WHEN DOES QUANTIZATION FORM REUSABLE CONCEPTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization encodes learned representations as discrete tokens. When do joint encoder and codebook updates organize them into reusable concepts? A preserved property such as shape may still require complex token combinations to read across colors and viewpoints. We study reuse through one rule shared across contexts. Our principle is that parameter updates move assignment boundaries, and inputs crossing them alter subsequent updates and shared-rule errors. Reference models show how boundary geometry constrains factor selection and relative update rates change local stability and relaxation. In a frozen residual model, we bound the time until one token test reads the target without error across contexts. Learned quantizers preserve the target perfectly at the same token budget, yet have certified minimum error-free rule lengths of 15 and 34 bits under one grammar. These results connect training conditions to factor organization and distinguish target recovery from concise reuse.

1 INTRODUCTION

Quantization represents inputs with discrete indices called tokens. Quantized autoencoders can encourage factor separation (Hsu et al., 2023), and discrete hidden units support inspection and control (Tamkin et al., 2024). A reusable interpretation should remain useful as other factors change, such as one shape rule applied across colors and viewpoints.

Retained information and the complexity of using it are distinct aspects of representation quality (Eastwood et al., 2023). Two quantizers can preserve the same target through different token organizations. One admits a single-layer test, while another requires conditional tests across layers (Fig. 1), increasing the shared rule's description cost. We study concise cross-context use as one testable aspect of concept reuse. At a fixed token budget, how do the data distribution and joint encoder and codebook updates form and maintain organizations supporting such rules?

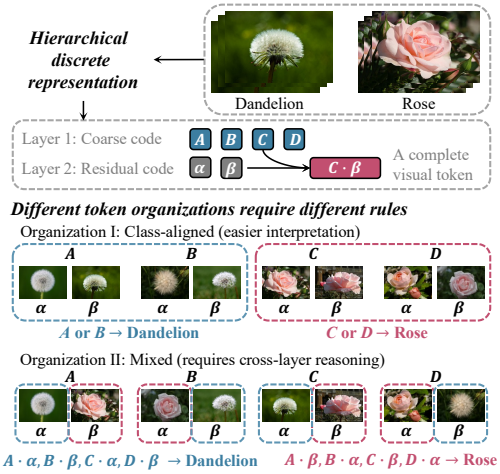


Figure 1: Shared rules for preserved flower identity (schematic), using single-layer or cross-layer tests.

Our starting point is feedback between assignments and updates. Parameter updates move assignment boundaries, exchanging inputs between tokens. The affected inputs alter subsequent updates and can correct or introduce errors of a shared rule. Boundary geometry and relative update rates shape this feedback. Tracking exchanges by target value and context connects organization formation to its use.

We use complementary settings to study factor selection and the cost of reading a preserved target. A two-factor reference varies boundary geometry while fixing selected moments and the distor-

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.

tions of two reference partitions. This yields local stability conditions and a relaxation time scale. Learned residual quantizers test how training organizes information retained in complete token tuples. Assignment exchanges connect the update mechanism to fixed-rule risk, while endpoint rule selection measures the access each organization supports. A frozen residual reference then links the distribution created by a preceding layer to the time an error-free token test becomes available.

Two quantizers trained from a common initialization preserve the target perfectly at the same token budget, with certified minimum rule costs of 15 and 34 bits under one grammar. Paired interventions show how short-rule access can improve or deteriorate between reliably readable endpoints. Residual experiments verify arrival bounds and show how preceding-layer motion changes formation time and token polarity. On frozen Shapes3D representations, combined response and assignment features improve intervention ranking. An electrocardiogram (ECG) measurement interface supports reliable shared-rule reuse across PTB-XL patients and phase contexts.

Our contributions are factor-specific stability conditions beyond matched distortion summaries, a connection between assignment feedback and shared-rule risk with learned cost certificates, and a residual rule-formation bound. These results connect the conditions that form and stabilize an organization to the costs and timing of using it.

2 Related Work

Unsupervised disentanglement requires inductive biases (Locatello et al., 2019). Hsu et al. (2023) combine learnable scalar quantization and strong weight decay, evaluating readout explicitness. Under structural assumptions, independent density discontinuities identify quantized factors through smooth transformations with smooth inverses (Barin-Pacela et al., 2024). These approaches seek factor distinctions that remain consistent as other factors vary. Codebook Features uses discrete hidden units for inspection and control (Tamkin et al., 2024), while concept bottlenecks use annotated concepts for test-time intervention (Koh et al., 2020). Matryoshka sparse autoencoders use nested reconstruction objectives to preserve features at several abstraction levels (Bussmann et al., 2025). We study how specified training updates form and maintain organizations supporting shared rules across contexts.

Straight-through and vector-quantization (VQ) studies examine surrogate gradients, encoder–codebook mismatch and collapse (Yin et al., 2019; Huh et al., 2023; Fifty et al., 2025; Zhu et al., 2025). Lu et al. (2026) study encoder drift, codebook lag, quantization error, encoder-rate interventions and assignment churn. Building on mode dynamics (Saxe et al., 2014), Zhao et al. (2026) analyze mode suppression through a rate-distortion-optimal continuous channel and discuss finite codebook adaptation. Our matched-moment reference gives factor-specific stability conditions for finite hard partitions with jointly updated code vectors. Loss-weighted assignment exchanges connect this feedback to shared-rule risk.

Readout capacity shapes probing (Belinkov, 2022). DCI separates disentanglement, completeness and informativeness (Eastwood & Williams, 2018), and DCI-ES adds explicitness and size, measuring explicitness by required functional readout capacity (Eastwood et al., 2023). Probe controls assess memorization (Hewitt & Liang, 2019), usable information restricts predictor families (Xu et al., 2020), and minimum-description-length (MDL) probing measures label-transmission cost given representations (Voita & Titov, 2020). Building on this distinction between retained information and ease of use, we study shared-rule description length and worst-context risk (Sagawa et al., 2020). Exact search certifies organization-dependent costs in learned quantizers, while the dynamics analysis asks how these organizations form.

3 Method

3.1 Quantized representations and shared rules

Let S be a target factor and N the remaining conditions. A fixed map $C = b(N)$ defines finitely many positive-mass evaluation contexts $\mathcal{C}_{\text{eval}}$.

An encoder maps input X to $Z = f_\theta(X)$. At each layer, residual vector quantization (RVQ) selects token T_ℓ and vector q_ℓ from K_ℓ codes (Lee et al., 2022; Zeghidour et al., 2021):

$$r_1 = Z, \qquad T_\ell = \arg\min_{0 \le k < K_\ell} \|r_\ell - c_{\ell k}\|^2, \qquad q_\ell = c_{\ell, T_\ell},$$
$$r_{\ell+1} = r_\ell - \mathrm{sg}(q_\ell), \qquad Q = \sum_{\ell=1}^{L} q_\ell, \qquad A = (T_1, \ldots, T_L). \tag{1}$$

Here $c_{\ell k}$ is code vector k, r_ℓ the residual, L the layer count, and sg stops gradients. Decoder D_ϕ receives only Q through the straight-through estimator (STE), $Q_{\mathrm{ST}} = Z + \mathrm{sg}(Q - Z)$ (Bengio et al., 2013). Squared reconstruction, codebook and commitment losses have weights 1, $\alpha > 0$ and $\beta \ge 0$ (App. B.2). Synchronous updates use codebook rate $\eta_C > 0$, encoder rate $\eta_E = \kappa \eta_C$ and time $t = j\eta_C$ after j steps. Labels enter readout fitting and diagnostics only.

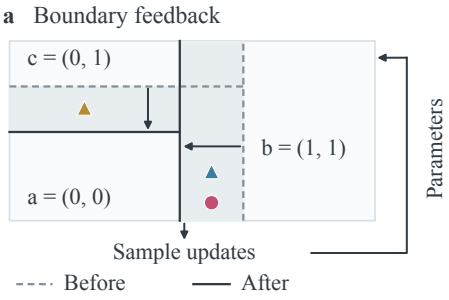
Let w collect model parameters, A_w their tuple assignment and $\mathcal{G}_B$ a finite rule family of length at most B bits, including constants. For loss $\ell \in [0, 1]$, zero-one in our experiments, define

$$R_n(g; w) = \mathbb{E}[\ell(g(A_w), S) \mid C = n], \qquad \mathcal{R}_S^{\mathrm{grp}}(B; w) = \min_{g \in \mathcal{G}_B} \max_{n \in \mathcal{C}_{\mathrm{eval}}} R_n(g; w). \tag{2}$$

A shared rule uses the same $g(A_w)$ across contexts without receiving C. A fixed rule also remains unchanged across states, while endpoint-selected rules are chosen separately at each endpoint. Categorical trees and tuple tables charge for operations, layers, category subsets and predictions, separately from token rate $\sum_\ell \log_2 K_\ell$. Layerwise token renumbering preserves costs. Small alphabets permit exact optima, while larger ones give attained risks from calibration-selected rules. Reliable recovery requires a frozen readout to meet a declared error threshold in every context, using exact risks or simultaneous held-out upper bounds. A separately fitted predictor (probe) or independently evaluated rule can establish recovery. At matched recovery, rule costs compare access to the retained target.

3.2 Boundary feedback and factor selection

Updates move assignment boundaries, and exchanged inputs alter subsequent updates (Fig. 2a).



b Fixed-rule errors

$g(T_1, T_2) = T_1$ in every context

Target	Tuple	Loss	Effect
▲ 1	a → b	1 → 0	−1 Corrected
● 0	a → b	0 → 1	+1 Introduced
▲ 1	a → c	1 → 1	0 Same output

Figure 2: Effects of an assignment exchange (schematic). (a) Moving boundaries alter tuples and sample updates. (b) A fixed rule can correct, introduce or preserve prediction errors.

Let $h(w)$ be the expected update per unit codebook time and $h_a(w, x)$ the sample update in cell $\mathcal{C}_a$. A perturbation u induces exchange velocity $\nu_{ab}[u]$, positive for switches from a to b. With density p, regular interfaces Σ_{ab} and zero-mass ties,

$$D_w h(w)[u] = \sum_a \int_{\mathcal{C}_a} D_w h_a(w, x)[u] p(x)\, dx + \sum_{a<b} \int_{\Sigma_{ab}} (h_b - h_a) p(x) \nu_{ab}[u](x)\, d\sigma(x). \tag{3}$$

Here D_w differentiates parameters and $d\sigma$ measures interface area. The terms describe within-cell changes and exchanges with all later RVQ assignments recomputed. Sample STE specifies h_a, while its population derivative also requires exchanges. This moving-region differentiation follows boundary-aware gradient principles (Lee et al., 2018), with assumptions and derivation in App. A.2.

Fix $Z = X = (S, N)^\top$, with independent $S \sim \mathrm{Unif}\{-1, 1\}$ and $N \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$. Target-aligned codes $\{\pm(1, 0)\}$ attain globally optimal two-code distortion one. Context-aligned codes

$\{\pm(0, \sqrt{3}/2)\}$ have distortion $5/4$ and minimum target error $1/2$, yet attract nearby population Lloyd updates (Lloyd, 1982). Each Lloyd update replaces a code by its assigned-input mean. Training can favor a higher-distortion organization that does not identify the target.

For a symmetric density f_N continuous near zero, a small target-coordinate tilt of the context-aligned codes has multiplier $\Lambda_N = 2f_N(0)\mathbb{E}[S^2]/\mathbb{E}|N|$, with $\mathbb{E}|N| > 0$. The tilt grows for $\Lambda_N > 1$ and decays for $\Lambda_N < 1$. A density family indexed by $-1 < \xi < 1$ fixes $\mathrm{Var}(N) = 1$, $\mathbb{E}|N| = \sqrt{3}/2$ and both branch distortions, yet yields $\Lambda_N = 2(1 + \xi)/3$. Thus matched moments allow different local dynamics.

Joint training couples encoder shear, which mixes the two coordinates, to code-vector tilt. Our reference uses a symmetric positive-definite linear encoder with determinant one, its inverse as decoder and two independent code vectors $c_\pm$, with $\alpha = 1$, $\beta = 1/4$. At the context-aligned equilibrium, eliminating shear from the two-dimensional linearization gives

$$\ddot{d} + [\mathsf{A}(\xi)\kappa - \mathsf{B}(\xi)]\dot{d} + c_0\kappa(1 - 2\xi)d = 0. \tag{4}$$

Here $d = (c_{+x} - c_{-x})/2$ is code-vector tilt and dots denote time derivatives at fixed $\kappa > 0$. Linearization yields $\mathsf{A}(\xi), \mathsf{B}(\xi), c_0$, with $\mathsf{A}(\xi), c_0 > 0$. The $\dot{d}$ coefficient controls damping and the d coefficient the restoring response. All other modes decay, giving strict full-state linear stability exactly when

$$\xi < \tfrac{1}{2}, \qquad \kappa > \max\{0, \mathsf{B}(\xi)/\mathsf{A}(\xi)\}. \tag{5}$$

Geometry fixes the restoring sign and also enters damping. At fixed geometry, rates can stabilize this context organization, which does not identify S. Positive blockwise rates preserve strict linear stability at a fixed critical point of smooth-energy gradient flow. The transition here depends on STE feedback. For the uniform source, $\kappa_H = \mathsf{B}(0)/\mathsf{A}(0) \approx 0.74864$. Just above it, the tilt's oscillation amplitude relaxes on scale $2/[\mathsf{A}(0)(\kappa - \kappa_H)]$. Independent decoders introduce neutral modes, and Euler steps impose additional stability conditions (App. B).

3.3 Assignment changes and shared-rule risk

The same exchange can correct or introduce an error, depending on the target (Fig. 2b). Replacing update jumps in Eq. (3) by loss jumps gives, for discrete S,

$$D_w R_n(g; w)[u] = \sum_{a<b,s} [\ell(g(b), s) - \ell(g(a), s)] \int_{\Sigma_{ab}} \rho_{s,n}(x)\nu_{ab}[u](x)\, d\sigma(x), \tag{6}$$

where $\rho_{s,n}(x)dx = \Pr(X \in dx, S = s \mid C = n)$. The same exchanges weight update jumps for training and loss jumps for rule risk. Prediction-preserving changes contribute zero, so tuple-change counts (token churn) alone cannot determine risk response. Finite sources admit exact endpoint loss sums.

A rule frozen before intervention tests whether an interpretation remains usable. Along a smooth untreated trajectory,

$$\delta\dot{w}_t = J_t\delta w_t + b_t, \qquad J_t = D_w h(w_t, \kappa_t), \qquad \delta R_n(g; w_T) = D_w R_n(g; w_T)\delta w_T. \tag{7}$$

Here δw_t is the first-order state change, T the endpoint time and b_t the forcing. The first equation propagates the intervention through subsequent training. The second converts the endpoint displacement into a first-order change in the fixed rule's context risk. Pulses change the field, while displacements set δw_0. Full response features include assignment exchanges, while Branch holds assignments fixed. Backward path modifies the training field, supplying the same forcing to both predictors. Features use the initial checkpoint and untreated continuation. Predictors are calibrated separately and evaluated on treated confirmation endpoints.

With distribution, contexts, loss, alphabet and rule family fixed, the old optimal rule's worst-context risk change bounds the optimal-risk change from above. Improving that rule therefore certifies improved optimal risk at the same budget (App. C.3).

3.4 Residual geometry and rule formation

Freeze the encoder and a first layer mapping N to $m\,\mathrm{sign}(N)$, with $m = \mathbb{E}|N| > 0$. Layer two receives (S, R), $R = N - m\,\mathrm{sign}(N)$. For independent symmetric factors, bounded S and f_N

continuous near $\pm m$,

$$f_R(0) = 2f_N(m), \qquad \Lambda_{\text{res}} = \frac{4f_N(m)\mathbb{E}[S^2]}{\mathbb{E}\big||N| - m\big|}, \qquad \dot{v} = \frac{\alpha}{2}(\Lambda_{\text{res}} - 1)v + o(v). \tag{8}$$

Here f_R is the residual density and $v = d/c_y$ the tilt of second-layer codes $c_\pm = \pm(d, c_y)$, expanded at $v = 0$, $c_y = \mathbb{E}|R| > 0$. Subtraction moves density near $\pm m$ to the boundary, allowing opposite growth at equal variance. The gradient rate includes cell probability $1/2$.

For independent $S \sim \text{Unif}\{-1, 1\}$ and $R \sim \text{Unif}[-m, m]$, $m = \sqrt{3}/2$, frozen encoder and first layer, $0 < v_0 < m$ and $0 < c_{y,0} \leq m/2$, the first time $v_t \geq m$ under gradient flow satisfies

$$T_{\text{hit}} \leq \frac{6}{5\alpha} \log \frac{m}{v_0}. \tag{9}$$

When v reaches m, the target contribution dominates the residual in the assignment score $R + vS$. Thus $\text{sign}(R + vS) = S$ almost surely, giving an error-free shared token test. The bound links residual geometry to shared-rule formation time. Earlier-layer motion changes this process through residual geometry and signed forcing (App. D).

4 Experiments

Controlled sources connect boundary feedback to learned rule costs and residual rule formation. Frozen Shapes3D representations and ECG morphology measurements test shared-rule reuse across contexts. Seed-bootstrap 95% confidence intervals (CIs) summarize training variation, preserving pairs and contexts. Recovery uses exact risks or context-adjusted held-out upper bounds, with ECG bounds computed across patients. Blue/rose marks favorable/adverse estimates.

4.1 Learned organizations at matched recovery

We train tanh encoders and decoders on $X = (S, N, \tau SN) + U$, with binary target S, four values of N, interaction strength τ and coordinate noise U. Codebooks of sizes four and two supply three-bit tuples. Crossing 32 seeds, three relative rates, two interaction strengths and two noise levels gives 384 runs. The noiseless eight-point source permits exact evaluation of all 256 binary functions.

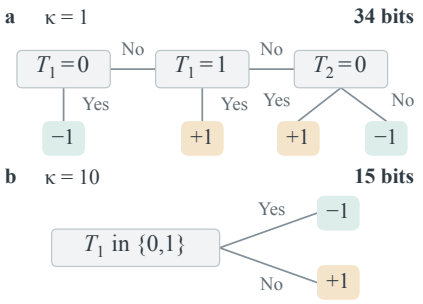


Figure 3: Minimum-length zero-error rules from a common initialization, using three-bit tuples. Branches show outcomes and leaves predict S.

At $\tau = 0.5$ without noise, two runs from a common initialization preserve the target perfectly but require different rules (Fig. 3). For $\kappa = 10$, one first-layer test on membership in $\{0, 1\}$ reads the target. For $\kappa = 1$, three conditional tests across both layers are required. They cost 15 and 34 bits, respectively. Exhaustive search certifies both minima: worst-context risk stays at 0.5 from the three-bit constant rule to each threshold, then drops to zero. Thus at $B = 16$, risks differ despite similar reconstruction mean squared errors (MSEs), 0.00887 and 0.00975.

Across all conditions, 375/384 runs pass the recovery check. All paired mean-risk CIs for raising κ from 1 to 10 include zero. The exact pair demonstrates how training can organize a preserved target at different rule costs. We next isolate the feedback governing factor selection.

4.2 Boundary feedback and factor selection

We test the selection mechanism by varying boundary geometry at fixed variance, absolute first moment and reference distortions. The Lloyd multiplier measures how strongly a tilt changes a cell's target-coordinate mean relative to the code-vector scale. It ranges from 0.0667 to 1.2667,

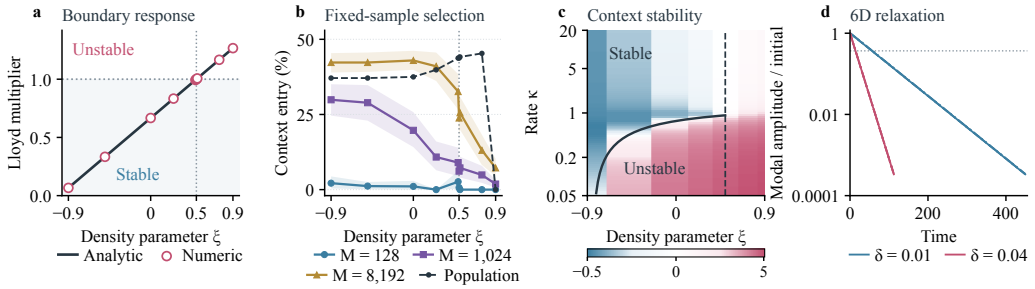


Figure 4: Boundary feedback and relaxation. (a) Lloyd gain. (b) Fixed-sample entry (64 seeds, 95% CIs). (c) Largest growth rate at the context-aligned equilibrium. (d) Normalized modal amplitude. M: sample size. $\delta = \kappa - \kappa_H$.

crossing the stability threshold of one (Fig. 4a). These matched summaries do not determine local attraction.

The six-dimensional spectrum over nine source parameters and 81 rates agrees with the stability condition away from numerically neutral boundaries. The target-aligned equilibrium remains locally stable throughout this grid, so either branch can attract nearby states where both are stable. Finite-sample Lloyd updates select between them according to initialization and sampling (Fig. 4b, 64 seeds, 16 starts each, 60 steps).

Near the uniform-source threshold, small perturbations of the six-dimensional population field relax on the predicted time scale (Fig. 4d). Relaxation takes approximately 57 time units at $\kappa - \kappa_H = 0.01$ and 14 at 0.04, showing how relative rates control the persistence of a local disturbance.

4.3 Training interventions and cross-context reuse

In the noisy task, each saved checkpoint, or parent, starts Control, Slow, Fast, Backward path (Backward) and Random direction (Random) runs. Control uses $\kappa = 1$, with Slow/Fast encoder-rate pulses to 0.1/10. A 12-bit second-layer token test stays fixed across arms. Each endpoint separately selects a rule at $B = 16$.

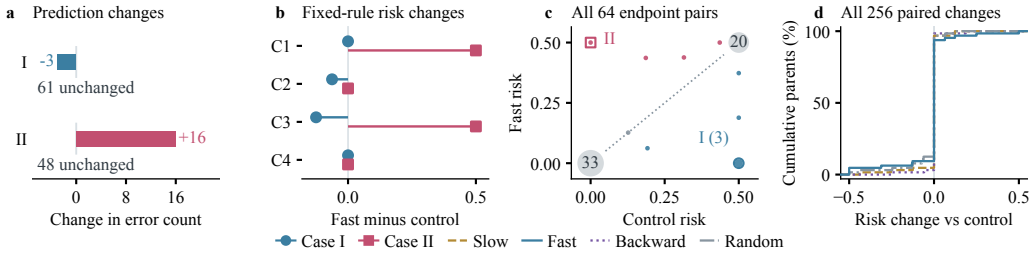


Figure 5: Fixed-rule exchanges and endpoint outcomes. (a,b) Error-count and context-risk changes in two illustrative cases (64 inputs each). (c) Endpoint risks (64 pairs, $B = 16$), with count-scaled bubbles. (d) Paired risk-change distributions.

The two cases in Fig. 5a,b show how exchanges change rule use at reliably readable endpoints. In case I, three tuple exchanges correct errors and eight preserve predictions. Fixed-rule worst-context risk stays at 0.5, while endpoint-selected risk falls from 0.5 to zero and attained zero-test-error length from 34 to 14 bits. For the fixed rule, corrections improve two contexts, while the two unchanged contexts still determine the maximum risk. The new organization supports a shorter rule with zero held-out error in every context. In case II, sixteen exchanges introduce errors, leaving 48 predictions unchanged. Fixed-rule risk rises from 0.5 to one, endpoint-selected risk from zero to

0.5, and attained zero-test-error length from 15 to 34 bits. Exchange counts use fixed inputs and endpoint risks use held-out data.

Across all 64 parents, Fast lowers endpoint risk in six, leaves it unchanged in 54 and raises it in four (Fig. 5c,d). Neither rate pulse establishes a mean reduction (Table 1). Random displacement lowers risk by 0.0226 $[0.0032, 0.0460]$ but fails the ± 0.02 probe-risk equivalence criterion. This variation motivates predicting which interventions improve a fixed rule's use.

Table 1: Worst-context endpoint risks ($B = 16$, 64 paired seeds). Δ: treatment minus Control, with 95% seed CIs. Pass: context-adjusted probe upper risk ≤ 0.05 in every context. Eq.: probe-risk equivalence within ± 0.02. MSE: mean squared reconstruction error per coordinate.

		Probe risk		Rule risk			
Intervention	Pass	Mean	Δ [95% CI]	Mean	Δ [95% CI]	MSE	Eq.
Control	49/64	0.1007	–	0.2149	–	0.2860	Ref.
Slow	49/64	0.1036	$+0.0029$ $[-0.0039, 0.0106]$	0.2059	-0.0090 $[-0.0277, 0.0056]$	0.2888	Yes
Fast	51/64	0.0996	-0.0011 $[-0.0150, 0.0117]$	0.1973	-0.0177 $[-0.0536, 0.0146]$	0.2682	Yes
Backward	50/64	0.0967	-0.0040 $[-0.0100, 0.0000]$	0.2131	-0.0018 $[-0.0124, 0.0078]$	0.2815	Yes
Random	48/64	0.1075	$+0.0068$ $[-0.0039, 0.0213]$	0.1924	-0.0226 $[-0.0460, -0.0032]$	0.2863	No

We evaluate fixed-rule forecasts on the controlled task and Shapes3D (Burgess & Kim, 2018). For Shapes3D, we train linear encoders, residual codebooks and decoders on frozen 16-dimensional image representations. Shape is the four-class target, with hue–orientation contexts. Factor labels do not enter quantization-model losses. Probes use linear readouts or fitted output layers on fixed hidden features.

Forecasts use untreated continuations and intervention definitions, with disjoint seeds for fitting, tuning, interval calibration and confirmation. Four interventions are evaluated across four contexts per controlled-task seed and 138 contexts at two token rates per Shapes3D seed. Full and Branch use the respective response features from Sec. 3.3. Shapes3D Full additionally uses boundary mass and churn, giving a comparison of combined feature sets (Table 2). Ranks compare intervention–context responses in the small model and context-averaged interventions at each Shapes3D token rate.

Table 2: Fixed-rule forecasts. MAE: mean absolute error (10^{-3}). Rank: Spearman correlation (n: defined seeds). Sign err.: decrease/tie/increase error. Differences use shared valid seeds. Brackets: paired 95% seed CIs.

	Small model (64 seeds)			Shapes3D (32 seeds)		
Predictor	MAE	Rank (n)	Sign err.	MAE	Rank (n)	Sign err.
Full	25.65	0.591 (46)	0.880	4.39	0.598 (32)	0.677
Branch	24.90	0.640 (39)	0.879	4.51	0.280 (32)	0.687
Churn	28.84	0.106 (46)	0.915	7.32	0.007 (32)	0.724
Zero	27.59	– (0)	0.149	5.49	– (0)	0.522

Full − comparator		Small model		Shapes3D	
Comparator	Metric	n	Difference [95% CI]	n	Difference [95% CI]
Branch	MAE	64	0.7461 $[-2.4980, 4.4703]$	32	-0.1209 $[-0.1455, -0.0959]$
	Rank	39	-0.0403 $[-0.1871, 0.1075]$	32	0.3185 $[0.2268, 0.4036]$
	Sign err.	64	0.0010 $[-0.0166, 0.0205]$	32	-0.0104 $[-0.0152, -0.0057]$

On the controlled task, Full improves ranking over Churn by 0.4845 $[0.2954, 0.6709]$ across 46 seeds with defined correlations for both predictors. Paired CIs for its mean absolute error (MAE) and ranking differences from Branch include zero. On frozen Shapes3D representations, the combined predictor lowers MAE by 2.7% and raises ranking correlation from 0.280 to 0.598.

Ranking gains coexist with limited direction accuracy. Full predicts case I's nonzero context-level improvements but misses case II's increases (Table S3). On Shapes3D, its decrease/tie/increase error is 67.7%, versus 52.2% for always predicting no change. Prediction intervals and calibration details appear in App. E.4.

For Shapes3D endpoint rules at $B = 256$, six-bit Fast lowers ordinary-test risk by 0.0126 $[0.0029, 0.0267]$, while the twelve-bit CI includes zero. Across 138 contexts, no Control, Slow or Fast seed meets the 0.05 probe-recovery criterion.

Frozen selected rules also transfer to two six-context holdouts, excluded from readout fitting or from both quantization-model and readout fitting. At twelve token bits, mean worst-context rule risks are 0.034 and 0.049 (Table 3). The fixed rules meet the record-level upper-risk criterion in 20/32 and 7/32 seeds, compared with 1/32 and 0/32 for probes. Each passing rule supplies both a reliable readout and a concise interpretation on its held-out contexts. These results evaluate reuse over the six held-out contexts in each split, separately from the 138 ordinary-test contexts.

Table 3: Shapes3D holdouts ($B = 256$, 32 Control seeds). Entries: worst-context risk [95% seed CI]. P|R: passing probe|rule seeds under a nominal record-level upper-risk threshold of 0.05 in every evaluated context.

	6-bit tokens			12-bit tokens		
Split (contexts)	Probe [95% CI]	Rule [95% CI]	P|R	Probe [95% CI]	Rule [95% CI]	P|R
Ordinary (138)	$0.439[0.408, 0.470]$	$0.575[0.540, 0.607]$	0 | 0	$0.175[0.159, 0.192]$	$0.225[0.207, 0.245]$	0 | 0
Readout (6)	$0.212[0.178, 0.247]$	$0.321[0.274, 0.368]$	1 | 0	$0.236[0.202, 0.271]$	$0.034[0.021, 0.050]$	1 | 20
Joint (6)	$0.214[0.178, 0.256]$	$0.336[0.293, 0.381]$	0 | 0	$0.170[0.141, 0.200]$	$0.049[0.037, 0.062]$	0 | 7

Structural comparisons and rule-budget curves cover 14 conditions (App. E.5). The continuous autoencoder (AE) passes the ordinary-test probe check in all 32 seeds. Its selected rule reaches risk 0.0078 at $B = 4{,}096$. At $B = 256$, coefficient or normalization/grid costs admit only constants, illustrating how the declared rule language affects access costs across interfaces.

4.4 Residual Geometry and Rule Formation

Residual geometry determines whether a target-direction tilt grows after subtracting the preceding layer's code vectors. Uniform and gapped residuals share variance $1/4$ and preceding-layer distortion, yet their gradient growth rates are $5\alpha/6$ and $-\alpha/2$ (Fig. 6a,b, $\alpha = 1$). Their different boundary densities and absolute first moments produce growth in one and decay in the other. A nine-source sweep varies positive growth from 0.3514 to 1.3528.

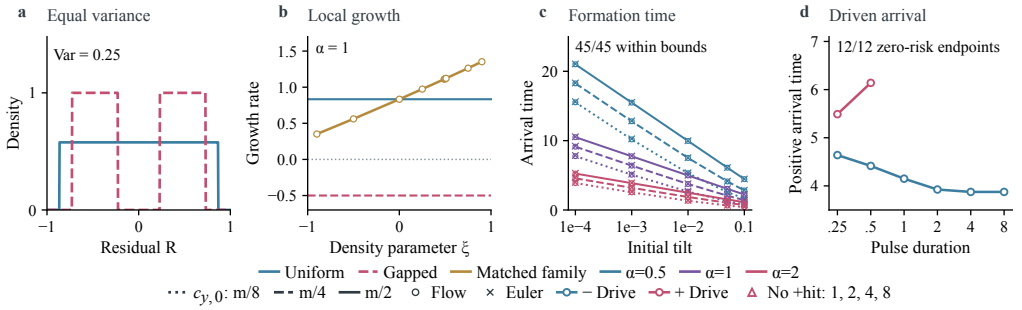


Figure 6: Residual geometry and formation. (a) Equal-variance densities. (b) Local growth. (c) Flow/Euler arrivals. (d) Signed-drive arrival. No-hit durations are listed in the legend.

All 45 deterministic configurations satisfying the reference assumptions obey both arrival bounds (Fig. 6c). For $v_0 = 10^{-3}$, $c_{y,0} = m/2$ and $\alpha = 1$, the target partition forms at time 7.741, below the bound of 8.117. One second-layer token test then reads the target without error in every context. Twelve zero-tilt controls never form that partition.

Preceding-layer motion tests signed forcing beyond the frozen reference. Despite matched initial target components and target-coordinate Jacobian entries, opposite paths yield arrival times 4.640 and 5.490 at duration 0.25 (Fig. 6d). Across six durations per direction, four opposing paths reverse token polarity and never hit the positive threshold, yet all 12 endpoints admit an error-free 12-bit rule. Here polarity specifies which token predicts each target class. At the reversed endpoints, swapping the two token predictions gives the error-free shared rule.

Joint training tests six pulse durations and two rates across 64 parents, ending at relative time 10. Endpoint rule risk is unchanged in 717/768 pairs, and all twelve mean-effect CIs include zero. Fixed readouts pass in 32/64 parents and 5–7 endpoints per condition, with intermediate failures on some paths. Longer pulses leave less recovery time before the shared endpoint (App. E.6).

4.5 Shared rules on ECG measurements

ECG measurements test reuse across patients, waveform phase and datasets. The QRS complex reflects ventricular depolarization. The target indicates whether QRS duration exceeds a threshold fixed by the calibration median, with contexts defined by its position within the window. An affine encoder and two learned code vectors quantize measured width. Fixed phase, amplitude and baseline quantizers complete a seven-bit tuple. The calibration-selected rule stays fixed across contexts and interventions. Risk weights patients equally after averaging their windows. Fig. 7a,b,d uses $B = 256$, while panel c compares encoder output z, code-vector output q and tuples.

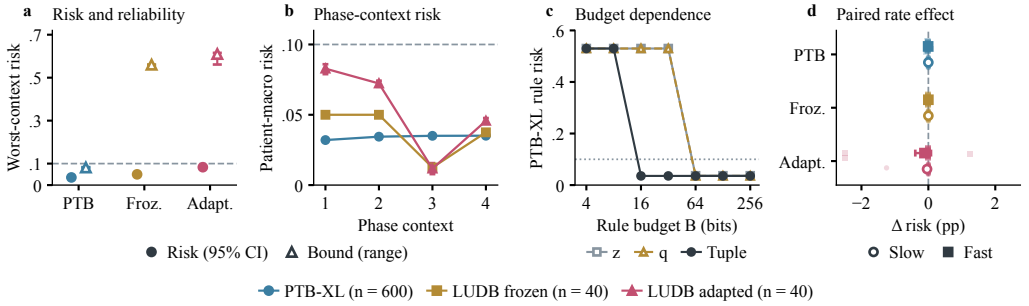


Figure 7: ECG reuse (32 seeds). (a) Patient-macro risk and upper bounds. (b) Context risks. (c) Tested rule budgets. (d) Paired rate effects. Bars: 95% seed CIs. Triangles in (a): patient-bound ranges.

On 600 PTB-XL test patients, the 13-bit rule attains risk 0.0355 and passes the 0.10 upper-risk criterion in all 32 seeds. It provides reliable reuse across patients and phase contexts, at a lower attained cost than the 42-bit continuous-coordinate tests under the same grammar. Frozen transfer to 40 LUDB patients attains risk 0.0500, with patient upper bound 0.5602, above the 0.10 reliability threshold. Refitting also fails this criterion, and all six paired rate-effect CIs include zero (App. E.7).

5 Conclusion and Limitations

Quantization training shapes both factor organization and the shared rules it supports. Boundary feedback determines local factor stability in our reference, and the same assignment exchanges affect training updates and rule errors. Learned quantizers exhibit different minimum rule costs despite perfect target recovery, while paired interventions show how short-rule access changes between reliably readable endpoints. The frozen residual reference connects this mechanism to a functional endpoint by bounding the time until an error-free shared rule becomes available. These results connect how factor organizations form during training to the cost and timing of their cross-context use.

The guarantees concern the specified reference updates and frozen residual inputs. Extending formation bounds to joint training requires control of evolving residual distributions and target recovery along the trajectory. Rule costs are relative to the declared grammar, with exact minima from exhaustive search and attained costs from candidate search. The larger-scale evidence concerns frozen Shapes3D representations evaluated at record level and ECG morphology measurements, with reliable reuse established within PTB-XL.

AI USE STATEMENT

We used generative AI tools to polish the writing and assist with literature retrieval and discovery. The authors reviewed the AI-assisted text and checked the cited sources. We take responsibility for the final content of this work, including its text, claims, and results.

ETHICS STATEMENT

We do not identify ethical concerns specific to this study. The experiments use controlled sources and existing datasets. The ECG analysis evaluates representation and rule behavior on PTB-XL and LUDB. Its duration threshold is defined from calibration data for this task and is not a clinical diagnostic criterion.

REPRODUCIBILITY STATEMENT

Sec. 3.1 specifies the quantizer, training objective and rule-risk definitions. Assumptions and derivations for the theoretical results appear in App. A–D. App. E provides model configurations, data splits, rule-selection procedures, intervention protocols and uncertainty estimates. Code is included in the supplementary materials.

REFERENCES

Mohit Bansil and Jun Kitagawa. Quantitative stability in the geometry of semi-discrete optimal transport. *International Mathematics Research Notices*, 2022(10):7354–7389, 2022.

Vitória Barin-Pacela, Kartik Ahuja, Simon Lacoste-Julien, and Pascal Vincent. On the identifiability of quantized factors. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 384–422, 2024.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101, 2025.

Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999. doi: 10.1137/s0036144599352836.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Cian Eastwood, Andrei Liviu Nicolicioiu, Julius Von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, and Bernhard Schölkopf. DCI-ES: An extended disentanglement framework with connections to identifiability. In *International Conference on Learning Representations*, 2023.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552.

Christopher Fifty, Ronald Junkins, Dennis Duan, Aniketh Iyengar, Jerry Liu, Ehsan Amid, Sebastian Thrun, and Christopher Re. Restructuring vector quantization with the rotation trick. In *International Conference on Learning Representations*, volume 2025, pp. 19153–19188, 2025.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173, 2017.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45463–45488, 2023. doi: 10.52202/075280-1969.

Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14096–14113, 2023.

Alena Kalyakulina, Igor Yusipov, Viktor Moskalenko, Alexander Nikolskiy, Konstantin Kosonogov, Nikolai Zolotykh, and Mikhail Ivanchenko. Lobachevsky University Electrocardiography Database. *PhysioNet*, 2021. doi: 10.13026/eegm-h675. URL https://doi.org/10.13026/eegm-h675. Version 1.0.1.

Alena I. Kalyakulina, Igor I. Yusipov, Viktor A. Moskalenko, Alexander V. Nikolskiy, Konstantin A. Kosonogov, Grigory V. Osipov, Nikolai Yu. Zolotykh, and Mikhail V. Ivanchenko. Ludb: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8: 186181–186190, 2020. doi: 10.1109/access.2020.3029211.

Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019. doi: 10.4171/jems/889.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11523–11532, 2022.

Wonyeol Lee, Hangyeol Yu, and Hongseok Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/tit.1982.1056489.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Hao Lu, Yongxin Guo, Onur Koyun, Zhengjie Zhu, Abbas Alili, and Metin N. Gurcan. Nsvq: Mitigating codebook collapse by stabilizing encoder drift in vector quantization, 2026. URL https://arxiv.org/abs/2606.11363.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, 2015.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. Physionet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Alex Tamkin, Mohammad Taufeeque, and Noah Goodman. Codebook features: Sparse and discrete interpretability for neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47535–47563, 2024.

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

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, 2022. doi: 10.13026/kfzx-aw45. URL https://doi.org/10.13026/kfzx-aw45. Version 1.0.3.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley J. Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507, 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994.

Xinyu Zhao, Nikita Karagodin, Hamed Hassani, Sinan Hersek, Paul Pu Liang, and Yury Polyanskiy. Continuous first, discrete later: Vq-vaes without dimensional collapse, 2026. URL https://arxiv.org/abs/2605.06870.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22968–22977, 2025.

A Shared rules and assignment boundaries

A.1 Shared rules and token relabeling

For the tuple alphabet and grammar in Eq. (2), let π be a bijective relabeling. If $g \mapsto g \circ \pi^{-1}$ and its inverse preserve the grammar and description length, then $(g \circ \pi^{-1})(\pi(A_w)) = g(A_w)$. This pairs rules with identical predictions and lengths, preserving every context risk and the optimal risk–length curve.

This applies to separate layerwise token permutations $\pi(a_1, \dots, a_L) = (\pi_1(a_1), \dots, \pi_L(a_L))$ when token tests retain their costs. General tuple permutations can change the subsets expressible by short layerwise tests and hence their minimum costs. A fixed rule with unchanged token literals can also have a different risk after relabeling.

A.2 Differentiating the training field

The tuple assignment cells are $\mathcal{C}_a(w) = \{x : A_w(x) = a\}$, with interfaces $\Sigma_{ab}(w)$. Consider a fixed input density p in dimension d_x. Near the reference state, assume the finite collection of active cell interfaces is piecewise continuously differentiable in (w, x). Each regular interface has a defining function ψ_{ab} with $\nabla_x \psi_{ab} \neq 0$, and ties have probability zero. Intersections of distinct active faces have codimension at least two and contribute $o(\epsilon)$ mass under a perturbation of size $O(\epsilon)$. The density has common continuous traces from both sides of each face. The integrands and their derivatives admit integrable bounds that justify differentiation under the integral (Mohamed et al., 2020). Compact support with bounded, regular interfaces is a sufficient local setting.

Choose $\psi_{ab} < 0$ on cell a and $\psi_{ab} > 0$ on cell b. The unit normal from a to b is $\nabla_x \psi_{ab} / \|\nabla_x \psi_{ab}\|$. Let x_ϵ track a point on the moving face while input samples remain fixed. Under state perturbation u, differentiating $\psi_{ab}(w + \epsilon u, x_\epsilon) = 0$ gives the geometric face velocity $V_{ab}[u]$ along this normal and the opposite-signed exchange velocity $\nu_{ab}[u]$:

$$V_{ab}[u] = -\frac{D_w \psi_{ab}[u]}{\|\nabla_x \psi_{ab}\|}, \qquad \nu_{ab}[u] = -V_{ab}[u]. \tag{10}$$

Positive ν_{ab} means that cell a shrinks and fixed inputs switch from a to b. The index order $a < b$ fixes an orientation and counts each face once. On a fixed smooth support of dimension k, gradients and surface measures are taken within that support. Here $d\sigma$ is its $(k-1)$-dimensional Hausdorff measure. This includes the two lines $S = \pm 1$ in the two-factor reference. Conditional distributions used to evaluate contexts must satisfy these assumptions on their own supports.

For fixed-assignment sample updates h_a, write the population update as $h(w) = \sum_a \int_{\mathcal{C}_a(w)} h_a(w, x) p(x)\, dx$. Differentiating each cell integral gives an interior contribution from $D_w h_a$ and a moving-boundary contribution. On a shared face, the two cells contribute $h_a p V_{ab}$ and $-h_b p V_{ab}$, summing to $(h_b - h_a) p \nu_{ab}$. Counting each face once gives Eq. (3), with the intersection and domination assumptions controlling the remainder. Interior and boundary contributions also arise in gradients of non-differentiable models (Lee et al., 2018). Related cell geometry is studied in semidiscrete optimal transport (Kitagawa et al., 2019; Bansil & Kitagawa, 2022).

For one layer, take $\psi_{ab}(w, x) = \|f_\theta(x) - c_a\|^2 - \|f_\theta(x) - c_b\|^2$. In RVQ, each regular face comes from a distance tie during sequential assignment. Both labels in $h_b - h_a$ are complete tuples, with later-layer assignments recomputed on each side. Sample STE backpropagation supplies the fixed-assignment integrand h_a. The population derivative additionally accounts for these exchanges.

B BOUNDARY FEEDBACK AND FACTOR SELECTION

B.1 MATCHED MOMENTS AND LLOYD DYNAMICS

Set $X = (S, N)^T$ with independent symmetric factors. Assume the target factor S is bounded, N is integrable, and its density f_N is continuous near zero. The context-aligned organization partitions inputs by the sign of N.

For codes $c_\pm = \pm(d, c_y)$ with $c_y > 0$, put $v = d/c_y$. The positive cell is $N + vS > 0$ and has probability $1/2$. Symmetry gives its conditional mean as $\mu(v) = \mathbb{E}[(S, N)^T \operatorname{sign}(N + vS)]$. Conditioning on S gives $\mu_S'(0) = 2f_N(0)\mathbb{E}[S^2]$, $\mu_S(0) = 0$, and $\mu_N(0) = \mathbb{E}|N|$. Each Lloyd step replaces the codes by these conditional means (Du et al., 1999), so the new tilt is $\mu_S(v)/\mu_N(v)$. For $\mathbb{E}|N| > 0$, its local multiplier is

$$\Lambda_N = \frac{d}{dv}\frac{\mu_S(v)}{\mu_N(v)}\bigg|_{v=0} = \frac{2f_N(0)\mathbb{E}[S^2]}{\mathbb{E}|N|}. \tag{11}$$

Thus a small tilt grows when $\Lambda_N > 1$ and shrinks when $\Lambda_N < 1$. Boundedness of S can be replaced by an integrable bound that justifies differentiation of the conditional integral.

For an equiprobable binary target $S \in \{-1, 1\}$, choose the density of N as

$$\begin{aligned} f_\xi(n) &= \frac{1}{2\sqrt{3}}\left[1 + \xi P_4\left(\frac{|n|}{\sqrt{3}}\right)\right]\mathbf{1}_{\{|n|\le\sqrt{3}\}}, \qquad -1 < \xi < 1, \\ P_4(u) &= 70u^4 - 140u^3 + 90u^2 - 20u + 1. \end{aligned} \tag{12}$$

The shifted Legendre polynomial satisfies $|P_4(u)| \le 1$ on $[0, 1]$ and $\int_0^1 u^j P_4(u)\,du = 0$ for $j = 0, 1, 2$. The density is therefore positive inside its support and preserves $\mathbb{E}[N^2] = 1$ and $m = \mathbb{E}|N| = \sqrt{3}/2$. Yet $f_\xi(0) = (1 + \xi)/(2\sqrt{3})$ gives $\Lambda_N(\xi) = 2(1 + \xi)/3$. The target- and context-aligned distortions below also remain fixed across the family, separating these distortion summaries from the local direction response.

At the uniform member $\xi = 0$, both factors have unit variance. For any binary assignment T, total covariance gives

$$\mathbb{E}\|X - \mathbb{E}[X \mid T]\|^2 = 2 - \operatorname{tr}\operatorname{Cov}(\mathbb{E}[X \mid T]). \tag{13}$$

The covariance on the right is positive semidefinite, has rank at most one, and is bounded by the identity matrix I_2. Every free two-code quantizer thus has distortion at least one. The target codebook $\{\pm(1, 0)\}$ attains this bound, whereas $\{\pm(0, m)\}$ has distortion $2 - m^2 = 5/4$. Both values depend only on the matched moments.

At the uniform context-aligned codebook, independent perturbations of both code vectors give horizontal and vertical blocks of the Lloyd update-map Jacobian:

$$J_x = \begin{pmatrix} 1/3 & -1/3 \\ -1/3 & 1/3 \end{pmatrix}, \qquad J_y = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{pmatrix}. \tag{14}$$

The horizontal difference tilts the threshold, while a common vertical displacement translates it. The eigenvalues $\{0, 0, 1/2, 2/3\}$ all lie inside the unit circle, proving local attraction in all four coordinates.

B.2 THE STE FIELD AND ITS CONTEXT-ALIGNED EQUILIBRIUM

For the general RVQ interface in Eq. (1), we adapt the reconstruction, codebook and commitment objective of van den Oord et al. (2017) to residual codebooks:

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{2}\|X - D_\phi(Q_{\mathrm{ST}})\|^2 + \frac{\alpha}{2}\sum_\ell \|\operatorname{sg}(r_\ell) - q_\ell\|^2 + \frac{\beta}{2}\sum_\ell \|r_\ell - \operatorname{sg}(q_\ell)\|^2\right]. \tag{15}$$

The reconstruction term uses one overall STE; codebook mismatch and commitment use the indicated stop-gradient paths. Relative to codebook rate $\eta_C > 0$, synchronous encoder and independent-decoder rates are $\eta_E = \kappa\eta_C$ and $\eta_D = \kappa_D\eta_C$.

The reference in Sec. 3.2 sets $W = \exp H$ with $H = \left(\begin{smallmatrix} a & b \\ b & -a \end{smallmatrix}\right)$ and decoder W^{-1}. Its six-dimensional state is $w = (a, b, c_{-x}, c_{-y}, c_{+x}, c_{+y})$, with independent code vectors. The nearest-code index T defines $Q_{\mathrm{ST}} = WX + \mathrm{sg}(c_T - WX)$ and reconstruction $\widehat{X} = W^{-1}Q_{\mathrm{ST}}$. Eq. (15) uses one layer, $\alpha = 1$ and $\beta = 1/4$. Synchronous updates use codebook time and relative encoder rate $\kappa > 0$. The constraint $a^2 + b^2 \le \gamma_{\max}^2$, with $\gamma_{\max} > \log 2$, includes both reference equilibria and is inactive locally.

Put $e = W^{-1}c_T - X$ and $\theta = (a, b)$, and write $\partial_j W = \partial W/\partial\theta_j$. The reconstruction and commitment gradients, g^{rec} and g^{com}, and the codebook gradient g_{c_k} are

$$g_j^{\mathrm{rec}} = -\mathbb{E}[e^T W^{-1}(\partial_j W)e], \qquad g_j^{\mathrm{com}} = \beta\mathbb{E}[(WX - c_T)^T(\partial_j W)X],$$
$$g_{c_k} = \mathbb{E}[(c_k - WX)\mathbf{1}_{\{T=k\}}], \qquad \dot\theta = -\kappa(g^{\mathrm{rec}} + g^{\mathrm{com}}), \qquad \dot c_k = -g_{c_k}. \tag{16}$$

For the reconstruction term, the STE gives $\partial_j Q_{\mathrm{ST}} = (\partial_j W)X$, whereas differentiating the tied inverse contributes $-W^{-1}(\partial_j W)W^{-1}c_T$. Their sum is $-W^{-1}(\partial_j W)e$, which proves the first expression. The two stop-gradient losses give the remaining expressions directly.

For every density in Eq. (12), the context-aligned equilibrium is

$$a_N = \frac{1}{2}\log\frac{3+\sqrt{10}}{2}, \qquad b = 0, \qquad c_\pm = \pm(0, e^{-a_N}m), \qquad m = \frac{\sqrt{3}}{2}. \tag{17}$$

At this state the assignment is the sign of N. Write $q = e^{a_N}$, $r = q^{-1}$, and $V = 1 - m^2 = 1/4$. The codebook gradient vanishes because the codes are conditional encoded means. The encoder b gradient vanishes by symmetry. The encoder a gradient is $-1 + V + \beta(q^2 - r^2V)$, whose vanishing is equivalent to $4q^4 - 12q^2 - 1 = 0$. This proves that the location depends on the matched moments rather than ξ. Since S is independent of this assignment, its minimum target error is $1/2$.

The target equilibrium is $a_S = \frac{1}{2}\log\beta = -\log 2$, $b = 0$, $c_\pm = \pm(e^{a_S}, 0)$. Its numerical spectrum is reported in App. E.3.

B.3 Full-state linear stability

Full-state stability requires every perturbation mode to decay. Use the coordinates

$$d = \frac{c_{+x} - c_{-x}}{2}, \qquad c_y = \frac{c_{+y} - c_{-y}}{2}, \qquad u_x = \frac{c_{+x} + c_{-x}}{2}, \qquad u_y = \frac{c_{+y} + c_{-y}}{2}. \tag{18}$$

At the context-aligned equilibrium, the linearization splits into three blocks: direction (b, d), scale (a, c_y), and common translation (u_x, u_y). The direction block couples encoder shear b to the target-direction code component d. The following coefficients abbreviate the matrix entries, with $a = a_N$:

$$\chi = \frac{\sinh a}{a}, \qquad A_0 = -\frac{3(10q^2 - 1)}{64a^2q^2}, \qquad t_\xi = \frac{1+\xi}{\sqrt{3}},$$
$$B_\xi = \chi m t_\xi\big(q + (1-\beta)r\big). \tag{19}$$

Differentiation of Eq. (16), including its moving conditional thresholds, gives

$$J_{\mathrm{ang}} = \begin{pmatrix} \kappa(A_0 - B_\xi q\chi) & \kappa(\beta\chi m - B_\xi q^3/m) \\ \frac{1}{2}\chi(m + q^2 t_\xi) & \frac{1}{2}(q^4 t_\xi/m - 1) \end{pmatrix}, \tag{20}$$

and

$$J_{\mathrm{rad}} = \begin{pmatrix} -\kappa\beta[2q^2 + (2 - m^2)r^2] & -\kappa\beta rm \\ -rm/2 & -1/2 \end{pmatrix}, \qquad J_{\mathrm{trans}} = \mathrm{diag}\left(-\frac{1}{2}, -\frac{1-\xi}{4}\right). \tag{21}$$

The terms $t_\xi = 2f_\xi(0)$ enter through the boundary exchange. On the symmetric subspace, write $c_\pm = \pm c$ with $c = (d, c_y)^T$. The input-space threshold is $N + vS = 0$, with $v = (Wc)_1/(Wc)_2$. When $W = I$, as in the frozen residual reference, this reduces to $v = d/c_y$. At equilibrium its differential is $dv = q\chi\, db + (q^3/m)\, dd$, and the conditional horizontal mean has derivative $t_\xi\, dv$. Substituting this derivative into the encoder and codebook fields gives Eq. (20). Differentiating the moment terms with the context-aligned assignment fixed gives the radial block. Translating both

codes leaves the horizontal common mean with decay rate $1/2$, while the vertical threshold response changes its rate to $(1-\xi)/4$.

Using $4q^4 - 12q^2 - 1 = 0$, define the trace and determinant coefficients by

$$\mathsf{A}(\xi) = -(A_0 - B_\xi q\chi), \qquad \mathsf{B}(\xi) = \frac{1}{2}(q^4 t_\xi/m - 1), \qquad c_0 = \frac{9(4q^2-1)}{512 a_N^2 q^2}. \tag{22}$$

The direction block has trace and determinant

$$\operatorname{tr} J_{\mathrm{ang}} = -\mathsf{A}(\xi)\kappa + \mathsf{B}(\xi), \qquad \det J_{\mathrm{ang}} = c_0\kappa(1-2\xi). \tag{23}$$

Since $A_0 < 0$ and $B_\xi > 0$ for $-1 < \xi < 1$, both $\mathsf{A}(\xi)$ and c_0 are positive. Numerically, $\mathsf{A}(\xi) \simeq 3.55907 + 2.12644\xi$, $\mathsf{B}(\xi) \simeq 2.66447 + 3.16447\xi$, and $c_0 \simeq 0.20408$. The scale block has negative trace and positive determinant $\kappa\beta(q^2 + Vr^2)$. Both translations also have negative eigenvalues. The direction block is therefore decisive: its eigenvalues have negative real parts exactly when its trace is negative and its determinant is positive. This two-dimensional Hurwitz criterion proves the full-state condition in Eq. (5). At equality, linear stability does not decide the nonlinear behavior.

For fixed ξ and κ, the characteristic identity $J_{\mathrm{ang}}^2 - (\operatorname{tr} J_{\mathrm{ang}})J_{\mathrm{ang}} + (\det J_{\mathrm{ang}})I = 0$ gives Eq. (4) for each direction coordinate. Its coefficients describe damping and restoring feedback in this linearization.

B.4 Rate scaling and local relaxation

To separate STE feedback from positive rate scaling, consider the gradient flow $\dot{w} = -P\nabla E(w)$. Here E is twice continuously differentiable and P is a symmetric positive-definite block-rate matrix. At a fixed critical point w^*, let $H_E = \nabla^2 E(w^*)$. Its Jacobian satisfies

$$P^{-1/2} J P^{1/2} = -P^{1/2} H_E P^{1/2}, \qquad J = -P H_E. \tag{24}$$

The right-hand side is symmetric and congruent to $-H_E$, so positive block-rate changes preserve its inertia. They cannot change strict linear stability at this fixed critical point of the same energy. Finite-step stability also depends on the update map.

For the uniform source, $\xi = 0$, define

$$\kappa_H = \frac{\mathsf{B}(0)}{\mathsf{A}(0)} \simeq 0.74864186, \qquad \omega_H = \sqrt{c_0\kappa_H} \simeq 0.39087710. \tag{25}$$

At κ_H the angular eigenvalues are $\pm i\omega_H$, and the other four eigenvalues have negative real parts. Write $\delta = \kappa - \kappa_H > 0$. While $\mathsf{A}(0)^2\delta^2 < 4c_0(\kappa_H + \delta)$, the angular eigenvalues remain a complex-conjugate pair with real part $-\mathsf{A}(0)\delta/2$. The linear modal amplitude therefore decays as $\exp[-\mathsf{A}(0)\delta t/2]$, giving the e-fold relaxation time

$$\tau_{\mathrm{lin}} = \frac{2}{\mathsf{A}(0)\delta}. \tag{26}$$

This diverging timescale describes decay of the local oscillation envelope. App. E.3 compares it with the actual population field and reports numerical periodic orbits for the uniform source.

B.5 Independent decoder and discrete updates

For an independent decoder, replace W^{-1} by a freely trained 2×2 matrix D and put $e_D = Dc_T - X$. The overall STE gives

$$g_j^{\mathrm{rec}} = \mathbb{E}[e_D^T D(\partial_j W)X], \qquad \dot{D} = -\kappa_D \mathbb{E}[e_D c_T^T]. \tag{27}$$

The commitment and codebook fields are unchanged. The resulting ten-dimensional spectrum, including two numerically identified neutral modes, is reported in App. E.3.

Suppose the Jacobian J is Hurwitz, meaning that all its eigenvalues have negative real parts. Synchronous Euler updates require $|1 + \eta_C\lambda| < 1$ for every eigenvalue $\lambda \in \operatorname{spec}(J)$. Expanding the squared modulus gives

$$0 < \eta_C < \min_{\lambda\in\operatorname{spec}(J)} \frac{-2\operatorname{Re}\lambda}{|\lambda|^2}. \tag{28}$$

For the independent-decoder spectrum, this bound governs the nonneutral modes, while neutral modes retain Euler multiplier one.

For the tied-decoder uniform reference in the complex-conjugate angular regime, $|1 + \eta_C\lambda|^2 = 1 + \eta_C \operatorname{tr} J_{\mathrm{ang}} + \eta_C^2 \det J_{\mathrm{ang}}$. Substitution gives the angular stability threshold

$$\kappa > \kappa_H^{(\eta_C)} = \frac{\mathsf{B}(0)}{\mathsf{A}(0) - \eta_C c_0}, \tag{29}$$

provided the denominator is positive, the pair remains complex and the other discrete modes are stable. Across five step sizes from 10^{-4} to 10^{-2}, the threshold shifts from 0.748646 to 0.749071. Five rate offsets per step size give 25 agreeing spectral checks.

Rate changes govern the growth or decay of perturbations around an equilibrium common to those rates. Without a perturbation, that equilibrium remains a deterministic fixed point.

C Assignment changes and shared-rule risk

C.1 Loss-weighted exchanges and endpoint risk changes

For the discrete target, fixed rule and conditional density in Eq. (6), expand context risk over cells with integrand $\ell(g(a), s)\rho_{s,n}(x)$, summed over s. Assume each density satisfies the trace regularity in App. A.2. The integrand has no state dependence within a cell, so the transport argument gives Eq. (6) with loss differences as boundary weights. Exchanges between tuples with equal rule outputs contribute zero. Under zero-one loss, corrected errors have weight -1 and introduced errors have weight $+1$. For continuous-coordinate readouts, risk changes can also arise within cells or at readout thresholds.

For any two states w_0, w_1 evaluated on the same source, the endpoint risk change has the exact decomposition

$$R_n(g; w_1) - R_n(g; w_0) = \sum_{s,a,b} \big[\ell(g(b), s) - \ell(g(a), s)\big] \cdot \Pr(S = s, A_{w_0} = a, A_{w_1} = b \mid C = n). \tag{30}$$

Partitioning the loss difference by (S, A_{w_0}, A_{w_1}) proves the identity, including simultaneous switches of discrete source points. Applied to successive states, it also resolves intermediate changes whose sum gives the net endpoint effect.

C.2 Response of a fixed rule to intervention

On a smooth population trajectory $\dot{w}_t = h(w_t, \kappa_t)$, let $J_t = D_w h(w_t, \kappa_t)$. The propagation matrix $\Phi(t, s)$ carries a small state perturbation from time s to time t. It satisfies $\partial_t \Phi(t, s) = J_t \Phi(t, s)$ and $\Phi(s, s) = I$. For intervention forcing b_t in Eq. (7), variation of constants gives

$$\delta R_n(g; w_T) = D_w R_n(g; w_T)\Phi(T, 0)\delta w_0 + \int_0^T D_w R_n(g; w_T)\Phi(T, s) b_s \, ds. \tag{31}$$

A rate pulse has $b_t = \partial_\kappa h(w_t, \kappa_t)\delta\kappa_t$. An initial displacement instead enters through δw_0. For a specified differentiable interpolation of update fields h^ϵ, a graph perturbation supplies $b_t = \partial_\epsilon h^\epsilon(w_t)|_{\epsilon=0}$. Discrete training replaces Φ by products of actual update-map derivatives (Maclaurin et al., 2015; Franceschi et al., 2017; Baydin et al., 2018). Atomic switches and finite perturbations use the endpoint identity above. The experimental predictors approximate the response using features from the initial checkpoint and untreated continuation.

C.3 Optimal risk at a common budget

Compare endpoints using the same distribution, contexts, loss, alphabet and rule family at budget B, with $M(g; w) = \max_n R_n(g; w)$. For $g_i \in \arg\min_{g \in \mathcal{G}_B} M(g; w_i)$,

$$M(g_1; w_1) - M(g_1; w_0) \le \mathcal{R}_S^{\mathrm{grp}}(B; w_1) - \mathcal{R}_S^{\mathrm{grp}}(B; w_0) \le M(g_0; w_1) - M(g_0; w_0). \tag{32}$$

The lower bound follows from $M(g_1; w_0) \geq \mathcal{R}_S^{\mathrm{grp}}(B; w_0)$ and $M(g_1; w_1) = \mathcal{R}_S^{\mathrm{grp}}(B; w_1)$. Reversing endpoint roles proves the upper bound. Discrete sources and changes of the worst context are allowed. The maximum must be taken separately at each endpoint before subtracting risks.

For approximate feasible rules with gaps $\varepsilon_i = M(g_i; w_i) - \mathcal{R}_S^{\mathrm{grp}}(B; w_i) \geq 0$, subtract ε_1 from the lower bound and add ε_0 to the upper bound. Certified gap bounds then give corresponding bounds on optimal risk. Calibration-selected test risks quantify attained performance within the searched candidate family.

A minimum length B_* at risk threshold ϵ is certified by a rule g of length B_* with $M(g; w) \leq \epsilon$ and proof that $\mathcal{R}_S^{\mathrm{grp}}(B; w) > \epsilon$ at every smaller admissible budget. A feasible rule without the lower bound gives an attained cost.

D Residual growth and formation time

D.1 Residual geometry and local growth

With the encoder fixed, take independent symmetric factors (S, N). Assume bounded S, integrable N with density f_N, and $m = \mathbb{E}|N| > 0$. Freeze a first-layer codebook that quantizes the sign of N to $\pm m$. The scalar residual is $R = N - m\,\mathrm{sign}(N)$. Changing variables on the positive and negative branches of N gives $f_R(r) = f_N(r + m) + f_N(r - m)$ for $|r| < m$. If f_N is symmetric and continuous at $\pm m$, then $f_R(0) = 2f_N(m)$ and $\mathbb{E}|R| = \mathbb{E}\big||N| - m\big|$. Applying Eq. (11) to (S, R) gives the multiplier in Eq. (8) when $\mathbb{E}|R| > 0$.

For a symmetric pair $c_\pm = \pm(d, c_y)$ with $c_y > 0$, let $v = d/c_y$ and $\mu(v)$ be the positive-cell conditional mean of the fixed input (S, R). Each cell has probability $1/2$, so a squared codebook loss with weight $\alpha > 0$ gives

$$\dot{d} = \frac{\alpha}{2}(\mu_S(v) - d), \qquad \dot{c}_y = \frac{\alpha}{2}(\mu_R(v) - c_y), \qquad \dot{v} = \frac{\alpha}{2c_y}[\mu_S(v) - v\mu_R(v)]. \tag{33}$$

Linearizing at the context-aligned equilibrium $v = 0$, $c_y = \mathbb{E}|R|$ gives $\dot{v} = \alpha(\Lambda_{\mathrm{res}} - 1)v/2 + o(v)$. A synchronous Euler step has local direction multiplier $1 + \eta_C\alpha(\Lambda_{\mathrm{res}} - 1)/2$. These are gradient codebook updates.

To obtain opposite growth at equal residual variance, let $m = \sqrt{3}/2$, $d_{\mathrm{gap}} = \sqrt{33}/12$, and

$$N = U(m + V d_{\mathrm{gap}} + \epsilon), \qquad \epsilon \sim \mathrm{Unif}[-1/4, 1/4], \tag{34}$$

where U, V are independent equiprobable signs and are independent of ϵ and S. Since $m - d_{\mathrm{gap}} - 1/4 > 0$, $\mathrm{sign}(N) = U$ and $\mathbb{E}|N| = m$. Moreover,

$$\mathrm{Var}(N) = m^2 + d_{\mathrm{gap}}^2 + \frac{1}{48} = 1, \qquad R \overset{d}{=} V d_{\mathrm{gap}} + \epsilon, \qquad \mathrm{Var}(R) = \frac{1}{4}. \tag{35}$$

The residual has $|R| \geq d_{\mathrm{gap}} - 1/4$, so $f_R(0) = 0$ and $\Lambda_{\mathrm{res}} = 0$. For $|v| < d_{\mathrm{gap}} - 1/4$ and binary S, $\mathrm{sign}(R + vS) = \mathrm{sign}(R)$ and the positive-cell mean is $(0, d_{\mathrm{gap}})$. Lloyd therefore maps the direction to zero in one step. In contrast, the uniform source yields $R \sim \mathrm{Unif}[-m, m]$ and $\Lambda_{\mathrm{res}} = 8/3$. Both constructions match the first-layer context-aligned codebook, its distortion, and the residual variance. Their residual absolute means differ, at d_{gap} and $m/2$, respectively.

Eq. (33) therefore gives gradient growth rates $5\alpha/6$ and $-\alpha/2$ for the uniform and gapped residuals.

D.2 Continuous and discrete arrival bounds

Assume independent $S \in \{-1, 1\}$ with equal probabilities and $R \sim \mathrm{Unif}[-m, m]$, where $m = \sqrt{3}/2$, and freeze the inputs. Initialize symmetric codes with $0 < v_0 < m$ and $0 < c_{y,0} \leq m/2$. Let T_{hit} be the first continuous time with $v_t \geq m$, and J_{hit} the first Euler index with $v_j \geq m$. At this threshold the assignment separates the target classes almost surely. For gradient flow and, respectively, synchronous Euler steps with $0 < \eta_C\alpha/2 \leq 1$,

$$T_{\mathrm{hit}} \leq \frac{6}{5\alpha} \log \frac{m}{v_0}, \tag{36}$$

$$J_{\mathrm{hit}} \leq \left\lceil \frac{\log(m/v_0)}{\log(1 + 5\eta_C\alpha/6)} \right\rceil. \tag{37}$$

At arrival, $\operatorname{sign}(R + vS) = S$ almost surely, certifying a zero-error shared token test in every positive-mass context. The code vectors may continue moving.

For the continuous bound, direct integration for $0 < v < m$ gives

$$\mu_S(v) = \frac{v}{m}, \qquad \mu_R(v) = \frac{m^2 - v^2}{2m}, \qquad \dot{v} = \frac{\alpha v(2 - m^2 + v^2)}{4mc_y}. \tag{38}$$

Before arrival, $0 < \mu_R(v) \le m/2$. Variation of constants gives

$$c_y(t) = e^{-\alpha t/2} c_{y,0} + \frac{\alpha}{2} \int_0^t e^{-\alpha(t-s)/2} \mu_R(v_s)\, ds, \tag{39}$$

so $0 < c_y(t) \le m/2$ for every finite $t < T_{\text{hit}}$. Substituting into Eq. (38) yields $\dot{v} \ge \alpha(2 - m^2)v/(2m^2) = 5\alpha v/6$. Hence $v_t \ge v_0 e^{5\alpha t/6}$ until arrival, which forces v_t to reach m by the time in Eq. (36).

For the discrete bound, put $s = \eta_C \alpha/2 \in (0, 1]$. A superscript $+$ denotes the next iterate. One synchronous step before arrival is

$$c_y^+ = (1 - s)c_y + s\mu_R(v), \qquad v^+ = \frac{(1 - s)c_y v + s\mu_S(v)}{(1 - s)c_y + s\mu_R(v)}. \tag{40}$$

This convex combination preserves $0 < c_y^+ \le m/2$, including $s = 1$. Therefore

$$v^+ - v = \frac{s[\mu_S(v) - v\mu_R(v)]}{(1 - s)c_y + s\mu_R(v)} = \frac{s\, v(2 - m^2 + v^2)}{2m[(1 - s)c_y + s\mu_R(v)]} \ge \frac{5\eta_C \alpha}{6} v. \tag{41}$$

Induction gives $v_j \ge (1 + 5\eta_C \alpha/6)^j v_0$ while $v_j < m$, proving Eq. (37).

For arbitrary $c_{y,0} > 0$, set $c_{\max} = \max\{c_{y,0}, m/2\}$ and $\gamma_{\text{arr}} = \alpha(2 - m^2)/(4mc_{\max})$. The continuous solution and discrete convex combination preserve $0 < c_y \le c_{\max}$ before arrival. The same proofs then give

$$T_{\text{hit}} \le \frac{1}{\gamma_{\text{arr}}} \log \frac{m}{v_0}, \qquad J_{\text{hit}} \le \left\lceil \frac{\log(m/v_0)}{\log(1 + \eta_C \gamma_{\text{arr}})} \right\rceil. \tag{42}$$

All other source, initialization and step-size assumptions above remain in force. Exactly zero initial target direction stays on the residual axis by Eq. (33). Earlier-layer motion changes these frozen-input dynamics through the forcing derived next.

D.3 Earlier-layer motion and signed forcing

Let y_t contain the encoder and preceding-layer states that determine layer ℓ's input distribution. Let c_ℓ collect that layer's code vectors and $F_\ell(c_\ell, y)$ denote their deterministic update field. Suppose a differentiable equilibrium $c_\ell^*(y)$ exists locally for each frozen y, with $F_\ell(c_\ell^*(y), y) = 0$. The displacement from this moving equilibrium is $z_\ell = c_\ell - c_\ell^*(y_t)$. The chain rule and a local expansion give

$$\dot{z}_\ell = A_\ell(t) z_\ell - D_y c_\ell^*(y_t)\dot{y}_t + \mathcal{E}_\ell(z_\ell, t), \qquad A_\ell(t) = D_{c_\ell} F_\ell(c_\ell^*(y_t), y_t). \tag{43}$$

Here $A_\ell(t)$ gives the response with preceding states fixed, and $-D_y c_\ell^*(y_t)\dot{y}_t$ accounts for motion of the reference equilibrium. The remainder is locally $o(\|z_\ell\|)$, or $O(\|z_\ell\|^2)$ with a bounded second derivative. Let $\Phi_\ell(t, s)$ satisfy $\partial_t \Phi_\ell(t, s) = A_\ell(t)\Phi_\ell(t, s)$ and $\Phi_\ell(s, s) = I$. Variation of constants gives

$$z_\ell(t) = \Phi_\ell(t, t_0) z_\ell(t_0) + \int_{t_0}^t \Phi_\ell(t, s)[-D_y c_\ell^*(y_s)\dot{y}_s + \mathcal{E}_\ell(z_\ell(s), s)]\, ds. \tag{44}$$

The retained remainder makes this representation exact on the local trajectory. Matching a projected growth coefficient and its time integral still leaves signed forcing free to assist or oppose alignment. The full propagation matrix carries multidimensional perturbations. In joint training, y_t evolves through its feedback with the later layer. The controlled-path experiment isolates this driving effect by prescribing y_t externally.

E EXPERIMENTAL PROTOCOLS AND SUPPORTING RESULTS

E.1 EVALUATION UNITS AND RULE LANGUAGE

Seeds index independently initialized runs, with paired arms sharing a parent checkpoint and specified inputs. Contexts, budgets, anchors and repeated initializations within a seed are dependent observations. Bootstrap confidence intervals (Efron, 1979) resample seeds, preserving treatment–control and Full–comparator pairs. Boundary-estimation intervals quantify anchor uncertainty, prediction intervals cover a new seed's response set, and ECG risk bounds account for patient sampling.

Rule language. The categorical rule language uses binary trees whose nodes test whether one layer's token belongs to a nonempty proper subset of its categories. A test encodes its operation, layer, subset size and subset identity. These cost two, $\lceil \log_2 L \rceil$, $\lceil \log_2(K_\ell+1) \rceil$ and $\lceil \log_2 \binom{K_\ell}{m} \rceil$ bits, respectively. Here L is the number of layers, K_ℓ the layer-ℓ codebook size, and m the subset size. A leaf predicts one of K_S target classes and costs $2 + \lceil \log_2 K_S \rceil$ bits. Tree costs sum all test-node and leaf costs. A complete table over $D = \prod_\ell K_\ell$ tuples costs $2 + \lceil \log_2(D+1) \rceil + D(\lceil \log_2 D \rceil + \lceil \log_2 K_S \rceil)$ bits, including unused tuples. Shared subrules and uncharged ordinal thresholds are excluded.

For the 4×2 token interface with a binary target, exhaustive search enumerates all $2^8 = 256$ truth tables, minimizes tree cost over every valid split and compares the complete-table cost. It gives the exact worst-context risk at each budget, including unused tuples. The illustrated 15-bit rule costs $9 + 2 \times 3$ bits. The 34-bit rule costs $8 + 8 + 6 + 4 \times 3$ bits. Renumbering tokens within a layer preserves these costs. Larger interfaces use finite candidate families and report attained risks.

All shared rules exclude the context index. Recovery follows Sec. 3.1, with rule and probe pass counts reported separately. ECG uses patient-level bounds and parent-selected functions.

Endpoint evaluation. Small-model and Shapes3D endpoint comparisons use separate readout-fitting, calibration and test data. At each endpoint, calibration selects a fitted rule with lowest worst-context risk within budget and independently selects the probe. Both functions are frozen before testing. Small-model rule ties favor shorter candidates, then fixed candidate order, while probe ties use candidate order. Shapes3D candidate families and selection orders are specified in App. E.5. Continuous baselines receive their declared continuous interface.

Noiseless finite-source risks and recovery checks use exact population evaluation. Empirical readout checks use nominal Wilson bounds (Wilson, 1927), assuming independent evaluation records within each context. For empirical probe error rate p_n among m_n test records in context n, let z be the standard-normal quantile at $1 - 0.05/(2|\mathcal{C}_{\text{eval}}|)$. The context-adjusted Wilson upper bound is

$$U_n = \frac{p_n + z^2/(2m_n) + z\sqrt{p_n(1-p_n)/m_n + z^2/(4m_n^2)}}{1 + z^2/m_n}. \tag{45}$$

A probe passes when every upper bound is at most $\varepsilon = 0.05$, over four small-model contexts or 138 ordinary Shapes3D test contexts. The 2.5th and 97.5th percentiles of paired resampled seed means give 95% CIs for treatment-minus-control rule and probe risks. Small-model probe equivalence requires the probe-risk CI to lie strictly inside $(-0.02, 0.02)$.

Fixed-rule forecasts. For evaluation data $\mathcal{D}$ with inputs $X_{\mathcal{D}}$, let $\mathcal{D}_n$ contain the examples in context $C = b(N) = n$. The empirical context risk is

$$\widehat{R}_n^{\mathcal{D}}(g; w) = |\mathcal{D}_n|^{-1} \sum_{(x,s)\in\mathcal{D}_n} \mathbf{1}\{g(A_w(x)) \neq s\}. \tag{46}$$

In Algorithm 1, $\bar{w}_j$ is the untreated state and h_j its training update field evaluated on anchors. For a nonzero perturbation z, set $u = z/\|z\|$ and

$$\widehat{J}_j^c z = \frac{\|z\|}{2\delta}[h_j^c(\bar{w}_j + \delta u) - h_j^c(\bar{w}_j - \delta u)], \qquad \widehat{J}_j^c 0 = 0, \quad c \in \{\mathrm{F}, \mathrm{B}\}.$$

The Full field h_j^{F} recomputes assignments at each perturbed state. The Branch field h_j^{B} holds complete token tuples fixed at their assignments under $\bar{w}_j$.

A rate pulse supplies $b_j = \delta\kappa_j[h_j(\bar{w}_j; 1) - h_j(\bar{w}_j; 0)]$, where $\delta\kappa_j = \kappa_j - 1$ and the second argument overrides only the encoder rate. A displacement sets z_0 with $b_j = 0$. Backward path uses $b_j = h_j^{\mathrm{back}}(\bar{w}_j) - h_j(\bar{w}_j)$ on the same untreated state and anchors. In the small model, this removes the stop-gradient from earlier residual subtraction, preserving the overall identity STE and forward quantization rule. Full and Branch use the same intervention forcing throughout its duration. The small-model forecast evaluates endpoint risk, including finite assignment changes. The dataset protocols below specify intervention durations, forecast horizons and the ridge models that convert response features into predictions.

Algorithm 1 Fixed-rule response computation

Require: Parent w_0, fixed rule g, anchors $\mathcal{D}$, horizon J, step η_C, radius δ, forcing b_j, displacement z_0.
Ensure: Full and Branch response features.

$\bar{w}_0 \leftarrow w_0$, $z_0^{\mathrm{F}} \leftarrow z_0^{\mathrm{B}} \leftarrow z_0$.
for $j = 0, \ldots, J-1$ **do**
  Freeze assignments in h_j^{B} at $A_{\bar{w}_j}(X_{\mathcal{D}})$.
  Evaluate b_j at untreated state $\bar{w}_j$.
  for $c \in \{\mathrm{F}, \mathrm{B}\}$ **do**
    $v_j^c \leftarrow \widehat{J}_j^c z_j^c$.
    $z_{j+1}^c \leftarrow z_j^c + \eta_C(v_j^c + b_j)$.
  end for
  $\bar{w}_{j+1} \leftarrow \bar{w}_j + \eta_C h_j(\bar{w}_j)$.
end for
for $c \in \{\mathrm{F}, \mathrm{B}\}$ and each context n **do**
  $x_n^c \leftarrow \widehat{R}_n^{\mathcal{D}}(g; \bar{w}_J + z_J^c) - \widehat{R}_n^{\mathcal{D}}(g; \bar{w}_J)$.
end for
return $\{x_n^{\mathrm{F}}, x_n^{\mathrm{B}}\}_n$.

E.2 Learned rule complexity

The source is

$$X = (S, N, \tau SN) + U, \qquad S \in \{-1, 1\}, \quad N \in \{-3, -1, 1, 3\}/\sqrt{5},$$

with eight equally weighted factor combinations and independent coordinate noise $U_j \sim \mathrm{Unif}[-\nu, \nu]$. The four contexts correspond to the four values of N. Class indices zero and one encode $S = -1$ and $S = +1$, respectively.

The discovery sweep and small-model interventions use $3 \to 16 \to 3$ tanh encoders and independent decoders. Two residual codebooks contain four and two three-dimensional code vectors, giving 248 parameters. Training follows Eq. (15) with $\alpha = 1$, $\beta = 1/4$, coordinate and layer sums, and a sample mean. Stochastic gradient descent (SGD) evaluates all gradients at the same pre-update state, without momentum, weight decay or dead-code resets.

Table S1: Exact assignments and risk–length frontiers for Fig. 3. Cells are zero-based token tuples (T_1, T_2); contexts 1–4 index $N = (-3, -1, 1, 3)/\sqrt{5}$.

	$\kappa = 1$		$\kappa = 10$	
Context	$S = -1$	$S = +1$	$S = -1$	$S = +1$
1	$(2, 1)$	$(2, 0)$	$(1, 1)$	$(2, 0)$
2	$(0, 1)$	$(3, 0)$	$(0, 0)$	$(3, 1)$
3	$(0, 0)$	$(1, 1)$	$(0, 1)$	$(3, 0)$
4	$(3, 1)$	$(1, 0)$	$(1, 0)$	$(2, 1)$
Risk	Budget range		Budget range	
$1/2$	$3 \le B < 34$		$3 \le B < 15$	
0	$B \ge 34$		$B \ge 15$	

The sweep crosses $\tau \in \{0, 0.5\}$, $\nu \in \{0, 0.05\}$ and $\kappa \in \{0.1, 1, 10\}$, with 32 seeds per setting. Training uses 20,000 updates at codebook step size 0.001, corresponding to codebook time 20, and $\kappa_D = 1$. Discovery batches contain eight inputs without noise and 128 with noise. Zero-noise risks use the exact eight-atom population. Noisy discovery and endpoint-intervention readouts use 8,192 fitting, 4,096 calibration and 16,384 test inputs, balanced over the eight factor combinations. The test split thus contains 4,096 inputs per context. Intervention updates instead use 64 fixed inputs (App. E.4). The noiseless pair in Fig. 3 shares one initialization. Table S1 gives all eight assignments and exact risk–length frontiers. Table S2 combines all 12 conditions with paired rate effects.

E.3 Geometry and training stability

The geometry sweep varies ξ over nine values while fixing probability mass, variance, absolute first moment and both reference distortions. Finite-sample Lloyd updates compare a fixed empirical

Table S2: Learned organizations (32 seeds per setting, $B = 16$). (a) All 12 conditions: Probe and Rule are mean worst-context risks; Pass counts successful probe checks. (b) Paired $\kappa : 1 \to 10$ risk changes and decrease/tie/increase counts. Brackets: 95% seed CIs.

(a) Discovery sweep

		No noise ($\nu = 0$)				Noise ($\nu = 0.05$)			
τ	κ	Pass/32	Probe	Rule [95% CI]	MSE	Pass/32	Probe	Rule [95% CI]	MSE
0	0.1	32	0.000	0.375 [0.297, 0.438]	0.005	32	0.004	0.290 [0.213, 0.368]	0.047
	1	28	0.062	0.312 [0.234, 0.391]	0.031	32	0.004	0.291 [0.213, 0.369]	0.046
	10	31	0.016	0.328 [0.250, 0.406]	0.010	32	0.004	0.234 [0.156, 0.311]	0.046
0.5	0.1	32	0.000	0.328 [0.250, 0.406]	0.005	32	0.004	0.382 [0.306, 0.443]	0.052
	1	31	0.016	0.344 [0.266, 0.422]	0.014	32	0.003	0.336 [0.260, 0.412]	0.052
	10	30	0.031	0.250 [0.156, 0.344]	0.025	31	0.018	0.230 [0.153, 0.306]	0.059

(b) Paired rate effects

	No noise ($\nu = 0$)				Noise ($\nu = 0.05$)			
τ	ΔRule [95% CI]	Decrease	Tie	Increase	ΔRule [95% CI]	Decrease	Tie	Increase
0	0.016 [−0.094, 0.125]	5	21	6	−0.058 [−0.179, 0.063]	20	1	11
0.5	−0.094 [−0.203, 0.016]	10	18	4	−0.106 [−0.226, 0.015]	21	0	11

sample with fresh batches. Each of 64 seeds supplies 16 initial code-vector pairs, with all four coordinates independently drawn from $\mathcal{N}(0, 1)$. After 60 Lloyd updates, a codebook enters a reference branch if its Frobenius distance, minimized over the two vector orders, is at most 0.1. Empty cells retain their previous code vectors. Fig. 4b and Fig. S1c average these within-seed entry frequencies for fixed samples and fresh batches, respectively.

The joint-training sweep evaluates all six eigenvalues for both factor-aligned branches at nine source parameters and 81 encoder rates from 0.05 to 20. The target-aligned decay rates include a repeated eigenvalue and agree across source distributions to 10^{-12}. With an independent decoder at $\xi = 0$, all six settings have two numerically neutral modes. At decoder rates $\kappa_D = (0.1, 1, 10)$, the largest nonneutral real parts are $(0.515528, 0.446724, 0.753823)$ for $\kappa = 1$ and $(-0.007263, -0.076945, -0.25)$ for $\kappa = 10$. Both J and J^2 have numerical nullity two. Their nullspace matches the tangent space of a two-parameter equilibrium family within projection error 5.4×10^{-9}. One tangent changes the assignment threshold, so stable nonneutral modes do not establish strict stability of an individual organization.

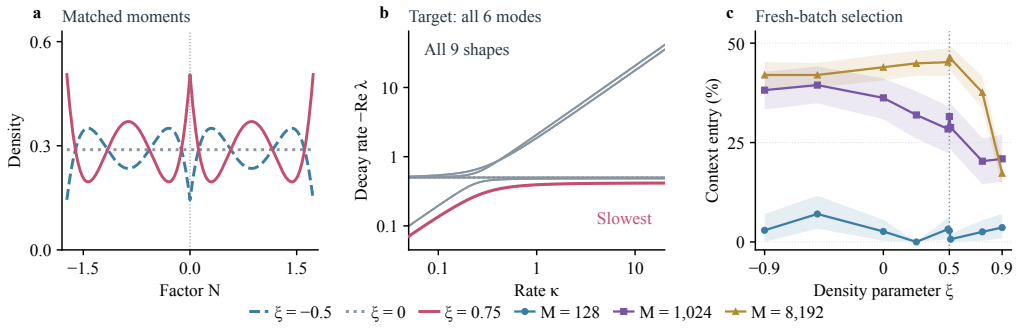


Figure S1: Source geometry and stability. (a) Matched-moment densities. (b) Six target decay modes across nine source shapes. (c) Fresh-batch entry (64 seeds, 95% CIs). ξ: density parameter. M: sample size.

Finite-perturbation relaxation. For the uniform source, we integrate the six-dimensional population field with independent code vectors and the specified STE paths. At each of $\delta = \kappa - \kappa_H \in \{0.01, 0.04\}$, four perturbation directions have state-displacement norm 0.00460226. Relaxation

time is the time for the oscillatory modal amplitude to fall by a factor of e. The four times span 56.27–57.40 at $\delta = 0.01$ and 14.13–14.39 at 0.04, close to Eq. (26).

The 48 configurations also include zero and larger perturbations, with and without a rate pulse. All 16 zero-perturbation configurations remain at equilibrium, while larger perturbations can leave the radius-0.5 neighborhood. Six-dimensional shooting finds unstable periodic orbits at $\delta = 0.005, 0.01, 0.02$, with maximum Floquet moduli $1.39, 2.18, 9.62$ and residuals at most 1.32×10^{-10}. At $\delta = 0.04$ and 0.08, the shooting computation leaves the local chart or fails to converge. These integrations test local return near equilibrium (App. B.4).

E.4 Fixed-rule response prediction

Small-model interventions. The controlled task fixes $\tau = 0.5$ and $\nu = 0.05$, averaging updates over 64 fixed inputs, eight per factor combination. From a parent at time 2, each arm continues to time 4 with step size 0.005. Slow and Fast set κ to 0.1 and 10 over $[2, 2.2]$, then restore it to one. Backward path remains active throughout continuation. Random direction applies a one-time displacement of norm 0.3477377, calibrated on eight separate seeds. Each of 64 confirmation seeds supplies all five arms. Each endpoint selects its own rule at $B = 16$ and its own probe, following App. E.1. Table 1 reports all endpoint comparisons.

Small-model forecast. All parents use the same 12-bit rule, fixed before intervention: predict class zero when the second-layer token is zero and class one otherwise. Four interventions and four contexts give 16 treatment-minus-control error responses per seed. A constant rule supplies a zero-response control. Algorithm 1 uses central differences at normalized radius 0.003 along the untreated 248-parameter trajectory. Other features measure boundary mass that can change the rule output, forecast token changes, parameter displacement or static reconstruction error. Features are standardized on fitting seeds and passed through ridge regression, with penalties 0.1 for Full and Branch and 10 otherwise. Disjoint fitting, tuning, interval-calibration and confirmation cohorts contain 64, 32, 64 and 64 seeds.

Endpoint organization examples. The two cases in Fig. 5 illustrate opposite changes in short-rule access within the 64-parent confirmation cohort. They were selected after evaluation. Table S3 links the forecasts, exchanges and endpoint rules for these same parents. Forecasts use untreated continuations and previously fitted regression coefficients. Exchange counts use 16 inputs per context, and held-out responses use 4,096. Reported lengths are attained zero-test-error costs, whereas Table S1 certifies population minima on the noiseless source.

Table S3: Forecasts and endpoints for the two cases in Fig. 5. Test changes use 4,096 inputs per context; exchange counts use 16. All four endpoints have zero probe error and upper risk 0.0015.

(a) Fixed 12-bit rule: Fast minus Control risk

Case	Context	Full	Branch	Observed test	Exchange sum
I	1	−0.0007	+0.0014	0.0000	0.0000
I	2	−0.2433	−0.1586	−0.0559	−0.0625
I	3	−0.2433	−0.2787	−0.1265	−0.1250
I	4	−0.0007	+0.0014	0.0000	0.0000
II	1	−0.0007	+0.0014	+0.5000	+0.5000
II	2	−0.0007	+0.0014	0.0000	0.0000
II	3	−0.1827	−0.1186	+0.5000	+0.5000
II	4	−0.0007	+0.0014	0.0000	0.0000

(b) Endpoint risks and attained zero-test-error rule lengths

Case	Arm	Fixed-rule risk	Selected risk ($B = 16$)	Attained bits
I	Control	0.5000	0.5000	34
I	Fast	0.5000	0.0000	14
II	Control	0.5000	0.0000	15
II	Fast	1.0000	0.5000	34

Boundary estimation and sampling resolution. The anchor analysis crosses total sample counts $M \in \{1{,}024, 4{,}096, 16{,}384, 65{,}536\}$ with $\delta \in \{0.01, 0.003, 0.001\}$. Anchors are balanced over eight factor combinations, with independent noise within each combination and smaller samples nested within larger ones. We perturb the encoder along its normalized parent gradient, computed on the 64 training inputs, and evaluate the fixed rule at both perturbed states. All later residual tokens are recomputed after an earlier assignment changes. Paired risk differences divided by 2δ estimate finite-difference quotients. Empirical-Bernstein intervals account for the balanced strata, giving simultaneous 95% coverage over four contexts for each seed and setting (Maurer & Pontil, 2009). Fig. S2 summarizes the resulting 256 seed–context cases per setting. No observed crossing means no token-tuple change between perturbed states, while a resolved sign requires the quotient interval to exclude zero. Perturbation sensitivity is the mean absolute quotient change on common anchors relative to $\delta = 0.001$.

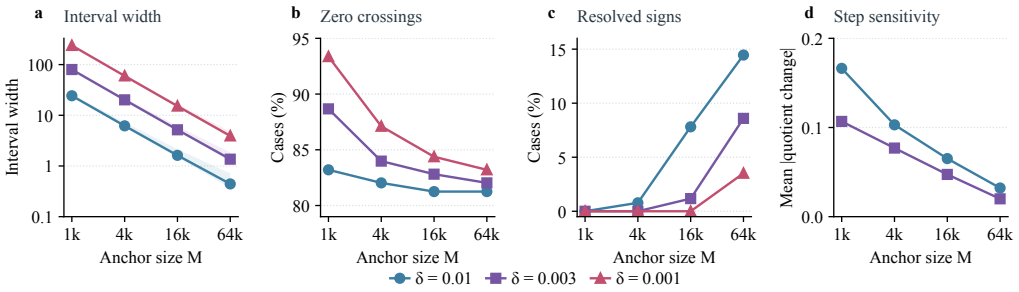


Figure S2: Boundary estimation (64 seeds, four contexts). (a) Mean interval width. (b) No crossings. (c) Resolved signs. (d) Perturbation sensitivity. Intervals: anchor uncertainty. Shading in (a): seed–context 10th–90th percentiles.

Shapes3D forecast. The parent fixes a function predicting the most frequent target class for each token tuple in readout-training data. Arms share the parent and minibatch sequence. Two token rates, four interventions and 138 contexts yield 1,104 responses per seed. Training and calibration anchors supply a 16-step forecast with central Jacobian-vector products at normalized displacement 10^{-3}. Predictor inputs exclude treated endpoints and test responses. Disjoint fitting, tuning, interval-calibration and confirmation cohorts contain 64, 32, 64 and 32 seeds.

Full combines the full-response feature with boundary mass, churn, displacement and static reconstruction. Branch combines the fixed-assignment response with displacement and static reconstruction. Other predictors combine their named feature with static reconstruction. All include intervention and token-rate terms, with ridge penalty 100. Backward path removes the encoder commitment term throughout the pulse and recovery phases at $\kappa = 1$.

Metrics and simultaneous intervals. Metrics average seeds equally. The small-model analysis ranks 16 arm–context responses within each seed. Shapes3D averages responses over contexts, ranks four interventions at each token rate, then averages the defined rate-level correlations. Rank correlations are jointly defined for 39 small-model seeds for Full versus Branch and 46 for each other fitted comparator. Shapes3D uses all 32 pairs. Paired differences use these shared seeds and can differ from differences between marginal means. Sign error treats negative, zero and positive responses as distinct categories. The Zero baseline predicts no change. Its constant predictions give undefined rank correlation, and we report it without a calibrated interval.

The split-conformal score is the maximum absolute prediction residual across a seed's complete response set (Lei et al., 2018). The 59th and 62nd of 64 ordered calibration scores set radii for nominal 90% small-model and 95% Shapes3D coverage, respectively. A seed is covered only when every response is covered. Reported widths precede display clipping.

Table S4: Fixed-rule prediction results. MAE: 10^{-3}. Rank: Spearman correlation (n: defined seeds). Sign err.: decrease/tie/increase error. Covered: seeds with every response covered. Width: risk units before clipping. Brackets: paired 95% seed CIs.

	Small model (64 seeds)					Shapes3D (32 seeds)				
Predictor	MAE	Rank (n)	Sign err.	Covered	Width	MAE	Rank (n)	Sign err.	Covered	Width
Full	25.65	0.591 (46)	0.880	61/64	0.9985	4.39	0.598 (32)	0.677	31/32	0.2863
Branch	24.90	0.640 (39)	0.879	61/64	0.9972	4.51	0.280 (32)	0.687	30/32	0.2570
Boundary	28.00	-0.110 (46)	0.936	61/64	1.0123	7.10	-0.216 (32)	0.758	31/32	0.6598
Churn	28.84	0.106 (46)	0.915	63/64	1.0222	7.32	0.007 (32)	0.724	31/32	0.7605
Displace.	28.19	0.023 (47)	0.922	63/64	1.0023	7.77	-0.206 (32)	0.748	31/32	0.6809
Static	28.11	0.019 (47)	0.918	61/64	1.0013	6.79	-0.096 (32)	0.723	31/32	0.6550
Zero	27.59	– (0)	0.149	–	–	5.49	– (0)	0.522	–	–

Full − comparator		Small model		Shapes3D	
Comparator	Metric	n	Difference [95% CI]	n	Difference [95% CI]
Branch	MAE	64	0.7461 [−2.4980, 4.4703]	32	−0.1209 [−0.1455, −0.0959]
	Rank	39	−0.0403 [−0.1871, 0.1075]	32	0.3185 [0.2268, 0.4036]
	Sign err.	64	0.0010 [−0.0166, 0.0205]	32	−0.0104 [−0.0152, −0.0057]
Churn	MAE	64	−3.1975 [−6.9973, 0.8560]	32	−2.9306 [−3.8179, −2.2436]
	Rank	46	0.4845 [0.2954, 0.6709]	32	0.5915 [0.4579, 0.7151]
	Sign err.	64	−0.0352 [−0.0586, −0.0127]	32	−0.0471 [−0.0539, −0.0398]

E.5 SHAPES3D TRAINING, HOLDOUTS AND RULE BUDGETS

Input and training. Shapes3D (Burgess & Kim, 2018) contains 64×64 RGB renderings varying floor hue, wall hue, object hue, scale, shape and orientation. We use the four shape classes as targets and object hue and orientation to define contexts. All seed-specific quantization models receive the same frozen 16-dimensional image representation. Reconstruction MSE averages squared error over these 16 coordinates. The models learn 16×16 linear encoders and decoders, with two residual codebooks containing eight or 64 vectors each at six or twelve token bits. Squared losses are averaged over the 16 coordinates, with encoder commitment applied to the final residual $Z - Q$. The step size is 0.04 and commitment weight is 0.25, increased to one for Strong commitment. The initial shared checkpoint receives 16 updates. Pulse conditions add 16 warm-up, eight pulse and eight recovery updates, totaling 48. Slow and Fast use $\kappa = 0.1$ and 10 during the pulse, then return to one. Control uses one throughout.

For exported hue and orientation coordinates $h, o \in [0, 2\pi)$, the context map is $C = 15\lfloor 10h/(2\pi) \rfloor + \lfloor 15o/(2\pi) \rfloor$, giving 150 contexts. The coordinates are continuous encodings of the source factor categories. Each of the 138 ordinary contexts supplies 64 readout-training, 128 calibration and 512 test records, totaling 8,832, 17,664 and 70,656, respectively. Target-class frequencies are not forced to balance. Quantization-model training additionally includes the 3,072 readout-only holdout records. Target and context labels are excluded from quantization-model losses. All comparisons concern downstream training on this fixed representation.

Structural comparisons and budget dependence. Table S5 reports all 14 methods at both token rates on the same 32 confirmation seeds. The continuous AE exposes its latent vector, while discrete methods expose their token tuples. Fixed scalar uses Cartesian scalar levels, the central construction of finite scalar quantization (Mentzer et al., 2024). Frozen encoder selects the lowest training reconstruction loss among eight codebook restarts of 32 steps each. This uses 272 updates including initialization, compared with 48 for the pulse conditions. The Frozen prototypes and Gradient prototypes pair shares its encoder and initial codebooks, changing whether codebooks are updated. Continuous pretraining adds 16 continuous-AE updates before 16 quantized fitting updates.

Fig. S3 evaluates rules selected independently by calibration at each of 14 budgets, allowing nonmonotonic test curves. The continuous AE's first nonconstant candidate costs 1,091 bits, explaining its constant-only behavior at $B = 256$.

Table S5: Shapes3D comparisons ($B = 256$, 32 paired seeds). MSE: reconstruction error. Rule: worst-context risk. Δ: treatment minus Control [95% seed CI]. AE has no token rate.

	6-bit setting			12-bit setting		
Method	MSE	Rule	ΔRule [95% CI]	MSE	Rule	ΔRule [95% CI]
Control	0.0734	0.5750	Ref.	0.0245	0.2253	Ref.
Slow	0.0734	0.5809	$+0.0059$ [-0.0006, 0.0178]	0.0245	0.2253	0.0000 [-0.0002, 0.0002]
Fast	0.0736	0.5624	-0.0126 [-0.0267, -0.0029]	0.0244	0.2252	-0.0001 [-0.0010, 0.0009]
Backward path	0.0734	0.5752	$+0.0002$ [-0.0005, 0.0009]	0.0245	0.2256	$+0.0002$ [-0.0001, 0.0006]
Random direction	0.0758	0.5737	-0.0013 [-0.0197, 0.0181]	0.0245	0.2252	-0.0002 [-0.0012, 0.0008]
Frozen encoder	0.0683	0.5070	-0.0681 [-0.1225, -0.0120]	0.0235	0.2176	-0.0078 [-0.0320, 0.0156]
Continuous pretraining	0.0733	0.5827	$+0.0077$ [0.0010, 0.0197]	0.0245	0.2260	$+0.0007$ [0.0002, 0.0012]
Strong commitment	0.0735	0.5798	$+0.0048$ [-0.0025, 0.0173]	0.0245	0.2250	-0.0004 [-0.0009, 0.0001]
Frozen prototypes	0.0736	0.5856	$+0.0106$ [0.0036, 0.0226]	0.0246	0.2264	$+0.0011$ [0.0004, 0.0018]
Gradient prototypes	0.0733	0.5826	$+0.0076$ [0.0010, 0.0197]	0.0245	0.2256	$+0.0003$ [-0.0002, 0.0009]
Single VQ	0.0470	0.2290	-0.3460 [-0.3854, -0.3080]	0.0134	0.7988	$+0.5735$ [0.5545, 0.5914]
Fixed scalar	0.1161	0.5253	-0.0497 [-0.1157, 0.0214]	0.1145	0.4883	$+0.2630$ [0.2084, 0.3204]
Learned scalar	0.1161	0.5254	-0.0496 [-0.1152, 0.0179]	0.1145	0.4883	$+0.2630$ [0.2076, 0.3187]
Continuous AE	$< 10^{-9}$	0.7988	$+0.2238$ [0.1913, 0.2579]	$< 10^{-9}$	0.7988	$+0.5735$ [0.5543, 0.5920]

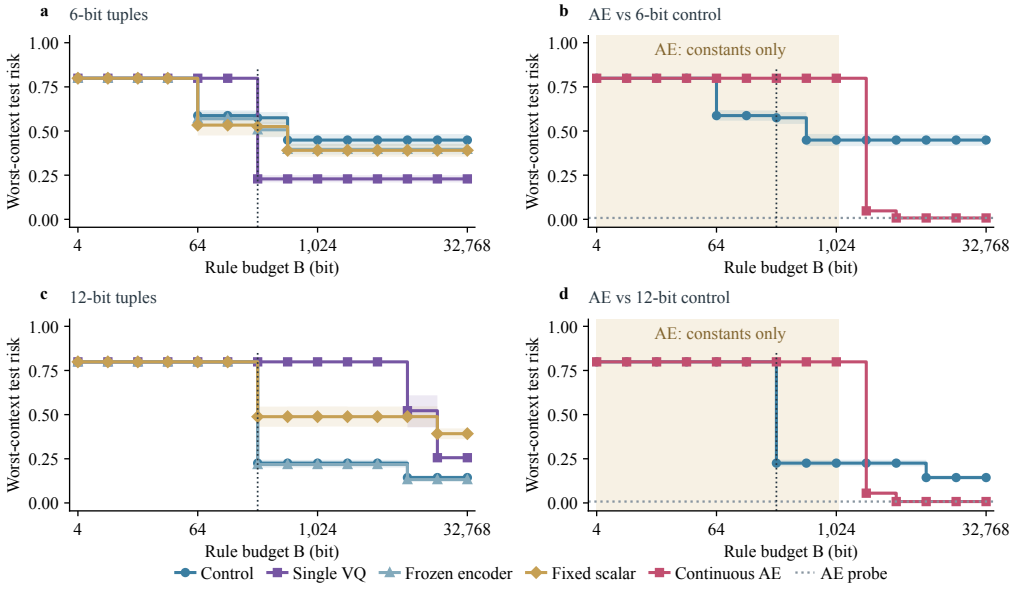


Figure S3: Rule-budget curves (32 seeds, 95% CIs). (a,c) Discrete methods. (b,d) Continuous AE comparisons. Dotted references: $B = 256$ and AE probe risk. Gold shading: constant AE rules only.

Holdouts and readout fitting. Contexts $\{30, 31, 32, 45, 46, 47\}$ enter quantization-model training but are excluded from readout fitting. Contexts $\{0, 1, 2, 15, 16, 17\}$ are excluded from both. Each holdout contains six contexts of 512 records, totaling 3,072 records. Prediction and structural-comparison metrics use the 138 ordinary-test contexts, also with 512 records each. Each run reuses its calibration-selected rule and probe across all three splits. Holdout risks take a maximum over six contexts, compared with 138 for the ordinary test. The readout holdout is transductive because its

inputs enter quantization-model training. Record IDs are disjoint across splits. Shapes3D Wilson bounds use these records as evaluation units and do not adjust for possible clustering by source image. Holdout pass counts apply the context correction separately to each fixed function, seed and budget.

The probe family contains linear fits and fixed hidden-feature maps with one hidden layer of width 128 or two of width 256. Both hidden-feature maps have fitted output layers. Fixed index lists select eight distinct training records per ordinary context. Four per context form a 552-record subset, and four more extend it to 1,104 records. Calibration selects among the resulting six candidates by worst-context risk, then mean risk and candidate order.

Categorical rules include constants and maps on the first-layer token, last-layer token and complete tuple. Each continuous interface has four constants, six axis trees and two oblique classifiers. Axis trees use a 256-level grid and leaf limits 4, 16 and 64, each fitted on 4,416 or 8,832 records. Oblique candidates use all 8,832 records and 8- or 16-bit coefficients. Within budget, selection minimizes the chosen calibration objective (worst-context or mean risk), then the other risk, length and candidate order. Nonconstant continuous candidates pay a 1,033-bit normalization/grid header. Trees with J leaves cost $1033 + 4J + 14(J - 1)$ bits, and oblique candidates cost 1,590 or 2,134 bits. Constants cost four bits.

A bitmask alternative charges K_ℓ membership bits per layer while retaining operation, layer and output costs, and reselects among the same candidate functions on calibration data. At $B = 256$, five of 896 method–rate–seed selections change, all at six token bits. Thus the selected rules are largely unchanged by this alternative encoding at the evaluated budget.

E.6 Residual formation and joint training

Frozen-input reference. The uniform and gapped references share variance 0.25 but have absolute first moments 0.4330 and 0.4787. Of 72 deterministic arrival configurations, 45 satisfy the principal nonzero-initialization conditions and obey both continuous and discrete bounds. Euler uses step size 0.01, with numerical runs capped at codebook time 100. Twelve zero-tilt controls never develop a target-direction tilt. The other 15 have $c_{y,0} > m/2$ and use the extended bound with $c_{\max}$.

Controlled preceding-layer motion. With a frozen encoder, positive and negative target-direction paths move the preceding codebook before returning it to its initial value. The pulse has amplitude ± 0.05, starts at time one and uses ramp duration 0.25, while the context component stays at m. For uniform input and duration 0.25, both paths start at $d_0 = 0.004330127$. Their target-coordinate Jacobian entries $J_{dd}(t)$ match on the evaluation grid to 4×10^{-12}. Integrating over the full grid gives $\int_0^{21.25} J_{dd}(t)\, dt = 17.7091$ for both paths. The paths match this projected growth measure while supplying opposite signed forcing. All twelve uniform-source paths across six durations admit a 12-bit zero-risk rule at the endpoint. Four opposing paths, with durations 1, 2, 4 and 8, reverse token polarity and never reach the positive-direction threshold by horizons 22, 23, 25 and 29. The reversed token test first has zero risk on the evaluation grid at times 9.7, 6.0, 5.4 and 5.5, respectively. It remains error-free at all subsequent evaluations. These times measure formation of a rule of either polarity on a grid spaced by 0.1.

Joint training. The pulse-duration experiment uses 64 parents trained to time 2 in the 248-parameter model. Continuations on 64 fixed inputs use step size 0.01 for 10 additional time units. Relative encoder rate $\kappa = 0.1$ or 3 is applied for 0.25, 0.5, 1, 2, 4 or 8 time units, then restored to one. Readouts use 128 parent-calibration and 512 test inputs, balanced over factor combinations. Parent-calibrated rules and probes remain fixed across arms and times. The probes use categorical tables on token interfaces and trees with maximum depth eight and minimum leaf size two on continuous interfaces. Success requires the fixed rule's empirical risk and fixed probe's upper risk bound both to be at most 0.05. We evaluate success at relative times $0, 0.25, 0.5, 1, 2, 4, 8, 10$, with unobserved events censored at the common horizon. Endpoint success is evaluated separately.

Table S6: Joint-training readout counts (64 paired seeds). Endpoint: passes at relative time 10. Ever: passes at any evaluation time, including the parent.

	$\kappa = 0.1$		$\kappa = 3$	
Duration	Endpoint	Ever	Endpoint	Ever
0.25	5/64	34/64	5/64	32/64
0.5	6/64	35/64	5/64	33/64
1	6/64	35/64	6/64	33/64
2	6/64	34/64	6/64	33/64
4	7/64	34/64	6/64	34/64
8	7/64	34/64	5/64	34/64

Table S6 reports endpoint and ever-passing counts across the twelve conditions. Longer pulses leave less recovery time before the shared endpoint. Four seed–condition pairs pass at both endpoints but fail at an intermediate evaluation. No condition yields simultaneous reliable readout from the encoder, entering residual and complete token tuple at the endpoint. Joint-training outcomes therefore mix changes in recovery with changes in rule formation.

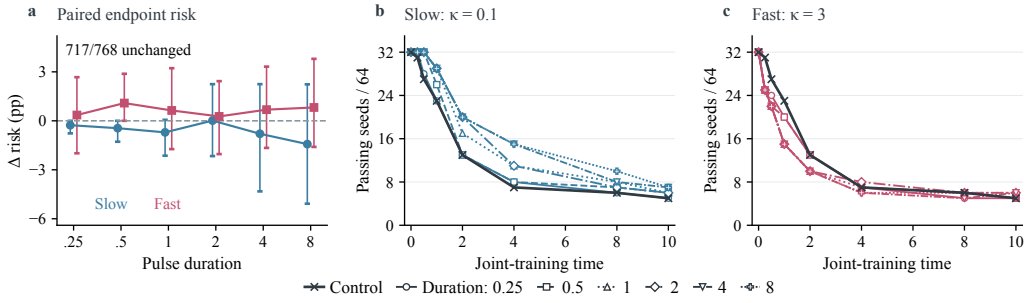


Figure S4: Joint-training responses (64 paired seeds). (a) Endpoint risk changes with 95% seed CIs. (b,c) Passing seeds under Slow/Fast pulses. Lines connect evaluation times after the parent checkpoint.

E.7 ECG MORPHOLOGY, PATIENT EVALUATION AND TRANSFER

Table S7 reports shared-rule reuse across patients and datasets, separating training-seed variation from patient-sampling uncertainty.

Table S7: ECG reliability ($B = 256$, 32 Control seeds). Risk: worst-context patient-macro mean [95% seed CI]. Bounds span seeds. Pass: upper risk ≤ 0.10 in every context.

Evaluation	Patients	Rule risk [95% CI]	Patient upper bound	Passes	Rule bits
PTB-XL	600	0.0355 [0.0354, 0.0358]	0.0810–0.0835	32/32	13
LUDB, frozen	40	0.0500 [0.0500, 0.0500]	0.5602	0/32	13
LUDB, adapted	40	0.0832 [0.0797, 0.0859]	0.5612–0.6154	0/32	13

Task and patient splits. The ECG experiment uses the public 12-lead datasets PTB-XL v1.0.3 and LUDB v1.0.1 (Wagner et al., 2022; Kalyakulina et al., 2021), distributed through PhysioNet (Wagner et al., 2020; Kalyakulina et al., 2020; Pollard et al., 2026). Inputs are morphology measurements from fixed one-second, 12-lead windows at 500 Hz. The binary target indicates whether lead-II QRS

duration exceeds 110.749 ms, the PTB-XL calibration median. Four contexts partition the QRS center's position within the unchanged window at 0.231637, 0.444497 and 0.766753 seconds. Windows retain their original duration and waveform timing. Evaluation requires complete QRS coverage, an available annotation and quality at most the PTB-XL calibration 95th percentile, 0.023184 mV. The target threshold, context boundaries and eligibility criteria remain fixed for LUDB. Representation training uses all training windows.

The analyzed PTB-XL cohort contains $4{,}800/600/600$ training/calibration/test patients, and the LUDB cohort contains $120/40/40$. Patient identifiers are disjoint across splits, with eight windows per patient. Evaluation retains $4{,}380/4{,}800$ PTB-XL test windows and $290/320$ LUDB test windows. The same patients are reused across seeds and arms.

Measurement interface and fixed rules. Acquired QRS width is centered and scaled by its mean and population standard deviation over all training windows in the fitting domain, including evaluation-ineligible windows. These statistics remain fixed during calibration and testing. Normalized width enters an affine encoder, two learned code vectors and a width decoder. Its distribution uses 64 support bins fitted on the domain's training windows. Each parent uses fixed positive, smoothed random bin weights shared across arms. Seeds vary initialization and bin weights on the same patient cohort. Fixed, training-fitted phase, amplitude and baseline quantizers complete a seven-bit $2 \times 4 \times 4 \times 4$ tuple. Training uses step size 0.01, $\alpha = 1$ and $\beta = 0.25$. Parents receive 200 updates to time 2, then each arm receives 200 more. Slow/Fast set κ to $0.1/10$ for the first 20 continuation steps, then restore it to one. Each fitting domain has 16 calibration and 32 independent confirmation seeds.

The model reconstructs acquired widths without class labels. Readout selection uses annotation durations, and test errors use reference durations, each thresholded as above. Calibration-patient risk selects a parent rule, frozen across arms and subsequent times. Candidates comprise two constants and a single-coordinate test in either polarity. A tuple test costs 13 bits, while a continuous-coordinate test costs 42, including a 32-bit threshold. The probe uses the same family and predicts identically to the selected $B = 256$ rule. Lengths are attained costs within this family. Fig. 7c covers both transitions. Functions, lengths and risks remain unchanged at seven larger tested budgets through $32{,}768$. Context and budget panels use 10,000 seed-bootstrap replicates.

Frozen transfer applies the PTB-XL model, normalization and rule to LUDB. Adaptation refits on LUDB training/calibration patients with the same measurement protocol. Both use the same LUDB test patients. Rate effects in Fig. 7d use paired seed CIs on these fixed patients; all six include zero.

Patient risk and upper bounds. Within context $C = n$, average window losses within patients, then weight patients equally. Let m_n count contributing patients, $\widehat{R}_n^{\mathrm{pat}}$ their mean loss and $\widehat{\sigma}_n^2$ their unbiased sample variance. For $G = |\mathcal{C}_{\mathrm{eval}}| = 4$ and $\gamma = 0.05$, the conservative empirical-Bernstein bound is

$$U_n^{\mathrm{pat}} = \min\left\{1, \widehat{R}_n^{\mathrm{pat}} + \sqrt{\frac{2\widehat{\sigma}_n^2 \log(3G/\gamma)}{m_n}} + \frac{3\log(3G/\gamma)}{m_n}\right\}, \qquad m_n \geq 5. \tag{47}$$

This exceeds the bound of Maurer & Pontil (2009), since $3/m_n \geq 7/[3(m_n - 1)]$ and $\log(3G/\gamma) > \log(2G/\gamma)$ for $m_n \geq 5$. Smaller contexts receive upper bound one, and absent contexts fail. Every evaluated context has at least 40 patients. The bound assumes independent patients within each context-specific population, allowing dependent windows within patients. Recovery requires both patient and window-level upper bounds at most 0.10 in every context, additionally correcting over distinct selected rules for length evaluation. Seed-bootstrap CIs describe training variation separately from patient uncertainty. On LUDB, the 13-bit rules have empirical risk below 0.10, while patient uncertainty keeps all tested budgets through $32{,}768$ bits above the reliability threshold.