one hyperparameter each. Driving the entropy weight to zero reduces realized occupancy by -0.93 bits and worsens reconstruction; raising it leaves occupancy and reconstruction essentially unchanged; lowering the codebook commitment leaves occupancy essentially unchanged. The contrast is therefore not an artifact of one hyperparameter setting. We make no directional claim from these arms.

E Robustness and controlled training contrasts

E.1 Capacity-conditioned association

We re-read two frozen populations under a predeclared control scheme. Population A is the prognostic population of section 5.4: 32 transitions, 5 initializations, 4 regimes $\times$ 3 quantizer families. Population B is the cross-generator population of section E.2: 24 transitions, 3 initializations, 2 generators $\times$ 2 noise levels $\times$ 2 mechanisms. The analysis uses existing fits. The two reconstructed ledgers agree with the published records to numerical precision:

population	transition-field comparisons	maximum absolute deviation
A	192	2.6×10^{-15}
B	216	1.9×10^{-15}

The reconstruction reads the underlying records independently of the published aggregates; A's uncontrolled correlation also agrees to this precision.

The control scheme is S0 uncontrolled; S1 within-stratum mean; S2 control for lower nominal bits; S3 adds nominal step; S4 (primary) adds realized occupancy gain; S5 adds family indicators; S6 is S0 on the rung-disjoint subset. The seed axis is enumerated exactly: 126 atoms on A and 10 on B.

Population A.

spec	value (n)	exact seed interval
Uncontrolled (S0)	-0.5381 (32)	$[-0.5693, -0.4677]$
Within-stratum mean (S1)	-0.6667 (32)	$[-0.7440, -0.5476]$
Lower nominal bits (S2)	-0.5121 (32)	$[-0.5526, -0.3647]$
Nominal step (S3)	-0.5459 (32)	$[-0.5900, -0.3698]$
Realized occupancy gain (S4)	-0.5458 (32)	$[-0.5856, -0.3696]$
Family indicators (S5)	-0.5466 (32)	$[-0.5623, -0.3440]$
Rung-disjoint subset (S6)	-0.5008 (20)	$[-0.5895, -0.2992]$

Population B.

spec	value (n)	exact seed interval
Uncontrolled (S0)	-0.2870 (24)	$[-0.3887, -0.1348]$
Within-stratum mean (S1)	-0.7429 (24)	$[-0.7429, -0.4429]$
Lower nominal bits (S2)	-0.0450 (24)	$[-0.2147, +0.0361]$
Nominal step (S3)	-0.0554 (24)	$[-0.2290, +0.0440]$
Realized occupancy gain (S4)	$+0.0259$ (24)	$[-0.2004, +0.2199]$
Family indicators (S5)	$+0.4171$ (24)	$[-0.0641, +0.4171]$
Rung-disjoint subset (S6)	-0.2471 (16)	$[-0.4676, -0.0412]$

On A, all 126 exact seed atoms of S4 are negative (range -0.6383 to -0.2746); on B, 63% of the exact weight is positive and the support covers zero (range -0.2004 to $+0.2199$). The pooled-vs-stratified difference is descriptive, not a variance decomposition: A's between-stratum correlation is -0.600 and its within-stratum mean is -0.667 (same sign), whereas B's between-stratum correlation is $+0.400$ and its within-stratum mean is -0.743 (opposite signs), consistent with the weaker pooled correlation -0.2870. At matched capacity, the mechanism with the more negative excess has the higher utility gain in 19 of 28 pairs on A ($\rho = -0.352$, sign-test $p = 0.087$) and 10 of 12 on B ($\rho = -0.336$, $p = 0.039$).

The design limits how finely these effects can be separated. A's design has rank 4 of 4 with residual df 28; B's S4 has residual df 20. Nominal step is a function of the rung pair (VIF ≈ 2.8 on A, 4.0 on B), so S3 adds little over S2. Realized occupancy gain is family-dependent, so S4 and S5 remove part of the mechanism contrast by construction. The seed enumeration resamples initializations only; the panel axis is not recomputed for the partial statistics. The auxiliary incremental-value regression is unaffected and still reports no improvement from adding the state account; the estimands differ.

E.2 The cross-generator study

We test the allocation findings under two additional latent dynamics: a nonlinear saturating recurrence and a seasonal process with stochastic volatility. Both use additive Gaussian observation noise. This is a separate population from the prognostic study of section 5.4; its results do not enlarge that population or its interval. The design was predeclared locally (Table 11).

Design. The nonlinear recurrence combines a cyclic permutation with a saturating nonlinearity, driven by standardized Student-t innovations with a finite fourth moment. The second is a seasonal recurrence with a lag of one day-cycle and a stochastic-volatility state. We denote the two generators g_1 (nonlinear) and g_2 (seasonal). In full, with $P(s_1, s_2, s_3, s_4) = (s_2, s_3, s_4, s_1)$ and t_7 a unit-variance Student-t innovation with seven degrees of freedom,

$$s_t = 0.20\, s_{t-1} + 0.80\, \tanh(1.50\, P s_{t-1}) + 0.35\, \xi_t, \qquad \xi_t = \sqrt{5/7}\, t_7, \tag{20}$$

$$h_t = 0.95\, h_{t-1} + 0.15\, u_t, \qquad u_t \sim \mathcal{N}(0,1), \tag{21}$$

$$s_t = 0.25\, s_{t-1} + 0.55\, s_{t-24} + 0.20\, \sin(2\pi t/24 + \phi) + 0.25\, e^{h_t/2}\, \eta_t, \qquad \eta_t \sim \mathcal{N}(0,1), \tag{22}$$

with $\phi \sim \mathrm{Unif}(0, 2\pi)$ per panel, entity and channel. Both processes run for a burn-in of 2048 steps before recording panels; the split grid is 8/8/32 panels with 64 entities and 256 time steps. The latent volatility and phase are never given to the estimator. They share one observation model, $x_t = s_t + \sigma_\varepsilon \varepsilon_t$, whose noise scale and per-channel normalizer are computed on the training split

alone and then frozen. The non-Gaussianity and the heteroskedasticity therefore sit in the latent state, not in the observation noise the estimator assumes. The study trained 96 codecs at one budget and target SNRs 0.1 and 1.0 (2 generators $\times$ 2 observation-noise levels $\times$ 2 quantizer families $\times$ 3 initializations $\times$ 4 widths), and measured every transition with 2048 sampled states and 1024 antithetic pairs per state. Its endpoint is the *stratified* form defined in the protocol: the within-noise-level correlation taken separately, and the two averaged with equal weight. That endpoint is denoted Q_g, with $H_1 : Q_g < 0$; it differs from the pooled single-ρ estimand of section 5.4.

Results and interpretation. The stratified allocation–utility association is negative on both generators: $Q_g = -0.657$ with interval [-0.829, -0.514] for the nonlinear process, and $Q_g = -0.829$ with interval [-0.829, -0.029] for the seasonal process. The corresponding H_1 verdicts are supported and supported. Median marginal noise shares are 0.899 and 0.900, and all 24 transitions lie below the reference: 12 of 12 and 12 of 12, respectively.

Each generator has 6 of 6 transitions valid at each noise level. Fractions of bootstrap replicates retained are 1.000 and 1.000; no draw is invalid. The pooled association, however, does not survive capacity conditioning (section E.1).

The two generators are fixed design choices. Their intervals describe evaluation uncertainty within these processes, not variation over a population of dynamical models. Within each observation-noise level, the 6 transitions are three consecutive rungs on each of two quantizer ladders. A shared capacity trend, with excess decreasing and utility gain increasing, can itself produce a negative correlation. Part of Q_g therefore reflects ladder position rather than a cross-sectional relation. Pooling the two observation-noise levels, which the predeclared statistic deliberately does not do, collapses the correlation to -0.364 and -0.238, the visible size of the degeneracy. This SNR pooling is distinct from the capacity-conditioned analysis in Section E.1, which conditions on ladder position and realized occupancy rather than averaging over observation-noise levels.

The reference changes interpretation but adds little ranking separation on these populations: the excess and measured share are close to rank-identical (0.958 and 0.979), and the reference share is not constant, so the ledger stays readable. No incremental ranking information from the reference is demonstrated on these two mechanisms. The raw gains correlate with utility in the opposite direction to the share: the reconstruction and noise gains are strongly positive against utility, while the share and the excess are negative. A transition that spends a larger *fraction* of its gain on noise having a lower utility gain is a statement about where the capacity went, not about how much noise was absorbed in total, even when its absolute noise gain is larger.

Replica-budget calibration. An independent pilot of at most 8 codecs calibrated the replica budget against the criterion that the interval of the 2048-minus-4096 corrected state gain lies inside a predeclared band. The corrected Monte-Carlo bias fell 3.67×10^{-4}, 9.16×10^{-5} and 4.58×10^{-5} across the three rungs, and 2048 replicas were adopted. The finest rung is not a known truth, so this bounds run precision and does not establish analytic accuracy.

E.3 Reader robustness and penalty sensitivity

We re-evaluate utility with nonlinear readouts and then widen the ridge penalty grid. These checks test whether the allocation–utility relation depends on the original reader choice.

Nonlinear readouts. The prognostic population (B1, 220 existing codecs) is re-read with ridge regression, a two-layer MLP and a one-layer GRU. All see the same window block. They share a training budget and 2 paired initializations, with context lengths 4 and 32 for the prognostic and cross-generator populations, respectively. The same comparison is made on the cross-generator population (B2, 96 codecs).

Both nonlinear readers support the negative direction on the prognostic population at per-interval alpha 0.025. All 4 cross-generator claims are supported at alpha 0.0125.

population / dynamics	readout	stratified estimate	interval
prognostic	MLP	-0.720	[-0.804, -0.530]
	GRU	-0.702	[-0.792, -0.506]
cross-generator / nonlinear	MLP	-0.629	[-0.857, -0.600]
	GRU	-0.629	[-0.886, -0.629]
cross-generator / seasonal	MLP	-0.657	[-0.829, -0.043]
	GRU	-0.743	[-0.914, -0.086]

Within the prognostic population, correlations are negative in all 4 regimes for both of the 2 nonlinear readers: -0.762 to -0.667 for the MLP and -0.786 to -0.571 for the GRU.

These readouts use a shared per-channel scaler, whereas the original prognostic readout uses a per-feature scaler. B1 therefore evaluates existing codecs on fresh test panels under a changed interface, rather than reproducing the original endpoint. B2 reuses the cross-generator codecs, test panels and audit ledger, changing only utility; it is not an independent replication. B1 and B2 are analyzed separately. Their direction statistics retain the shared capacity-trend limitation described in section E.2.

Ordering. A further comparison uses no excess at all: whether the readers agree on the *ordering* the capacity ladder induces. Kendall's τ_b between the utility transition gains of each reader pair, averaged with equal weight over strata, is positive with a lower interval bound above the predeclared 0.5 bar in all 6 intervals that were computed; their lower bounds run from $+0.571$ to $+0.833$, at a per-interval alpha of 5.56×10^{-3}. The 0.5 threshold is a practical criterion chosen for this study. Agreement on capacity ordering leaves state fidelity and task utility as distinct quantities. The protocol declared a family of 9 ordering claims — three populations crossed with three reader pairs — and only 6 were computed: the second population averages its two generators into one point per pair, so the per-generator ordering intervals the protocol names were never produced. The interval alpha is the Bonferroni allocation for the declared 9-claim family applied to the 6 computed claims, giving a conservative allocation. The three missing claims remain not computed.

The protocol also declares a band of $|G_U| \leq 0.005$ in rank-information-coefficient units inside which a transition's utility gain is treated as sign-unstable. On the first population 36 reader-transition entries fall inside it, concentrated on the widest rung of each ladder where the gain is smallest. They are flagged and kept. The band never enters the Spearman or τ_b statistics, and it is never used to drop a transition from any endpoint. The learned readers themselves are fitted for at most 2000 updates. On a target that is nonlinear in the window the MLP reached a mean per-step cross-sectional rank information coefficient of 0.764 against the ridge readout's 0.489, so the learned readers are not merely undertrained. The GRU did not outperform the ridge readout on this target.

The extended penalty grid. The ridge penalty is selected from a fixed grid of 9 points on inner training panels. In 71 of the 96 codecs the selected penalty is the largest grid point, so the selection is at the boundary rather than interior. This shows that the optimum on the searched grid lies at its edge; a wider grid could change the selection. The preregistered grid is retained for the primary analysis, and the 96 frozen codecs of the cross-generator study were re-read with the ridge penalty selected from a grid extended by three decades: `logspace(-4, 7, 12)` against the frozen `logspace(-4, 4, 9)`. The extended grid is a strict superset, so the two selections are nested; features, scaler, inner split, selection metric, target, test panels and the audit ledger are unchanged, and only the refit changed. The frozen-grid refit reproduces the selected penalty and `reader_mean_rank_ic` exactly (maximum absolute deviation 0.0 over 96 codecs) and the published boundary count of 71 of 96.

On the extended grid the largest point is selected 0 of 96 times, and 21 penalties move, all to the three new points. Mean inner RankIC changes by $+1.02 \times 10^{-4}$, mean test utility by -3.52×10^{-5}, and the maximum absolute test-utility change is 1.29×10^{-3}. Rank agreement is 0.9997 over the 96 utilities and 0.9948 over the 24 transition utility gains. Both generators' stratified Q_g are identical to four decimals (-0.6571 for g1 and -0.8286 for g2); pooled g1 moves from -0.3636 to -0.3287 and pooled g2 is unchanged at -0.2378. A smaller amplitude is not the mechanism, because RankIC is invariant to positive rescaling; a larger penalty changes the coefficient direction, and the measured direction agreement is 0.995. The claim remains limited by the shared per-channel scaler, the two learned readers and the near-rank identity between the measured share and the excess. The frozen grid's numbers remain the published ones; this is a post-hoc sensitivity.

These checks cover two nonlinear reader families and the stated penalty grids. The interval correction and execution records are documented in section H.2.

E.4 TOKENIZER CONTROLS: SCALAR BINNING AND CAUSAL FILTERING

A training-free scalar ladder and a limited causal encoder test whether noise dominance depends on learned quantization or a pointwise receptive field. The scalar study is descriptive; the temporal comparison was predeclared locally.

Scalar-binning control. A two-stage residual scalar quantizer with fixed grids—uniform (equal-width bins over the training $[0.001, 0.999]$ quantile range) and quantile (equal-frequency bins)—was applied to the frozen Exp 39 generators (24 transitions) and the three frozen VAR regimes (Exp 16) (18 transitions), 42 transitions in all. No model is trained. The estimator, noise replay, second-moment-matched reference and ridge readout are the frozen ones; the reader sees the dequantized reconstruction (4 dimensions per step) instead of a 64-dimensional latent, which is declared. Each of the 4 channels and each of the 2 stages uses 1–4 bits, so the nominal rate is $8b$ bits per time point and the ladder's capacity levels are 8, 16, 24 and 32; the realized occupancy is the joint entropy of the packed 4×2-digit per-step code, not the sum of the per-channel entropies. The readout keeps a 32-step history of that reconstruction, a different interface from the four-step latent history the core reader uses, so this control's utility column is not on the same footing as the learned arms'.

quantity	scalar ladder	Exp 39 learned arms (frozen)
transitions	42	24
median measured noise share A_ε	**0.9985**	0.899 / 0.900
share range	0.885–1.006	—
share above one half	**42/42**	23/24
excess below the reference	**20/42**	24/24
excess above the reference	**22/42**	0/24
median excess	$+0.0001$	negative (all)
pooled $\rho(E_\varepsilon, G_U)$	**+0.378**	-0.364 / -0.238

The breakdown by quantization scheme and environment is:

group	median share	median excess	below / above ref.	pooled ρ
uniform	0.9995	$+0.0009$	5 / 16	$+0.699$
quantile	0.9885	-0.0103	15 / 6	$+0.484$
cross-generator	0.9995	—	9 / 15	$+0.575$
VAR regimes	0.9877	—	11 / 7	$+0.236$

The noise dominance reproduces and is stronger: a codec with no learning at all shows the pattern at least as clearly. "Below the reference" does not reproduce (20/22 split, median excess $+0.0001$), and the excess–utility association changes sign ($+0.378$). This control is not an implementation of Chronos: the official quantizer was not read for this study, and the ladder is a two-stage residual scalar quantizer written here. The comparison against the learned arms is a descriptive nearest-neighbour match on the realized occupancy increment alone: it matches neither occupancy level nor any rate–distortion condition, and its only rule is to take the closest row in the same cell. Over the 24 frozen Exp 39 pairs the absolute gap between the two increments runs from 0.210 to 4.896 bits per time step, median 3.229. These gaps limit the comparison to descriptive nearest neighbors; they do not establish cross-mechanism rate equivalence.

Causal temporal encoder control. 96 codecs—the frozen Exp 39 grid, 2 generators $\times$ 2 noise levels $\times$ 2 mechanisms $\times$ 4 widths $\times$ 3 seeds—were retrained with one change: the encoder gains a depthwise causal 8-tap filter, initialized to identity, after its GELU. Every shared module starts from the pointwise model's own initialization; the token rate per time step is unchanged (one token per step, $2w$ bits); the decoder, quantizers, objective, budget, seeds, data, reader recipe and measurement are frozen. The pointwise arm is the frozen Exp 39 population, verified by hash (192 assets). The window-wise noise replay gives both arms the same window ends and the same noise, so the comparison is paired at the evaluated step.

	pointwise (frozen)	temporal (new)
parameter count	27,972–158,532	29,700–160,260(+1,728)
test reconstruction MSE, mean	0.24575	0.24564
paired test MSE ratio, median / mean	—	**0.9998** / 0.9999
paired test MSE ratio, range	—	0.958–1.033
median measured noise share	0.8993	0.8968
share range	0.4985–0.9798	0.4964–0.9797
transitions below the reference	24/24	24/24
median excess	-0.0527	-0.0517
median utility gain	0.0131	0.0134
median reconstruction gain	0.0775	0.0766
median occupancy gain (bits)	1.7662	1.7688
pooled $\rho(E_\varepsilon, G_U)$	-0.2687	-0.2635
descriptive Q_g, g1 / g2	-0.6571 / -0.8286	-0.6000 / -0.8286

The temporal arm's training-time validation reconstruction MSE averages 0.24640; the frozen Exp 39 records do not preserve the pointwise arm's validation MSE, so it is not compared with it. The fit comparison in the table is instead the like-for-like test-side measurement: both columns are produced by this study's own measurement, on the same test windows and paired over the 96 codec pairs. On that measurement the arms are closely matched (paired ratio median 0.9998, mean 0.9999, range 0.958–1.033).

The primary endpoint is the SNR-stratified mean paired change in the marginal noise share, $\Delta\bar{A}_{\varepsilon g} = \frac{1}{2}\,\mathrm{mean}\{\Delta A_\varepsilon \mid \mathrm{SNR}\ 0.1\} + \frac{1}{2}\,\mathrm{mean}\{\Delta A_\varepsilon \mid \mathrm{SNR}\ 1.0\}$ with $\Delta A_\varepsilon = A_\varepsilon^{\mathrm{temporal}} - A_\varepsilon^{\mathrm{pointwise}}$, expressed in share units (97.5% paired bootstrap over initialization and panel, with a shared draw for both arms and a Bonferroni allocation across the two generator claims): g1 gives $\Delta\bar{A}_{\varepsilon g} = -0.00071$ with interval $[-0.00158, +0.00026]$, and g2 gives -0.00020 with $[-0.00171, +0.00090]$; both are inconclusive. The intervals constrain the paired effect on the noise share to roughly two tenths of a percentage point on this grid. The Q_g correlations in the table are descriptive secondary statistics of a different estimand, the allocation–utility stratified rank correlation. The frozen endpoint record names both quantities `DeltaQ_g`; the two are not one quantity, and the share change above is the one the interval bounds. The ledger is otherwise unchanged: 24/24 remain below the reference, the median share moves by 0.0025, the median occupancy gain by $+0.0026$ bits, and the association with utility survives (pooled -0.263 against -0.269; stratified Q_g moves only on g1). Fit quality is closely matched on the paired test measurement, so differential fit quality does not explain the near-zero point estimate; the paired endpoint nevertheless remains inconclusive. This comparison covers one architecture change, two controlled generators and one budget. Patch-based, convolutional and spectral tokenizers remain untested.

E.5 Matched-pair negative control

The matched-pair study attempts to train two arms with different allocations but similar reconstruction gains and realized occupancy increments, with a symmetric relative reconstruction tolerance of 0.10 and an absolute joint-entropy tolerance of 0.05 bit. At three initializations the group-level object matched on both tolerances (relative reconstruction difference 0.053, joint-entropy difference 0.0144 bits), although only 1 of the three per-initialization pairs did. At ten initializations the group-level match fails both tolerances (0.202 and 0.0912 bits) with 1 of 10 per-seed pairs matching. Those tolerances are the pilot's own decision rule and were originally explored with some results already visible; they are exploratory tolerances rather than formal thresholds. The extended study's interpretation rule was predeclared with three outcomes, and because no group-level match holds at ten initializations the outcome is *inconclusive on matching*.

The reported occupancies use paired-stage joint entropy. The merged-histogram correction and the pilot's timing and source limitations are recorded in section H.2.

E.6 Allocation at matched reconstruction and occupancy diagnostics

This paired training contrast tests whether the audit distinguishes allocations when reconstruction and occupancy diagnostics agree. Its protocol, matching bands and decision table were fixed before

the formal fits. A separate pilot selected the regularization strength using matching statistics only, without reading allocation shares.

We use the nonlinear generator of section E.2 at target SNR 1.0, residual binary quantization, one capacity span (per-stage widths $4 \to 7$, a nominal step of six bits per token) and one training budget. Five paired initializations, shared across the two conditions and the two capacities, give 20 formal codecs; the data seeds are new and disjoint from every earlier population.

Training conditions. For a state s let $x^{(1)} = s + \varepsilon^{(1)}$ and $x^{(2)} = s + \varepsilon^{(2)}$ be independent draws of the declared observation-noise law. Both conditions train on both views against the *observation*,

$$L(\lambda) = \tfrac{1}{2}\big[L_{\mathrm{rec}}(x^{(1)}) + L_{\mathrm{rec}}(x^{(2)})\big] + L_{\mathrm{q}} + \lambda \left\| f(x^{(1)}) - f(x^{(2)}) \right\|^2, \tag{23}$$

where L_{q} is the quantizer term the frozen objective already carries. The plain arm is $\lambda = 0$; the consistent arm adds the consistency term at $\lambda = 0.10$, the same value at both capacities, chosen by the pilot as the largest candidate whose reconstruction gain and level differences sat inside the bands. The state s is never a reconstruction label in either arm.

Matching. The three bands were frozen before the formal test: a relative difference of at most 0.10 in the transition reconstruction gain, at most 0.05 bit in the realized occupancy increment, and at most 0.15 in the reconstruction error at each capacity. The gate is applied on the development panels, and there the two arms sit inside all three: their reconstruction gains differ by 0.7%, their occupancy increments by 0.0006 bits, and the two capacities' reconstruction-error levels by 2.6% and 4.6%. These are practical matching bands with a deterministic decision rule, not a formal equivalence test. Both arms satisfy the rule, giving the verdict *matched*.

Result. The ledger below is measured on the test panels. The gate above is applied on the development panels; on the test panels the same descriptive verdict, *matched*, holds.

	plain	consistent
reconstruction gain G_x	0.1096	0.1089
occupancy increment ΔH (bits)	2.333	2.332
state gain G_s	0.0250	0.0290
noise gain G_ε	0.0846	0.0800
noise share A_ε	0.772	0.734
excess E_ε	-0.159	-0.197

The arms allocate that gain differently: the consistent arm moves share from the noise component to the state component, a paired change of -0.038 with a 95% interval $[-0.048, -0.027]$ over the 5 initializations. Both arms sit below the same matched reference (0.931), so the excess moves with the share. Exact enumeration of the $n = 5$ seed bootstrap — all 126 multinomial atoms — leaves no atom on the other side of zero (support -0.055 to -0.018). The paired utility change is -0.00003 rank information coefficient, which is consistent with no movement and is reported without a direction.

This is a controlled demonstration for one generator, mechanism, capacity span, budget and intervention. It establishes a detectable allocation difference at matched diagnostics, with no directional utility claim. It establishes neither a general capacity-selection rule nor a causal effect of allocation on utility. The earlier matched-pair study remains inconclusive on matching (section E.5).

F Additional utility analyses and capacity selection

F.1 Prediction interface and decoding

Three secondary studies bound how the utility account may be read.

A preregistered prediction-interface study held tokenizers, information set and target fixed and compared the closed-form ridge readout against a decoder-only model over discrete tokens with a capacity-growing output layer, and against a supervision-source control. Under the corrected scoring convention 4 of 18 transitions confirm a negative interface difference at the Bonferroni-adjusted level and 15 of 18 have negative point estimates, with no positive confirmation; intervals containing zero

are reported as unconfirmed. The scoring convention that produces this count was declared after the test ran, layered on a preregistered contrast, and the original convention confirmed five transitions. A single-arm increment is not evidence for the contrast in either version.

Replacing sequential greedy decoding with an expectation over the fine stage was tested as a development pilot. Its predeclared scaling criterion required a negative greedy capacity gain together with a full-expectation recovery of at least 0.01 rank information coefficient. The greedy gain was -0.0217, but the recovery was -0.0154, so the criterion was not met and the pilot was not scaled up. This does not show that decoding is irrelevant; a mechanism claim would need fresh training and independent confirmation.

Scores use average ranks for ties, zero for exact constants and the full date axis; non-finite inputs are rejected. The post-hoc scoring correction and its effect on earlier values are documented in section H.2.

F.2 Predictive and state accounts

The indirect reference. The ledger's reference of section 3 is a second-moment-matched object at the model's own occupancy, and it answers a reconstruction question: a code of x_t is asked to reconstruct x_t. The selection rule of section F.4 needs a different object — the error with which a code of the *current* observation conveys the *same-time* state s_t — and that is the indirect, or remote-source, Gaussian rate–distortion problem, which is closed form. Decompose $s = Ax + n$ with $A = \Sigma_s \Sigma_x^{-1}$ and n independent of x. The unpredictable part n survives every code, and the encoder's remaining task is the rate–distortion problem for x under the weighted distortion $\mathbb{E}\|A(x - \hat{x})\|^2$, which reverse water-filling solves once the weights γ_i are read off from $\Sigma_x^{1/2} A^\top A \Sigma_x^{1/2}$. Writing $\Sigma_{s|x} = \Sigma_s - \Sigma_s \Sigma_x^{-1} \Sigma_s$ for the conditional state covariance, the reference error at rate R and its own water-filling level θ_R are

$$e_{\mathrm{ind}}(R) = \frac{1}{d}\Big[\operatorname{tr} \Sigma_{s|x} + \sum_i \min(\gamma_i, \theta_R)\Big], \qquad R = \sum_{\gamma_i > \theta_R} \tfrac{1}{2} \log_2 \frac{\gamma_i}{\theta_R}, \tag{24}$$

so the rate water-fills over the γ_i, the eigenvalues of $\Sigma_x^{1/2} A^\top A \Sigma_x^{1/2}$, whereas the reconstruction reference of section B.4 water-fills over the eigenvalues of Σ_x itself. The two optimize different distortion objectives and their outputs are not interchangeable. At $R = 0$ Eq. (24) is the state variance $\operatorname{tr} \Sigma_s / d$, and as $R \to \infty$ it falls to the observation-noise floor $\operatorname{tr} \Sigma_{s|x} / d$; both limits are asserted to 10^{-9} relative on every regime before the object is used. Its measured counterpart is the held-out linear readout of section F.4, whose error is an upper bound on the conditional mean's and therefore makes the measured value an upper bound on what the reference offers. The prognostic study of section 5.4 measures its state-information account against this same object, and the selection rule of section F.4 scores its capture against it; both use it for the same-time state, not for a future one. It is an observation on this population under a Gaussian-source assumption rather than a general invariance, and it supplies no quantity to the ledger. Two further studies ask whether a *predictive* account — one built on the conditional mean of the future rather than on the state — would separate from the state account on this generator.

The first study is post-hoc and from the same session (Table 11). It rebuilds the oracle conditional mean in closed form from the generator, and it reproduces the frozen state account before trusting anything new. Across 18 transitions the two accounts agree in sign 18/18 times, the within-cell rank correlation between them runs 0.939 to 1.000, and the ratio of the two accounts is a constant per cell with no trend along the ladder. This has a structural *explanation*, though the argument is not a proof: with a linear target read by a linear reader, the predictive account is the state account re-weighted by a fixed matrix. A fixed re-weighting on its own does *not* force two configurations to keep the same ordering — that additionally requires the error's anisotropy to stay fixed as capacity grows, which is a property of this population rather than of the algebra.

A preregistered follow-up tested an anisotropic low-variance target and predicted that a low-variance channel's ratio would rise with capacity. The prediction failed: 0 of 18 low-variance comparisons separate, the median absolute change in the ratio over those comparisons is 0.0024, below its own standard error at a median ratio of 0.94, and exactly 1 of 36 cell-by-channel comparisons flag, on the reconstruction-aligned channel with the opposite sign. That is what a two-standard-error screen

over that many tests is expected to produce by chance. On this generator, then, the two accounts were not separable, the measured anisotropy was stable along the ladder, and the preregistered low-variance-target test did not separate them. These findings concern this generator and target construction; they leave predictive-capacity rules on other generators unresolved.

F.3 Incremental predictive value

Adding the state account did not improve out-of-regime prediction on this grid. The incremental-value entry in Table 12 records the corrected auxiliary regression behind this result. Its endpoint is a difference of pooled out-of-fold mean squared errors, not a regression coefficient. After correcting the family labels, while retaining the specification, population, folds, penalty grid and replicate count, the regression gives a base out-of-fold MSE of 3.62×10^{-4} and an enhanced one of 5.32×10^{-4}, an MSE change of -1.69×10^{-4} against a superseded -6.29×10^{-4}, with an interval of $[-8.88 \times 10^{-4}, -1.18 \times 10^{-4}]$. Adding the state account increases held-out error: the MSE change is defined as base minus enhanced, so a negative value means worse prediction. The selection rules score a different object, the validation state-information capture, and do not outperform direct task validation on this grid.

The response of that regression is the seed-mean utility gain G_U of a transition, and its base design already carries the reconstruction gain G_x, the realized occupancy increment ΔH, the lower rung's nominal bits per stage, the step between the two rungs and the quantizer family. The test therefore asks whether the state account adds predictive value beyond these conventional diagnostics. The primary design adds G_s; the predeclared secondary design adds the excess share $A_\varepsilon - A_\varepsilon^{\mathrm{ref}}$ instead. G_ε is not entered, because $G_x = G_s + G_\varepsilon$ makes it collinear with the two accounts already present, and no combination is selected after the fact. The outer loop is leave-one-regime-out over 4 folds, so a whole ladder stays on one side of the outer boundary, and the penalty is chosen inside the training folds by five-fold cross-validation on a fixed grid with the same folds and seed for every design. The random inner folds can split transitions sharing an intermediate rung, a limitation retained from the declared design. Only the primary comparison has an interval; the secondary is a point estimate.

For the secondary account, the base is refit on the same sample as the enhanced model. Adding the excess share also increases error. The stored values are:

secondary comparison	out-of-fold MSE
base	0.0003623761875
enhanced	0.000603101609
change (base minus enhanced)	−0.0002407254215

The primary comparison improves in 3 of 4 folds, so its direction also varies across regimes. This grid establishes no stable incremental predictive value. That finding concerns added prediction under the specified regression, not the existence of the association reported in section 5.4.

F.4 The selection rules and the regret

State-information capture. The audit-derived rule scores each rung by a normalized drop in same-time state reconstruction error,

$$\text{capture} = \frac{\sigma_s^2 - \hat{e}_{\mathrm{meas}}}{\sigma_s^2 - e_{\mathrm{ind}}}, \tag{25}$$

where $\sigma_s^2 = \operatorname{tr} \Sigma_s / d$ is the state variance, i.e. the error of predicting s_t by its mean. Both numerator and denominator are therefore mean-squared-error drops relative to that variance baseline, and the quantity is a normalized variance-explained ratio, not a ratio of mutual informations.

Two objects in Eq. (25) are distinct from the ledger's, and the distinction is the point. The measured term $\hat{e}_{\mathrm{meas}}$ is the held-out ridge mean squared error of reading the *same-time* state s_t from the *same-time* code, fitted on the training panels and scored on the validation panels; its penalty is selected on an inner slice by state MSE, not by the per-step cross-sectional rank information coefficient the utility probe maximizes, so it is neither D_s nor a next-step utility probe. The reference term e_{ind} is the *indirect* (remote-source) Gaussian reference of section F.2: the error with which a memoryless

Gaussian code of the observation x_t at the rung's own realized joint-token entropy conveys s_t, the same-time state. It is not the reconstruction reference of section B.4.

Within that stated problem — the Gaussian surrogate, that rate and squared loss — e_{ind} is the best a memoryless code can do, so the denominator is the largest drop available and the ratio is normalized to $[0, 1]$ there. Read on the actual population it is bounded more weakly. 0 says only that this linear readout has no lower MSE than the variance baseline, which does not establish that the code carries no state information at all. 1 says only that the measured error equals the Gaussian reference, which is not global optimality on a non-Gaussian source. The value is not clamped. A ridge fitted in-sample and scored out of sample can fall marginally below the reference, so values slightly above one occur, and that over-shoot is not by itself evidence of a defect. The capture is a ratio, it is not the reference-relative excess E_ε, and it does not enter the ledger.

The five rules. Let b range over the available rungs and let $\bar{q}_b$, $\bar{r}_b$ and $\bar{c}_b$ be the mean over seeds at rung b of the validation rank information coefficient, the validation reconstruction error and the validation capture. *Largest* takes $\arg\max_b b$. *Lowest reconstruction error* takes $\arg\min_b \bar{r}_b$. *Highest validation utility* takes $\arg\max_b \bar{q}_b$. *One standard error* takes $b^\star = \arg\max_b \bar{q}_b$, the seed spread $s_b = \mathrm{std}(\{q_{b,\mathrm{seed}}\}, \mathrm{ddof} = 1)/\sqrt{n}$, the threshold $\theta_{1\mathrm{SE}} = \bar{q}_{b^\star} - s_{b^\star}$, and then the *smallest* rung with $\bar{q}_b \geq \theta_{1\mathrm{SE}}$. *Highest capture* takes $\arg\max_b \bar{c}_b$. All five depend only on validation-side scores.

Regret. With $\bar{U}_{\mathrm{test}}(b)$ the seed-mean test utility at rung b, the oracle rung is $b^\star = \arg\max_b \bar{U}_{\mathrm{test}}(b)$ and regret $= \bar{U}_{\mathrm{test}}(b^\star) - \bar{U}_{\mathrm{test}}(\hat{b})$. The table reports the mean of that quantity over the 12 cells together with the count of cells off the optimum. It is a descriptive mean: no interval is computed over cells, so the entries carry no paired uncertainty. Mean regret is 0.00116 for the largest rung, 0.00057 for the one-standard-error rule, 0.00126 for the audit-derived rule and 4.71×10^{-6} for reading the task on validation. Table 10 reports all five rules side by side.

selection rule	mean test regret	cells off best
largest available rung	0.00116	2
lowest validation reconstruction error	0.00116	2
one-standard-error rule	0.00057	6
highest validation utility	4.71×10^{-6}	**1**
highest state-information capture (audit-derived)	0.00126	3

Table 10: Capacity-selection comparison over 12 mechanism–regime cells. Mean test regret is the best test utility in each cell minus the utility of the selected rung, averaged over cells and initializations. Values are descriptive post-hoc means; between-rule differences are not tested.

G External and public-data studies

G.1 Beijing multi-site stress test

A second external study injects known i.i.d. Gaussian observation noise into the clean Beijing multi-site air-quality series (12 stations, 4 channels, 28513 hourly timestamps; per-season 60/20/20 splits; blocks of 168 retained hourly observations; train-only global per-station/channel normalization, no window statistics). A block is cut from the sequence of retained records and is not guaranteed to be 168 consecutive natural hours: a timestamp survives only if every station reports every channel after gap filling, so a block can span a real-time gap, and the readout's next step is the next retained record rather than necessarily the next hour. Three cells (clean control, target SNR 0.3 and 1.0) $\times$ two mechanisms (RBQ and VQ-VAE) $\times$ four widths $\times$ 2 seeds give 48 codecs under one shared budget; each noise cell has 24 transitions.

Endpoint and validity gates. The primary endpoint is the stratified Spearman $Q = \frac{1}{2}\rho(E_\varepsilon, G_U \mid \mathrm{SNR}\ 0.3) + \frac{1}{2}\rho(E_\varepsilon, G_U \mid \mathrm{SNR}\ 1.0)$ with $H_1 : Q < 0$; a 10000-draw season/panel/seed bootstrap gives $Q = -0.501$ with a 95% interval (2.5/97.5 percentiles) $[-0.574, -0.138]$, and 24/24 and 23/24 transitions fall below the reference at the two noise levels (per-cell $\rho = -0.545$ and -0.457). The clean cell's corrected noise account is exactly zero (worst per-model mean 0.0), the noisy-cell

reference residuals are positive semidefinite (minimum 0.815), the injected-noise means and variance ratios lie in $[-0.0050, +0.0126]$ and $[0.9951, 1.0094]$, and every fit shares the declared budget. The reported endpoint uses the season-independent noise realization and multiplicity-weighted bootstrap recorded in section H.2.

Time-axis construction. Missing values are filled by linear interpolation that fills at most six consecutive missing positions from each end of a gap, so a longer gap can be filled from both sides and "only gaps of six hours or less are filled" is not what the implementation guarantees; in the retained series a filled value is at most 6 records from an observed one (27962 values filled). Interpolation runs before the seasonal split, so a filled value inside one split can draw on an observation assigned to another split. The split-respecting sensitivity below measures the effect of removing this exposure. A post-hoc structural audit of the axis (`time_axis_audit.json`, written beside the frozen bundles and re-measuring no endpoint) finds 99.40% of the 28512 steps between adjacent retained records to be exactly one hour, with 172 steps crossing a gap and a largest gap of 889 hours. Of the readout's four-step windows and their next-step labels, at most 3.83% of the windows and 0.77% of the labels in a test split cross a gap (5.50% over every split, training included). The audit is a descriptive record of the construction: it is not a predeclared validity gate, and the reported endpoint is not re-estimated under a different time axis.

Robustness of the sign. The sign survives leave-one-season-out $[-0.564, -0.401]$, per-mechanism (-0.794 for RBQ and -0.378 for VQ-VAE) and leave-one-station-out $[-0.538, -0.456]$ checks. This is one city and one data domain with injected known noise, not a natural-noise identification, and it does not replace the inconclusive Saugeen endpoint.

Split-respecting fill. The interpolation above can draw across a split boundary, and the audit quantifies how often. A post-hoc sensitivity (Exp 50) removes that exposure by filling each split from its own observations only, keeping the archive, the split assignment and the measurement code unchanged; the fill runs one split at a time because masking the unassigned rows once still lets it bridge a boundary where two splits' edge rows are missing. No filled value then draws on an observation outside its own split. 29 of the 28513 retained records drop out because their fill is no longer reachable inside their own split — all of them in the training and validation splits, none in the test splits — and the endpoint moves from -0.501 to -0.480 with a 95% interval $[-0.593, -0.102]$. The sign is unchanged and the interval stays clear of zero in the same direction, so the cross-split fill is not driving the reading. This is a sensitivity on the same archive and the same measurement path, not an independent replication, and it does not replace the prespecified Beijing endpoint.

G.2 Saugeen stress test

The negative-association endpoint is inconclusive. It runs on 8 of the 9 designed cells — the predeclared rule excludes one whose reconstruction gain is not positive — where $\rho = -0.286$ with interval [-0.771, +0.571] and sign agreement 0.68, over 23741 daily points. It is not a replication success or failure.

The study injects known corruption into a real hydrological series. It compares a scratch-trained capacity ladder with a tokenizer pretrained on other public domains. The latter is our own pretraining, using the protocol's fallback because the official model zoo was unreachable. The executed design has the following limits.

item	executed design	interpretation
budget	Scratch: batch 512, 4000 updates, across forty-eight fits. Protocol: 1024, 5000. Pretrained arm: 1024, 5000, as planned.	The arms differ by 2.5 in sample-updates. Matching codebook size does not isolate pretraining.
source record	5 of nine source files differ from their recorded hashes; original versions are unrecoverable.	The executed versions cannot be verified against the current files. See section H.2.
side information	Window mean and standard deviation reach the decoder through a continuous path.	Occupancy measures token indices only; the scalar reference is a proxy.
VQ objective	Encoder-side weight 1; codebook-side weight 0.25. These swap the published paths, as checked by autograd.	The arm is architecture-inspired. Retraining with the published objective would be a separate study.

Budget entries come from the fit records; the written protocol is retained. The codec file matches the version identified by that protocol.

Side information and rate convention. The codec reconstructs as $\hat{x} = \mathrm{std}(x) \cdot \mathrm{decoder}(q(\mathrm{norm}(x))) + \mathrm{mean}(x)$, so the window's own mean and standard deviation enter through a continuous path that the token indices do not carry. A no-training test demonstrates it: holding the token indices fixed, translating a window by 1024 moves the reconstruction by the same amount and leaves the tokens and their dequantized vectors identical, with a residual below 10^{-12} in float64. The occupancy account is therefore a *token-index* occupancy that bounds the discrete code, not the pipeline's total rate. The scalar reference receives neither the side information nor the same reader interface, which makes it a declared proxy rather than the synthetic studies' comparison. One capacity doubling is 1 bit per token, 90 bits per window and 0.25 bit per time point; the nominal step and the realized occupancy increment are reported in the same per-window unit, with $\Delta H = R_{\mathrm{occ}}(C_2) - R_{\mathrm{occ}}(C_1)$. The corrected sign affects a secondary occupancy correlation, not the primary excess endpoint (section H.2).

Clean control. The corruption-free cells inject no observation noise, and the noise account is zero in all 12 clean fits of the scratch ladder. These zeros follow from the construction: without injected noise the estimator's noise term vanishes identically. The stored control flag is a tolerance test. Accuracy with nonzero noise is assessed separately by the failed finite-replica criterion in section C. The frozen multi-domain tokenizer's own clean fit is also zero and is not among the 12, because the recorded control counts the ladder's cells by name.

G.3 The public-data operational application

The operational application declares a state and a residual. The state is a causal per-channel seasonal mean — the mean of the same phase over the previous 7 days, at a seasonal lag of 96 rows for the dataset analysed — with a burn-in of 672 rows. The residual is resampled as independent noise per channel, and its lag-one autocorrelation is reported (0.793–0.990): the residual is temporally dependent, and independent resampling does not preserve that dependence. That is the diagnostic for how defensible that declaration is. Distortions are computed on 2048 sampled states at 512 replicas over 3 independent noise repeats.

Two public datasets were carried into the study; the predeclared guard on the state construction fails on the second, where the declared state does not track the observation, and only the first is carried forward. The reported implementation differs from the filter named in the protocol (section H.2). Because the split is a modeling choice, no number in this appendix is ground truth, and the state and noise accounts here are not comparable to the synthetic ledger's. On the dataset carried forward, the measured noise share of reconstruction error runs from 0.518 to 0.837, and the transition noise share of reconstruction gain from 0.344 to 0.663; measuring the predictive value of the state gain relative to the task increment would need more transitions than the study has.

G.4 Replay-law diagnostic

A stationary Gaussian process with every parameter declared was used to study the operational account under known truth: state AR(1) with $\phi = 0.9$; observation noise AR(1) with $\phi_e \in \{0, 0.5, 0.9\}$

plus a shared cross-channel factor; a lagged declared state built from the past; three codec families (pointwise and windowed nonlinear soft-threshold ladders, and a linear control); and four replay laws (iid, block, conditional, oracle). The frozen estimator is fed every draw unchanged. Two truths are computed independently of the estimator: the conditional truth $\mathbb{E}\|\hat{s} - \mathbb{E}[f(x) \mid \hat{s}]\|^2$ under the true law, and the physical truth with the true state in place of the declared one.

For the pointwise soft-threshold codec at the lower rung, the estimator's output D_s is 0.00518, 0.00579 and 0.00860 at $\phi_e = 0, 0.5, 0.9$, against conditional truths of 0.16563, 0.25201 and 0.15440—ratios of 0.0313, 0.0230 and 0.0557. Across all four replay laws and all three noise laws, target/estimator lies in 0.9873–1.0026, so the estimator faithfully converges to its own estimand; target/conditional truth lies in 0.0218–0.0606 for every law, including the conditional law (0.0301 at $\phi_e = 0$) and the oracle law (0.0441). The linear control's D_s is identical across all four replay laws (spread 0.0, 5.6×10^{-17}, 0.0 across the three noise laws), which is what the algebra requires.

The estimator forms $Y_j = (f(s + e_j) + f(s - e_j))/2$, so it converges to the conditional mean under the symmetrized replay law. A replay law whose draw carries a nonzero mean given the declared state—exactly what the conditional law adds—has that mean cancelled by the pairing. When the declared state is built from the past, the residual is not conditionally centred given it (the fitted residual predictor explains 7–11% of the residual variance in this process), and the measured state account is 2–6% of the natural estimand. At $\phi_e = 0.9$, block replay preserves the residual's within-window lag-one autocorrelation (0.79 against 0.80) and iid does not (0.0005), yet both give the same accounts: what matters is not the temporal structure of the replay but that the replay is drawn independently of the declared state. This known-truth diagnostic shows why an operational state account can be substantially below the declared-state fidelity. It does not estimate bias in any real dataset. The synthetic populations satisfy the distinct conditions of known latent state and conditionally zero-mean observation noise.

The implemented ladder differs from the protocol's description; the executed design and its recorded discrepancy are listed in section H.2.

G.5 The motivating production system

The benchmark was motivated by an equity-forecasting pipeline in which a tokenized representation is trained first and a downstream sequence model is trained on its tokens. Over a long continuation run of that system the tokenizer's own health indicators improved and its reconstruction error kept falling, while the downstream ranking quality did not follow. That is the observation that led to asking what a capacity increment preserves. No number from that system appears in this paper: its evaluation window has been consumed and its data come from Wind, so it serves only as motivation and carries no benchmark evidence. Its protocol and scoring conventions are described in section F.1 only to the extent that the scoring convention was later adopted by the benchmark.

H Study status and correction record

H.1 Study status and provenance

Table 11 records how each study was specified. A *preregistered* study fixed its population, statistic, interval construction and decision rule in writing before the corresponding models were trained. A *predeclared locally* study fixed the same things in a protocol frozen in this repository before its formal fits, but registered it nowhere outside the repository, so its predeclaration is a local record rather than a third-party one. A *post-hoc* study was specified after the data it reads already existed. A *characterization* reports measurements without a directional claim. A *failed-criterion* entry means a predeclared criterion was not met. *Superseded* means the value has been replaced and is not quoted as a current result. Table 11 gives the study behind each section of this paper and the way it was specified.

Every number in the main text is emitted by a generator script that reads the frozen results of those studies; none is transcribed by hand.

The replay-law diagnostic is a characterization of a method under known truth rather than a population measurement.

where	study	status	reads
replica ladder	calibration ladder	post-hoc	calibrated accounts
construct validity	known-fraction audit	preregistered	calibrated accounts
dose response	training-noise dose	characterization	dose ladder
Gaussian control	Gaussian construction	post-hoc	control fits
K1/K2	closed-form quantizer	preregistered (K2 failed)	analytic rows
ledger	extended ladder	characterization	all fits
prognostic	disjoint-population test	preregistered	fresh fits
cross-generator	nonlinear and seasonal dynamics	predeclared locally	fresh fits
reader robustness	MLP and GRU readouts	predeclared locally	frozen codes, fresh panels
regret	selection-rule replay	post-hoc	frozen population
causal isolation	matched pairs	preregistered (inconclusive)	matched fits
operational	public-data application	post-hoc	public tokenizers
external test	corrupted real series	preregistered (inconclusive)	external fits
prediction interface	interface study	prereg. contrast, post-hoc scoring	interface arms
decoding	decoding pilot	development pilot (criterion not met)	one panel
oracle-mean closure	oracle conditional mean	post-hoc (same session)	frozen population
anisotropic target	low-variance channel	preregistered (prediction failed)	frozen population
incremental value	auxiliary regression	post-hoc	frozen population
exact bootstrap	seed-support enumeration	post-hoc	frozen prognostic population
capacity conditioning	within-stratum/partial association	post-hoc	frozen prognostic + cross-generator populations
ridge sensitivity	extended penalty grid	post-hoc	frozen cross-generator codes
scalar-binning control	training-free scalar ladder	characterization	frozen generators + VAR regimes
temporal encoder	causal receptive-field control	predeclared locally	fresh fits
replay-law diagnostic	known-truth replay study	characterization	synthetic known-truth process
multi-site stress test	injected-noise multi-site series	predeclared locally	fresh fits
allocation discrimination	training contrast at matched diagnostics	predeclared locally	fresh fits
split-respecting fill	post-hoc sensitivity on the multi-site study	post-hoc	reused archive and measurement

Table 11: The studies this paper draws on and how each was specified. Where a superseded value is retained in the appendix, it is identified alongside its replacement.

H.2 Corrections, superseded values and execution records

Table 12 pairs superseded numerical values with their current replacements. The records below explain the affected calculations and execution limits. These analysis corrections use existing measurements; original result bundles are retained alongside the corrected outputs. Study-specific reruns and new fits are identified separately.

Quantity	Superseded value	Current value
Occupancy-to-utility rank correlation	-0.10	0.10
Joint-token entropy (pilot)	6.5728	11.5435
Analytic realized rate, finest step	3.0315	6.3692
Analytic noiseless reference share	0.9850–0.9970	0.9908–0.9999
Prognostic interval	[-0.584, -0.474]	[-0.569, -0.468]
Incremental-value MSE change	-6.29×10^{-4}	-1.69×10^{-4}

Table 12: Superseded and current values. Current analyses use the right-hand column; earlier values are retained here to document the corrections.

Superseded values. The Saugeen endpoint interval remains [-0.771, +0.571]; the prognostic interval correction did not change it.

Protocol deviations in the analytic calibration. Two are recorded for section 4: its coarsest replica rung ran at 4 replicas where the protocol declared 2, and its offset codec was outside the protocol, which section C explains was added because a zero-offset uniform quantizer has a nearly unbiased conditional mean that would have left the state account untested. The runner also recorded a supplementary screen that judges each component against twice the larger of the two component standard errors: all 48 comparisons pass it. That screen is lenient for whichever component has the

smaller standard error, and it is not a re-statement of the criterion. The failed predeclared criterion and its withdrawal consequence remain as reported in section C.

Reader-study intervals and execution records. The initial bootstrap reduced an aggregate already indexed by draw, leaving only the reference to vary and producing zero-width intervals. Recomputing the intervals with the corrected reduction leaves all point estimates and verdicts unchanged. Per-panel accounts, joint counts, reference matrices and readout scores are unchanged. The superseded endpoints are retained. Earlier checks missed this defect because they did not test whether a draw moves the measured side; the check now includes that assertion.

After B1 completed, the runner's lock path was amended for B2. B1's recorded hash therefore identifies its executed source, not the current file. The reader runner's prepare command can overwrite an existing protocol, unlike the cross-generator runner's. The amendment is recorded with the results; a newly stamped protocol does not verify an earlier execution.

Analysis corrections.

analysis	correction	consequence
analytic output rate	Reconstruction values had been used as integer bin indices, folding tail mass into the end bins for steps below one. Rates now use the output-bin probabilities.	The superseded finest-step rate 3.0315 becomes 6.3692. Reference accounts, noiseless shares and share ratios use the correction bundle.
prognostic interval	The original bootstrap averaged per-seed ratios; the corrected bootstrap recomputes the ratio of seed means.	The interval matches the reported estimand (section B.5); this is an implementation correction, not a second test.
incremental prediction	Mismatched family labels made one indicator identically zero and merged two reference categories. The same specification was rerun with the actual labels.	Population, folds, penalty grid and replicate count are unchanged. The corrected negative MSE change remains in Table 12.
Saugeen occupancy	The original analysis subtracted the higher-capacity occupancy from the lower. The corrected increment is $R_{\mathrm{occ}}(C_2) - R_{\mathrm{occ}}(C_1)$.	The descriptive occupancy–utility correlation changes from -0.10 to 0.10. The primary excess endpoint is unaffected.

Matched-pair occupancy and execution record. The pilot's implementation computed a quantity it called joint entropy by accumulating both quantizer stages into one histogram of length V and taking its entropy. That is the average marginal codebook entropy $H\big((p_{Z_1} + p_{Z_2})/2\big)$, not the joint entropy $H(Z_1, Z_2)$: it is blind to the dependence between stages and returns the same value whether the stages are identical, independent or anti-correlated (analytic values 1, 2 and 1 bit against a constant legacy value of 1 bit). The corrected computation pairs the stages as $z_1 V + z_2$; the reference configuration's corrected joint entropy is 11.5435 bits against 6.5728 bits for the merged quantity. The affected evaluation was rerun on the same frozen checkpoints, and no tokenizer was retrained.

The corrected evaluation's lock binds the twelve checkpoint hashes it consumes, the train/validation/test data fingerprints and the frozen source hashes. The pilot's upstream lock postdates its checkpoints by 573 seconds, so it certifies input identity rather than implementation precedence; the pre-crash runner and its original lock are not recoverable, and we do not reconstruct them. The original lock is unchanged.

Rank-scoring correction. The utility account's scoring convention is a rank information coefficient on the full date axis, with average ranks under ties, exact constants scored zero by declared convention, and non-finite input rejected. That convention was established by re-scoring an earlier study whose implementation used ordinal ranks (splitting ties by entity order) and skipped constants. Applying it to the population of section 5 across all 270 model-splits changed the per-model rank information coefficient by at most 4.50×10^{-5} (median 5.76×10^{-8}), with a largest single-cell change of 0.0430 in one high-tie model, and left every capacity comparison unchanged. The correction is real but numerically immaterial on this sample. It does not follow that the project's scoring is unified: that would require re-scoring every other delivered experiment as well. The historical implementations

remain frozen and keep their ordinal values; they are not edited, and their numbers are not comparable to the ones reported here.

Saugeen execution verification. The source drift summarized in section G.2 prevents verification of the affected executed files against their current versions. Their original bytes are unrecoverable. The unsealing step performs no hash check of the protocol sources, checkpoints or probes before it is written, so it records the time of unsealing rather than a verified input list. The pre-run protocol establishes predeclaration, while execution conformance and present-day reproducibility remain limited by these missing checks and sources.

Beijing superseded run and bootstrap. The superseded v1 run and the protocol-conforming v2 rerun both met the prespecified support criterion ($Q = -0.568$ with interval $[-0.668, -0.259]$ for v1); the v1 noise realization was shared across seasons while v2 uses an independent realization per season, and the first-pass bootstrap used equal panel weights while the reported bootstrap weights the per-panel reference curves by the draw's panel multiplicities, a correction that left the point estimate and verdict unchanged.

Operational state and replay-law specifications. The public-data study in section G.3 executed a different state filter from the one named in its protocol. Its reported values follow the executed split's lock. The replay-law diagnostic's protocol names two linear ladders; the runner instead uses two nonlinear soft-threshold ladders and a separate linear control. The reported replay results follow that executed design. Both original protocols are retained with their discrepancies documented.

I REPRODUCIBILITY

All results are produced from frozen result bundles by generator scripts; the manuscript's macros and tables are emitted by those scripts, and a read-only verifier re-derives each headline number from the bundles and fails if a generated file disagrees with one. Distortion accounts are reported on the calibrated replica basis throughout and never from the uncorrected decomposition; the superseded decomposition is retained in the record and is not quoted. The external application's budget, provenance, occupancy direction, side information and implemented objective are read from the correction bundle, and the generator requires that bundle and raises if it is missing or stale.

The regime-constant generator checks its constants against the defining source; a mismatch requires correcting the table. Two further checks verify the reference implementation and derived population summaries. The reconstruction reference's coordinate system of section B.4 is verified against a two-dimensional case with a non-diagonal observation second-moment matrix and a non-isotropic state second-moment matrix whose accounting is available in closed form: the frozen function reproduces $\theta = 1/2$, $D_x^{\mathrm{ref}} = 1/2$, $D_s^{\mathrm{ref}} = 85/1024$ and $D_\varepsilon^{\mathrm{ref}} = 427/1024$ to machine precision and is invariant to rotating the problem, and a deliberately coordinate-mixed expression fails the same check. The by-regime summary in Table 8 is emitted by a separate derived-view generator that reads only the two frozen transition tables and the incremental-value correction bundle, refuses to pool the two populations, refuses to drop the sub-half and above-one transitions, and raises rather than borrow an interval the secondary account does not have. Code and frozen protocols will be released with the paper; the public datasets are used under their own licenses and the external series is redistributed by its original repository rather than by this work. The accompanying source archive supports new runs of the supplied implementations and a small end-to-end demonstration; exact reconstruction of the numbers reported here additionally requires the original frozen result bundles, and the archive states whether those accompany it.

The release separates two routes. Given the original result bundles, the first reconstructs this paper's numbers from the frozen tables on CPU, without loading models or training; the verifier checks those reconstructed numbers. The second measures a *new* transition. It is delivered as a short script set rather than as a supported installer: it builds a controlled environment, trains a small codec pair at two capacities, runs the estimator of section B.1 and the reference of section B.4 on them, emits a record conforming to the ledger schema, and checks the result against known answers — an identity codec whose accounts must vanish, the two decomposition identities, the reference's realized rate against the rate requested of it, the zero-observation-noise boundary, and a reverse-water-filling solver written independently of the one under test. The codec pair it trains is far smaller than this paper's and shares

none of its data, so the numbers that run writes are new measurements and not results of this paper. The release also states the ledger schema, the conditions under which a measurement is not defined, and which assets are frozen results rather than re-measurable models. The two routes verify different operations: reconstruction of reported numbers from available bundles, and new measurement with the supplied implementations. Together they do not establish whole-pipeline reproducibility.

The second route is codec-specific: it dispatches on the codec class, so it covers the supplied RBQ, VQ-VAE and analytic uniform quantizer; measuring another codec requires a compatible extraction path. The demonstration runs the estimator, reference and ledger schema end to end on new data and checks the outputs against known answers.

Study names in the source archive. Experiment identifiers locate implementations and records; the descriptive names below are used in the text.

study	source experiment
cross-generator dynamics	Exp 39
nonlinear readouts / penalty sensitivity	Exp 40 / Exp 43
exact seed enumeration / capacity conditioning	Exp 41 / Exp 42
scalar binning / temporal encoder	Exp 44 / Exp 45
replay-law diagnostic	Exp 46
Beijing multi-site / split-respecting fill	Exp 47 / Exp 50
matched training contrast	Exp 48