AUDITING PLASTICITY DIAGNOSTICS WITH FUNCTION-PRESERVING RESCALINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can several plasticity diagnostics, read together, distinguish networks that implement the same function but respond differently to adaptation? We audit specified activation-, gradient- and spectral-based diagnostic formulas under function-preserving rescalings of ReLU networks. Every audited formula that normalises a per-unit signal by its layer mean, or reads normalised singular values, is exactly invariant in its ideal form to a layer-uniform rescaling; the normalisation, not the signal family, decides it—the absolute-cutoff feature rank reads the same singular values but can change. We identify this shared invariance across the formulas and, for specified full score vectors, characterise its exact extent within positive diagonal rescalings. On 40 rescaled pairs of 10 trained checkpoints (MinAtar Breakout DQN agents, ReLU MLP) the implemented scores agree within numerical precision except for a stabiliser-induced residual, and all audited unit-level threshold masks remain identical; yet the finite-budget adaptation to an action permutation differs under a protocol with SGD and a fixed rate-selection rule. For the *common* rescaling, evaluation on adaptation seeds unused for grid placement or rate selection yields a mean difference of $+0.317$ return units (95% checkpoint-cluster bootstrap interval $[0.228, 0.402]$), positive on every checkpoint. A separate analysis also finds a response gap after selecting a scalar learning rate independently for each arm on the tested finite grids. Under SGD the rescaling induces the known change in per-block effective learning rates; a corresponding layer-uniform effect is not detected in the Adam experiment, and a fixed-data, fixed-target experiment shows the optimisation difference without environment interaction. These formulas measure relative activity and representation dimension as designed; the limitation concerns their use to infer adaptation response: equality of the audited invariant readouts does not imply equal performance under the specified protocol.

1 INTRODUCTION

Diagnostics of plasticity—activation-based (the dormant-unit ratio), gradient-based (GraMa, the stagnant ratio) and spectral (effective ranks)—are computed from parameters and activations, monitored as proxies for the ability to adapt (Dohare et al., 2024; Lyle et al., 2022), and used to select units for resetting; libraries provide several side by side (Yuan et al., 2025), together with quantities such as the gradient norm that are not scale-normalised. That no single diagnostic suffices is known (Lyle et al., 2022); this paper asks, for one transformation that can be applied exactly, whether the diagnostics that share a within-layer normalisation jointly recover what each discards, and whether what they jointly discard matters for adaptation. The transformation is the function-preserving rescaling (1) of a ReLU network, which holds the realised function fixed to 1.2×10^{-15} while moving the parameter vector (Neyshabur et al., 2015; Lebeurrier et al., 2026) (Figure 1). Two established facts are premise and instrument, not claims: parameterisations agreeing on their outputs can adapt differently (Lyle et al., 2022; Lebeurrier et al., 2026), and rescaling changes optimisation dynamics across learning rates (Lebeurrier et al., 2026).

Where the information is lost. A diagnostic's value is the end of a chain—raw per-unit signal, normalisation within a layer, threshold, pooling—and equality can arise at any link. On the pairs our response experiment uses, every audited diagnostic that divides a per-unit signal by its layer mean, or one singular value by another, discards a layer's common scale at the normalisation link: the raw signals change by family-specific laws, the ideal normalised scores do not, the implemented scores

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

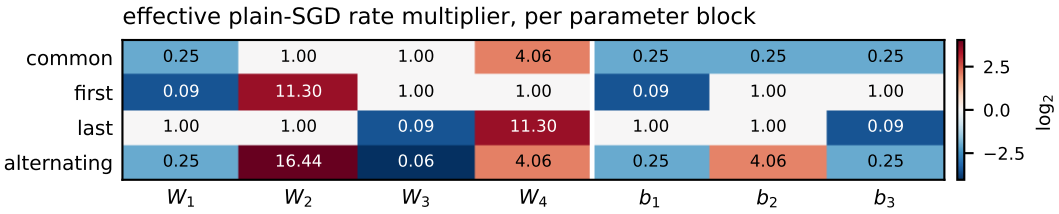

Figure 1: What one rescaling does to a plain-SGD step. A layer-uniform rescaling (Eq. 1) preserves the function and multiplies each of layer ℓ's raw signals by a factor common to the layer but specific to the signal (activations by c_ℓ, their gradients by $1/c_\ell$; Table 9), so whichever factor a signal carries cancels in the ideal normalisation and the audited scores, masks and fractions agree (algebraic, then audited: Section 4); that the finite-budget response changes is *empirical* (Section 5). Mapped back, the step's rate is multiplied by $(c_{\ell-1}/c_\ell)^2$ on W_ℓ and $1/c_\ell^2$ on b_ℓ—different, because a bias carries no incoming activation scale ($c_0 = c_{L+1} = 1$; cells are multipliers, colour their $\log_2$). Rows are Section 5's patterns from the experiment's own sampler, the closed form verified against realised gradients through the gauge code. Momentum transforms the buffer too; these are not Adam's.

agree with them under stated conditions, and the masks and fractions are unchanged. The blind spot is shared across the activation, gradient and spectral families because the normalisation rule, not the signal, decides it; the absolute-threshold feature rank, reading the same singular values with a fixed cutoff, does not share it. Any function of the ideal invariant vector is constant along that subgroup, so a response that varies there is unexplained by the collection, not merely by each member.

The paired comparison. On the same trained checkpoints the rescaling all of these diagnostics miss changes the held-out score of a fixed learning-rate selection protocol under SGD, and the change is re-measured on adaptation seeds never used to place the grid or select the rate (Figure 2). The object diagnosed is a finite-budget adaptation response under a fixed context (optimizer and its fresh state, rate-selection rule, task, budget); pairs exist on which the diagnostic vector does not determine it, along a direction that under SGD is a change of per-block effective learning rates. The invariance itself is not a defect (for relative activity or dimension it may be wanted); what it rules out is inferring an optimizer-dependent response from invariant readings.

Scope and contributions. (1) *An audit by stage* (Section 4): the formulas of Appendix B (five published diagnostics, one snapshot analogue, six controls of our own), pinned as formulas and implementations, audited link by link on the 90 representatives the response is measured on, with the conditions that qualify each equality stated.

(2) *A re-measured counterexample* (Section 5): within ten checkpoints a layer-uniform rescaling leaves the recorded diagnostics bit-identical while the protocol's held-out score differs, with intervals excluding zero; for the *common* rescaling the difference is re-observed on new adaptation seeds and remains after per-arm rate selection (post-hoc, finite grids).

(3) *Where it holds and fails* (Section 6, Figure 3): the *common* arm differs from the exact per-block-rate control (float64); the effect is not detected under Adam and its direction does not carry to a convolutional network; a rescaling comparable in size to the change ordinary training produces in one blind coordinate changes the response in either direction, and an optimisation difference appears on fixed data with no online feedback.

(4) *A practical limit* (Section 7): no function of the invariant diagnostics equals the score on these pairs (Propositions 2 and 3); keeping each normaliser with its signal preserves what normalisation removed without making it predictive; with the predictors tested, adding diagnostics did not beat the intervention's identity alone. What is shown is a non-identifiability of one protocol's response from the diagnostics, not that they fail to measure an intrinsic plasticity; we claim no new diagnostic, no repair, and nothing about ordinary training.

2 Related Work

Three axes; the closest works, by what each established, what is added and what stays open: Table 12 (Appendix B). *Rescaling and optimisation.* That parameterisation affects training dynamics

is established: rescaling-invariant updates (Neyshabur et al., 2015), equi-normalisation (Stock et al., 2019), teleportation (Armenta et al., 2020; Zhao et al., 2022), the symmetry's inductive bias (Phuong & Lampert, 2020), and path-conditioned rescaling built on the observation that rescaled parameterisations of one function train differently across learning rates (Lebeurrier et al., 2026). Relative layer scales set effective per-layer rates and the degree of feature learning (Kunin et al., 2024); Lyle et al. (2024) tie parameter norm and normalisation to the effective learning rate in RL; Baveja et al. (2025) derive a per-layer effective step-size bound and find no individual indicator reliably predicts loss of trainability. Our per-block reading (Section 5) is an instance of that mechanism, not a new one. Closest in object, Wang (2026b) (T. Wang, v2) fix a two-layer teacher–student network's initial function, vary a *per-neuron* gauge and prove which duplicate neuron acquires a feature under gradient flow and small-step gradient descent (no scalar-rate or clock matching): a dynamics theorem at initialisation with no diagnostic read; its gauges are neuron-wise rather than in H_{unif}, and its no-clock theorem is a continuous-time analogue of the scalar-recalibration question left open here on a finite grid.

Diagnostics and plasticity. Lyle et al. (2022) report feature rank as correlated with, and explicitly not sufficient for, learning progress; plasticity loss has several causes and occurs without saturated units (Lyle et al., 2023; 2025; Lewandowski et al., 2023; Abbas et al., 2023; Ma et al., 2024); all analyse mechanisms across training states, whereas we compare representatives of one state. Wang (2026a) (Q. Wang) formalise states indistinguishable under present-behaviour readouts that respond differently to future training; ours are gauge orbits, audited formulas and RL checkpoints. Wang et al. (2026) (J. Wang et al.) construct counterexamples to rank diagnostics and derive an optimisation-readiness score; their compared states share the initial function but are not gauge images of one another, and the diagnostic differs between them, whereas our question is whether a collection stays unchanged along one gauge orbit while the response changes—we propose no replacement. Nikishin et al. (2023) preserve predictions while changing what can be learned next; ours changes no trainable structure. Reparameterisation as an audit instrument follows Dinh et al. (2017), who move sharpness with the function fixed.

Metric definitions (sources in Table 6): the dormant ratio (Sokar et al., 2023; Xu et al., 2024) and its gradient-based successors (Liu et al., 2025; 2026) normalise a per-unit signal by its layer; the spectral ranks (Kumar et al., 2021; Roy & Vetterli, 2007) and the absolute-threshold feature rank (Lyle et al., 2022) read singular values; libraries aggregate them (Yuan et al., 2025); He (2026) question how much such descriptive diagnostics explain of optimisation dynamics. The audit concerns the specified formulas' measurement properties and neither reproduces nor refutes the reset or injection methods proposed alongside them; the stagnant-neuron row is a snapshot analogue and the GraMa reference-code row a transcribed formula, not the upstream package run.

3 SETUP

The gauge. For a ReLU network write $C_\ell = \mathrm{diag}(c_\ell)$ with every $c_{\ell,i} > 0$ and $C_0 = C_L = I$. Then

$$W_\ell \mapsto C_\ell W_\ell C_{\ell-1}^{-1}, \qquad b_\ell \mapsto C_\ell b_\ell \tag{1}$$

leaves every output exactly unchanged, since $\mathrm{ReLU}(Cz) = C\,\mathrm{ReLU}(z)$ for positive diagonal C. The rescalings form a group G; $H_{\mathrm{unif}} \subset G$ scales all units of a layer by one factor.

Proposition 1 (Conditional invariance)**.** *Let D be homogeneous of degree zero under a common positive scaling of the units of any single layer. Then $D(h\theta) = D(\theta)$ for all $h \in H_{\mathrm{unif}}$ and all θ.*

The proof is one line (Appendix C); what the proposition does not say is which diagnostics satisfy its hypothesis, or whether an implementation with a clamp or an ε realises it, and that is what the audit checks. Within G the converse holds at the score level.

Proposition 2 (Exact score-level blind subgroup within G; proof in Appendix C)**.** *Write $\mathrm{Blind}_G(D) = \{g \in G : D(g\theta) = D(\theta)\ \forall\theta\}$ and c_ℓ for g's factor on hidden layer ℓ. (i) For a per-unit signal $x_\ell \geq 0$ with $x_{\ell i}(g\theta) = c_{\ell i}^{p} x_{\ell i}(\theta)$, $p \neq 0$ (activations $p = 1$, activation gradients $p = -1$), attaining $x_{\ell i}, x_{\ell j} > 0$ for every pair, $s_\ell = x_\ell / \mathrm{mean}_k x_{\ell k}$ has $\mathrm{Blind}_G(s_\ell) = \{g : c_\ell$ constant$\}$, the other layers unconstrained. (ii) For $D = (s_1, \ldots, s_{L-1})$, $\mathrm{Blind}_G(D) = H_{\mathrm{unif}}$, and this persists when components whose blind sets contain H_{unif} are added. (iii) The last hidden layer's normalised singular-value spectrum has $\mathrm{Blind}_G = \{g : c_{L-1}$ constant$\}$ whenever a feature matrix with orthonormal columns is realisable.*

So no collection of such scores is blind to less than H_{unif}, and adding one per-layer normaliser restores what normalisation removed (not the rest of what an adaptation prediction needs). The classification concerns global symmetries of full score vectors for unit-equivariant signals; pooled or thresholded summaries, and signals mixing neighbouring layers' factors, may fail to distinguish more (Appendix C).

3.1 THE RESPONSE ESTIMAND

Definition 1 (LOO-selected grid performance)**.** Write $A(\eta, k; \theta)$ for the time-normalised area under the return curve of a 4,000-environment-step adaptation from θ at rate η with adaptation seed k: the mean over 20 equally spaced evaluations of the trailing-10-episode return of the *training* episodes themselves, exploration included, in return units (Section 5), and $\bar{A}(\eta; \theta) = \frac{1}{K} \sum_k A(\eta, k; \theta)$. Fix a grid $\mathcal{G}$ of 15 rates and a set S of training seeds. For a checkpoint θ_s the rate is selected on the *other* checkpoints of the same condition and evaluated on s:

$$\eta^{-s} = \arg\max_{\eta \in \mathcal{G}} \frac{1}{|S| - 1} \sum_{s' \neq s} \bar{A}(\eta; \theta_{s'}), \qquad P^{\text{loo}}(\theta_s) = \bar{A}(\eta^{-s}; \theta_s). \tag{2}$$

P^{loo} is not a peak but the held-out performance of a rate-selection rule fitted on the other checkpoints; a contrast in it mixes best-in-grid performance with how well a rate chosen elsewhere transfers (separated in Section 5). The estimand T is the protocol's expected output over the adaptation randomness at a fixed checkpoint, with K, the grid and the seed coupling part of the definition, conditioned on the realised selection data; one interval holds the selection fixed, a second re-runs it (band width and location: Definition 3, Table 13).

3.2 METRIC CONVENTIONS THAT DECIDE WHETHER INVARIANCE APPLIES

Appendix B states each audited formula as computed (its normalisation, threshold, bias convention, any ε, and whether it is the paper-defined formula, a library implementation or our analogue). Every result is read on three levels: *ideal invariance* of a formula (Proposition 1), the *condition* under which an implementation realises it, and *observed agreement* on the comparisons evaluated. Diagnostics are read on the checkpoint's current-task probe while the response is measured on the next task, as a diagnostic is read before adapting. Two within-layer normalisations carry a 10^{-12} clamp, verified inactive on every representative used (smallest layer means 7.8×10^{-2} and 3.0×10^{-4}); the update-activity ratio is a snapshot analogue of a training-window average, exactly invariant only at $\varepsilon = 0$, and a snapshot equality does not extend to the published temporal quantity; the rank rows are numerical ranks at a tolerance; GraMa is audited in its printed and its reference-code reading (Table 1).

4 THE AUDIT, BY STAGE

Within-layer-normalised scores satisfy Proposition 1's hypothesis because a common factor multiplies numerator and denominator alike; a ratio of singular values does because $\Phi \mapsto c\,\Phi$ scales every singular value by c; an absolute cutoff on those singular values, or a norm with no normalisation, does not. On gauge draws around one trained network the partition follows that rule exactly (Appendix B, Table 7; it replicates on a convolutional network, feature rank 1.283). Table 1 asks the sharper question on the pairs the response is measured on: *at which link* does each equality arise.

Layer-uniform pairs: lost at normalisation. Every layer-uniform pattern multiplies each hidden layer's activation matrix by one factor or leaves it unchanged (*common*: all three layers by $e^{0.7}$; *first*, *last*: one layer by $e^{0.7\sqrt{3}}$; *alternating*: $e^{\pm 0.7}$; Table 10). The raw per-unit signals change accordingly—activation means by up to $3.4\times$, gradient magnitudes by up to $4\times$, update activity by up to $16\times$, each by one common factor within a layer on all 40 pairs—by laws that differ from signal to signal (Table 9). For the ideal ratio-normalised formulas that factor cancels exactly; the implemented scores agree under the stated clamp conditions (largest change of any unit's score 2×10^{-15}), and the one additive stabiliser, the reference-code GraMa's $+10^{-9}$, leaves a residual 5×10^{-6} (its form: Appendix B). Every residual stays below the smallest distance of any unit's score to its threshold (10^{-4} and above, Table 11), a sufficient condition for unchanged masks, and they are

Table 1: Three states kept apart, on the 90 representatives of Section 5: whether the *ideal* formula ($\varepsilon = 0$, no clamp) is invariant on H_{unif}; the largest change of any unit's *implemented* score across a layer-uniform pair; and on how many pairs the reported *decisions* (mask unit for unit / pooled fraction, or the integer value) are equal, on the layer-uniform and on the anisotropic pairs. GraMa at the audit's $\tau = 0.25$; at the source's $\tau = 0.01$ the printed-equation reading is 0 on every identity checkpoint (both thresholds: Table 8). Thresholds, clamps, provenance and the *first*-pattern exclusion for the spectral rows: Appendix B.

			layer-uniform pairs		anisotropic
diagnostic	normalisation	ideal	score $\Delta_{\max}$	decisions =	decisions =
unit-based, 40 + 40 pairs; decisions: mask (unit for unit) / pooled fraction					
dormant ratio	activation / mean	yes	10^{-15}	40/40 / 40/40	2/40 / 3/40
GraMa, Eq. 2 reading	activation grad. / mean	yes	10^{-15}	40/40 / 40/40	0/40 / 0/40
GraMa, reference code	weight grad. / mean $+10^{-9}$	yes	5×10^{-6}	40/40 / 40/40	0/40 / 0/40
stagnant ratio (ε=0 analogue)	update activity / mean	yes	10^{-14}	40/40 / 40/40	0/40 / 0/40
spectral, last hidden layer (30 layer-uniform pairs rescale it; 40 anisotropic); decision: reported value					
cumulative spectral rank	singular-mass ratio	yes	0	30/30	0/40
entropy effective rank	singular-value ratios	yes	10^{-15}	continuous	continuous
feature rank	$\Phi/\sqrt{n}$, **absolute** cut	no	—	20/30	12/40
gradient norm, all parameters	none	no	0.58 (rel.)	0/40	0/40

unchanged: identical unit for unit on 40/40 pairs, fractions equal, no unit on a boundary. Exact score invariance (the ideal formulas) is thus distinguished from agreement of the reported decisions (the implementations). On the 30 pairs that rescale the last layer the ratio-based ranks are equal on all 30 and the entropy rank to 10^{-15}, while the absolute-threshold feature rank moved on 10—a singular value crossing a fixed cutoff, not a change of the spectrum's shape. The raw signals follow different laws (activations and activation gradients scale in opposite directions; weight-gradient signals also depend on neighbouring layers), and the shared blindness arises only after normalisation removes each signal's common layerwise factor: diversity of the signals does not by itself give sensitivity to these directions.

Three conditions. A rescaling confined to one layer does not reach tensors read in other layers (the *first* pairs, the spectral rows); at the source's $\tau = 0.01$ the printed GraMa reading is 0 on every identity checkpoint, a floor (Table 1 uses 0.25; both in Table 8); and the reference code's continuous score is not exactly invariant, only its thresholded fraction was equal.

Anisotropic pairs: differences survive every link. Under a per-unit rescaling the scores change by order one and the masks flip on 7% (dormant ratio) to 52% (stagnant ratio) of the units at the median pair; an equal pooled fraction with a different mask occurred on one of the 40 pairs—an existence case for the aggregation link, not evidence that pooling commonly erases unit information. Blindness outside G is not characterised.

5 The counterexample, re-measured on independent seeds

Setup. Checkpoints are MinAtar (Young & Tian, 2019) Breakout DQN agents (ReLU MLP, three hidden layers of 64) trained for 300,000 steps with Adam under an action-permutation switch every 50,000 steps; one final snapshot per seed. Adaptation starts from the representative on the next permutation with a fresh replay buffer and optimizer state (SGD, momentum 0.9; Adam in Section 6.1)—a restarted optimizer, not a continuation of the training one—for 4,000 steps after 1,000 warm-up steps, over 15 rates in factors of two from 1.95×10^{-6} to 3.2×10^{-2} and 6 adaptation seeds (Appendix D). Each of 10 checkpoints yields nine representatives—identity, four layer-uniform patterns at log-magnitude 0.7, four per-unit anisotropic draws (0.7 and 1.8, two directions)—each run through Definition 1. The training seed is the only independent unit; contrasts are formed *within* a checkpoint against its identity, with the pre-registered interquartile mean (Agarwal et al., 2021) and cluster-bootstrap intervals over the seeds.

Same diagnostic value, different response. Diagnostic equality is exact per pair; the response difference is an estimate. On the 40 layer-uniform pairs the two recorded diagnostics are bit-identical to the identity's, and every other invariant diagnostic is checked on the same pairs (Section 4); the

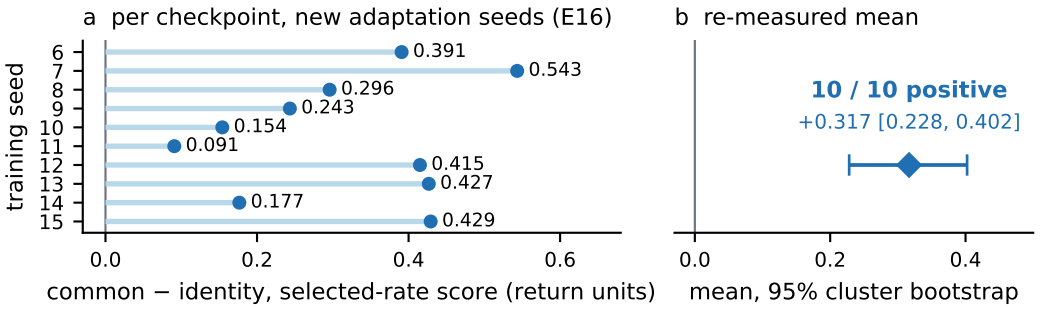

Figure 2: Independent-adaptation-seed re-measurement (E16): the *common* rescaling changes the selected-rate score while the audited invariant diagnostics agree. **(a)** *common* − identity per checkpoint; each arm's rate selected on adaptation seeds 0–5 on a fixed 17-rate grid, its score read on seeds 6–11, unused for placement or selection (no per-checkpoint interval is reported). **(b)** The arithmetic mean, 95% checkpoint-cluster bootstrap. E16 re-measures the comparison on the same ten trained checkpoints under new adaptation randomness; it is distinct from the pre-registered E9 family primary (Table 2), not a replication on new checkpoints or a new task.

pre-registered primary (E9) is the family contrast in the selected-rate level, $+0.230$ $[0.161, 0.299]$ return units (interquartile mean; arithmetic mean $+0.2359$; Table 2; the identity's selected score averages 2.12)—an aggregate effect, 3 of the 40 contrasts being negative (Figure 5, Appendix D); re-measured on new adaptation seeds, the *common* pattern's contrast is Figure 2. Across conditions (Table 16, Appendix E) the level moves in both directions: up under every layer-uniform pattern, down by -0.683 at the strongest anisotropic point, where the diagnostics move too.

Re-measured on independent seeds. The online selection-protocol contrasts are listed in Table 2 with what selects the rate and what scores it; the pre-registered primary is the family contrast above, the rest are analyses of the *common* pattern, each a different estimand, and no two are compared for size. Within-checkpoint selection and grid refinement around each arm's maximum leave the contrast, but the refinement positions were chosen with the evaluation seeds, so a final collection separated placement, selection and evaluation, its rules fixed before any row existed: the 17-rate grids *fixed*, the rate selected on the original six adaptation seeds alone, the score read on six adaptation seeds never used before, on the same ten checkpoints. The *common* − identity contrast is $+0.317$ $[0.228, 0.402]$ (interquartile mean $+0.325$ $[0.204, 0.420]$), positive on every checkpoint ($+0.091$ to $+0.543$; post-hoc exact sign and sign-flip tests, two-sided, p =0.002, a bootstrap-free directional check), "retained" under the pre-fixed rule—a re-verification on the same trained checkpoints under new adaptation randomness, not on new checkpoints or a new task, with no agreement of size claimed. **Net of evaluation noise** (retrospective, rates held fixed; Appendix D): with each checkpoint's noise variance estimated from its six per-seed paired differences, the noise-corrected response difference between these two representatives with identical diagnostics is 0.326 $[0.232, 0.404]$ return units (RMS over checkpoints; noise sd 0.111; an estimated 11.4% of the observed mean square is evaluation noise). Close to the mean contrast, it supports that the positive contrast is not evaluation noise; it is not a finding of heterogeneity across checkpoints, nor about natural states.

The contrast persists under greedy evaluation (post-hoc; Appendix D): five greedy episodes on a separate environment copy at each evaluation point give $+0.519$ $[0.405, 0.635]$ at the online-selected rate and $+0.495$ $[0.362, 0.630]$ at the greedy-selected one, positive on all 10 checkpoints—not confined to the exploration-inclusive online return.

What the rescaling does to the optimizer, exactly. Under plain SGD the rescaling is a change of coordinates, and mapping the update back to the original coordinates gives SGD with a per-block learning rate: for $\theta' = A\theta$, $\theta_{t+1} = \theta_t - \eta A^{-2}\nabla_\theta \mathcal{L}$: for *common* ($c = e^{0.7}$) $\approx 0.25\,\eta$ on the first layer's weights and the hidden biases, η on the hidden-to-hidden weights, $\approx 4\,\eta$ on the output weights (Appendix E). The response difference is thus a known kind of effect—a layerwise preconditioning that plain SGD does not undo and Adam's normalisation halves in log scale (Kunin et al., 2024; Lyle et al., 2024)—and the finding is that the audited diagnostics carry none of it. One scalar rate per arm does not undo it on the tested grids: selecting each arm's rate on three new seeds

contrast	rate selected on	scored on	status	value
layer-uniform family − identity	the other nine checkpoints, seeds 0–5 (LOO)	the held-out checkpoint, seeds 0–5	**pre-registered primary** (E9): the cross-checkpoint tuning-transfer score; IQM, selection fixed	$+0.230$ $[0.161, 0.299]$
the same	the same	the same	pre-registered mean, selection re-run in every resample	$+0.2359$ $[0.1707, 0.3173]$
common − identity	the same	the same	the E9 row for one pattern (mean)	$+0.351$ $[0.255, 0.449]$
common − identity	three of seeds 0–5, same checkpoint	the other three	post-hoc; 15 rates, then 17 (E13, rules fixed first)	$+0.376$ $[0.277, 0.468]$; $+0.359$ $[0.251, 0.476]$
common − identity	seeds 0–5, fixed 17-rate grid	seeds 6–11, never used before	**pre-specified re-measurement** (E16): the within-checkpoint selected score on new seeds; the abstract's number	$+0.317$ $[0.228, 0.402]$
common − identity	three of seeds 6–11, per arm	the other three	post-hoc (E19)	$+0.379$ $[0.324, 0.429]$

Table 2: The online selection-protocol level contrasts: the pre-registered family primary (E9) and the *common*-pattern analyses (return units, 10 checkpoints, 95% cluster-bootstrap intervals, arithmetic mean unless stated): what is compared, what selects the rate, what scores it, and its status; different estimands, not compared for size. The E21, E22, E23 and E25 contrasts are in Appendix D; widths, locations and curves in Table 16 and Figure 7.

and scoring on the other three leaves $+0.379$ $[0.324, 0.429]$ (post-hoc), which does not exclude a continuous scalar shift (the exact per-block control: Section 6.3).

6 Where it holds, and where it does not

6.1 A second optimizer: a research question

The same nine representatives under Adam change the adaptation optimizer and nothing else (Figure 3): the layer-uniform level effect is not detected ($+0.012$ $[-0.064, 0.051]$), the strongest anisotropic point lowers the level by -0.140, and the direct paired contrast $\Delta_{\mathrm{SGD}} - \Delta_{\mathrm{Adam}}$ is $+0.234$ $[+0.168, +0.310]$, all four patterns agreeing in sign. Adam's per-parameter scaling is a partial gauge correction (Appendix E); whether the difference is SGD-specific or a property of the representatives that Adam masks is open.

6.2 A second architecture (exploratory per-pattern evidence): the level's direction does not carry

The audit holds on a trained convolutional network (Appendix B). The response experiment was repeated on ten convolutional Breakout checkpoints (channel-uniform and per-channel rescalings, SGD, a 20-rate grid; pre-registered, stopped once and resumed as a continuation, D7; read once; Table 18). On all 40 channel-uniform pairs the recorded diagnostics are again bit-identical. The pre-registered *scalar-degradation* target (level down *and* band narrower together; thresholds 0.10 return units and 0.15 dex) is refuted: *first* lowers the level (-0.173 $[-0.248, -0.096]$) while the band widens and its location shifts, with no joint signature. But the MLP's positive level effect does *not* reproduce: the family (-0.039 $[-0.152, 0.097]$) and *common* ($+0.083$ $[-0.024, 0.258]$) contrasts are inconclusive (Figure 3); per-pattern readings are four uncorrected tests, no architecture interaction is tested and the grids differ.

6.3 The matched-optimizer control

A control that applies the rescaling's own elementwise rate to the momentum-buffered update commutes with (1) exactly, so on a rescaled representative it reproduces the identity's plain-SGD trajectory in the rescaled coordinates. Its float32 fidelity gate failed (copies diverged by 11% of the identity's range; Appendix E); re-collected in float64 with its scalar arms it passed (copies agree to 0.6%) and reproduced the identity level within the equivalence margin (offset -0.005). The *common* arm is $+0.162$ of the identity's range above the control (95% $[0.103, 0.212]$) and the contrast remained positive in float64 ($+0.346$ $[0.248, 0.435]$; float32 $+0.351$ $[0.255, 0.449]$): a per-block

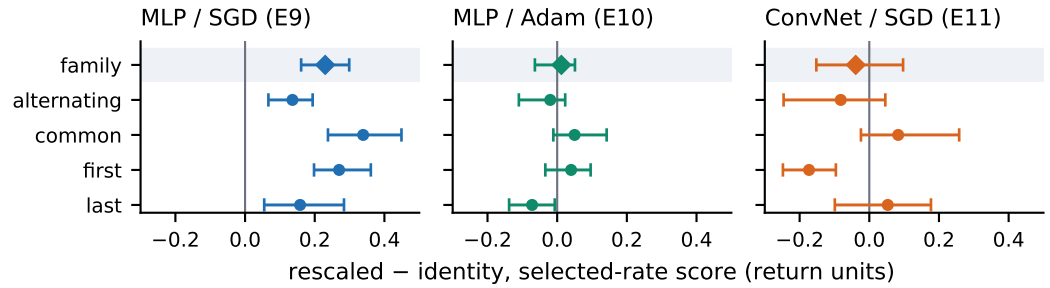

Figure 3: Optimizer and architecture scope: interquartile means of the paired selected-rate contrasts for the layer-uniform family (diamond) and its patterns, 95% checkpoint-cluster bootstrap intervals (Tables 16, 17, 18; ten checkpoints per setting, the ConvNet on a different grid). Per-pattern intervals are uncorrected, no architecture interaction is tested, and an interval crossing zero is not equivalence.

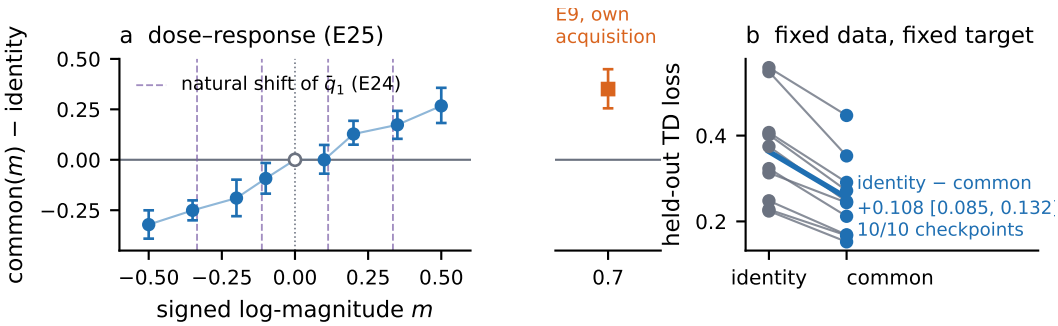

Figure 4: Two separate questions. **(a)** Intervention size (E25): *common*(m) − identity in the selected-rate score (return units; positive favours *common*) against m, ten checkpoints, 95% checkpoint-cluster bootstrap intervals, both arms of every contrast in one acquisition; hollow point: $m = 0$ by definition; no curve fitted. E9's $m = 0.7$ row is its own acquisition, in its own panel. Dashed: $\pm$ the natural shift of the first hidden layer's mean of q over a run (0.33) and its second half (0.11; Appendix F), a size reference for one coordinate, not a direction or frequency. **(b)** Fixed data, fixed objective (E26): each checkpoint's mean held-out TD loss over 3,000 SGD updates on identical random-policy transitions and minibatches, target frozen at the shared initial function, at the rate selected on the other checkpoints; paired difference identity − common, 95% interval. TD loss is not policy return; the panels share no axis.

conditioning change, not a float32-only discrepancy. The control covers *common* only and does not resolve the continuous scalar recalibration question.

6.4 Natural-scale magnitudes, and adaptation on fixed data

Two post-hoc collections, rules fixed before any row existed (Appendix D). **Dose–response.** The *common* rescaling at $m \in \{\pm 0.1, \pm 0.2, \pm 0.35, \pm 0.5\}$ on the ten checkpoints with E9's grid and seeds, against the identity re-collected in the same acquisition: the pre-fixed cross-environment gate failed (float32; Appendix D), so every E25 contrast is formed within one acquisition and E9's $m = 0.7$ row is quoted within its own. For scale, the rescaling at m shifts the first hidden layer's mean of q, one blind coordinate, by $-m$, against a natural shift of 0.33 over a run and at most 0.11 over its second half (Figure 4a; a size reference only). At $m = +0.35$ the contrast is $+0.173$ $[0.104, 0.242]$ (9 of 10 checkpoints positive) and at -0.35 -0.251 $[-0.300, -0.201]$: an intervention comparable in size to the observed change of that coordinate changes the response detectably, in either direction. At $|m| = 0.1$ the negative direction is detected (-0.093 $[-0.168, -0.016]$) and the positive is not (-0.000 $[-0.069, 0.074]$; no equivalence margin was registered). At the evaluated negative magnitudes the mean contrast is negative and more negative at larger $|m|$ (Figure 4a); the eight point estimates increase with m except between 0 and $+0.1$ (no curve is fitted). **Fixed data.** On identical random-policy transition sets and minibatch sequences with fresh optimizer state, 3,000 SGD updates at each rate of E9's grid with the target frozen at the shared initial function (one ob-

jective for both arms), the rate selected on the other checkpoints (Figure 4b): the identity's held-out loss is 0.363 and *common*'s 0.255, a paired difference of $+0.108$ $[0.085, 0.132]$, lower for *common* on all 10 checkpoints (final loss $+0.060$ $[0.041, 0.081]$; Figure 4b); with per-arm targets synced as in the online protocol, $+0.123$ $[0.095, 0.154]$. The fixed-data experiment thus shows an optimisation difference without online data feedback, consistent with the online difference but not its pathway (a TD objective on random-policy data, not policy quality). **An initial first-order score has limited predictive power.** With the rescaling read as per-block rates (Appendix E), the first-order approximation $\eta \sum_b p_b \|g_b\|^2$ to the decrease of the training objective at θ_0, tested as a proxy for these responses at each arm's selected rate (retrospective; Appendix D), does not beat the constant predictor on the fixed-data record (sign agreement 9/10 against 10/10; Spearman 0.09 $[-0.57, 0.72]$) and matches the sign of the dose–response point estimates at 1/8 of the eight magnitudes (descriptive: the points share checkpoints and one is near zero). A predictor's failure does not exclude the conditioning reading.

7 What this changes about reading a diagnostic

Proposition 3 (Stabilizer–response mismatch)**.** *Let $T(\cdot, O)$ be an adaptation estimand with fixed protocol context O (Definition 1) and $H \leq G$ a subgroup with (i) $D(h\theta) = D(\theta)$ for every $h \in H$ and (ii) $T(h_0\theta, O) \neq T(\theta, O)$ for some $h_0 \in H$. Then no φ satisfies $T(\cdot, O) = \varphi \circ D$ on the orbit $H\theta$.*

The proof is two lines (Appendix C); (i) is the bit-identical equality above, (ii) is supported by an interval excluding zero—statistical evidence, not a proof. D may be a vector; for nearly equal continuous scores a bound needs a Lipschitz restriction on the predictor (Corollary 2; Appendix D).

On this subgroup, combining diagnostics that share a scale-normalising invariance does not recover the discarded scale. Keeping the layer mean μ_ℓ with the score vector does ($x_{\ell i} = \mu_\ell s_{\ell i}$; Appendix B). Equal diagnostics do not license equal protocol performance, and responding to a rescaling is not predicting adaptation: a formula's design (an absolute cutoff against a ratio) decides its blindness, not the signal it reads. **Lower prediction error is not a better per-checkpoint choice, and with the predictors tested, adding diagnostics did not improve prediction over the intervention's identity alone** (post-hoc): a predictor given the pooled scale descriptor leaves a regret of 0.032 $[0.007, 0.065]$ among a held-out checkpoint's nine representatives, the same as the fixed rule "always choose *common*" (Table 14). A retrospective baseline names the source (Table 19, Figure 9): the intervention's identity alone reaches the lowest held-out MAE of every feature set; adding every audited diagnostic's value raises it, and adding the per-layer scale to both lowers the ridge MAE by 0.009 with an interval whose upper endpoint is 4×10^{-5} below zero (borderline; no improvement detected for k-NN; regret unchanged). Three questions stay apart: keeping the normaliser with its signal restores the discarded scale (an identity); whether that improves prediction (a limited ridge improvement that still does not reach the identity alone; nothing for k-NN); whether it improves the choice (no: the fixed rule matches); a worse fit with more features is not by itself an absence of information. **Tuning the rate separately for two states a diagnostic cannot tell apart did not measurably beat one shared rate**: $+0.004$ $[-0.016, 0.027]$ on the 15 rates both share, though they select different rates in two thirds of held-out splits—an advantage not established, not a cost shown absent (E30, post-hoc; Appendix D).

8 Limitations and conclusion

Breadth (Table 3, Appendix A). The audit covers Table 6's formulas on two architectures; the response rests on one task and one perturbation family under two optimizers, as a finite-horizon response above a pre-update reference, with 10 seeds; E25 and E26 cover one pattern, E26 on random-policy data. **Unresolved:** whether the difference is SGD-specific; a continuous scalar recalibration; blindness outside the diagonal group; how often this arises in ordinary training and with what consequence (Appendix F). No benchmark result is claimed.

Conclusion. The audited ratio-normalised diagnostics lose a layer's common scale at normalisation; on the same checkpoints that scale changes a fixed selection protocol's held-out score under SGD: a protocol-relative non-identifiability, not a degradation law.

AI USE STATEMENT

We used generative AI tools (Claude, Anthropic) to assist with research and manuscript preparation under author supervision. Assistance included drafting and revising experimental protocols, mathematical statements and proofs, implementation and analysis code, manuscript text, figures, and related-work comparisons, as well as supporting experiment execution and monitoring. The authors determined which experiments to conduct, checked the proofs, and reviewed and revised the manuscript and its interpretations.

Confirmatory, exploratory and post-hoc analyses are distinguished in the paper, with protocol changes and deviations documented in the appendices. Reported results are derived from recorded experiments, and automated checks verify specified manuscript quantities against those records. These checks do not constitute independent verification of every AI-assisted artifact. The authors retain responsibility for all scientific decisions, claims and final content.

ETHICS STATEMENT

This work analyses the measurement properties of existing diagnostics on a small-scale reinforcement-learning benchmark (MinAtar Breakout). It involves no human subjects, no personal data and no new dataset release; all checkpoints, adaptation records and analysis outputs were produced by the authors' own runs of publicly available software. We identify no application of these results that raises concerns under the Code of Ethics beyond those already present in the literature it audits.

REPRODUCIBILITY STATEMENT

Every quantity reported here is produced from a recorded collection by an analysis whose status—specified in advance, a subsequent correction, or exploratory—is identified in the deviation record rather than asserted in aggregate. The estimand definitions are in Section 3.1 and Appendix D, the metric definitions in Appendix B, and the proofs in Section 7 and Appendix C. Each confirmatory collection was frozen before it ran: the grid, the adaptation budget and seeds, and a *source digest* of the runner and estimator files are written to a freeze record the launcher refuses to collect without (E9 `a5bdf9d8d7f09d4c`, E10 `803ad2d22a9a9290`; the convolutional collection, stopped and resumed as a continuation, carried `d3b47ac2d06bf7d1`), and every row separately carries an *acquisition stamp* hashing the acquisition arguments it was collected under (E9 `53ec72e822ece6c6`, E10 `c37b1d05ffba2bdc`, E13 `2e4d003dfbe91ad5`; the later collections' stamps are in Appendix D). Re-running E9's identity arm in the present environment reproduces 48/90 of its per-seed values exactly and diverges elsewhere by float32 amplification (Appendix D); the frozen records are the reported ones. The E9 input file's sha256 begins `40349a24...` (the full digest is in `paper/figures/manifest.json` and the freeze record). The typeset figures are of two kinds: packages built from the reported values (`paper/figures/pkg/` and `pkg2/`, each with its provenance file and the CSV of values it draws; `scripts/check_figure_pkg2.py` checks those values against the macros, the generated tables and the read records) and record-generated figures (`scripts/e9_manuscript_figures.py`, compared to disk at every build; `scripts/e25_e26_figure.py`). Results were read in a pre-registered order—data quality, level, bandwidth, location, sensitivity, oracle comparison, diagnostic association—with each section's output recorded before the next was opened. Tables 16, 17 and 18 are generated from those recorded outputs; the effect sizes, intervals, bound figures and audit values quoted in prose are defined once in a single macro file with their provenance, and the test suite checks the layer-uniform level and bandwidth effects, their intervals and the two bound figures against the frozen records. Anonymous code, frozen protocol records and analysis artifacts are provided in the supplementary material.

REFERENCES

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents*, 2023.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2108.13264.

Marco Armenta, Thierry Judge, Nathan Painchaud, Youssef Skandarani, Carl Lemaire, Gabriel Gibeau Sanchez, Philippe Spino, and Pierre-Marc Jodoin. Neural teleportation, 2020. arXiv:2012.01118.

Gunbir Singh Baveja, Alex Lewandowski, and Mark Schmidt. A unified noise-curvature view of loss of trainability. In *NeurIPS 2025 Workshop on Optimization for Machine Learning*, 2025. arXiv:2509.19698 (the text consulted).

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017. arXiv:1703.04933.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Zhiqiang He. Rethinking plasticity in deep reinforcement learning, 2026. arXiv:2603.21173, preprint.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021. arXiv:2010.14498.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, 2024.

Arthur Lebeurrier, Titouan Vayer, and Rémi Gribonval. Path-conditioned training: A principled way to rescale ReLU neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2602.19799.

Alex Lewandowski, Haruto Tanaka, Dale Schuurmans, and Marlos C. Machado. Directions of curvature as an explanation for loss of plasticity. *arXiv preprint arXiv:2312.00246*, 2023. first posted 2023; version 4, October 2024, is the text consulted.

Jiashun Liu, Zihao Wu, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Ling Pan. Measure gradients, not activations! enhancing neuronal activity in deep reinforcement learning. In *Advances in Neural Information Processing Systems 38*, 2025. arXiv:2505.24061; the equation audited in this paper is Eq. 2 of the arXiv v2 text (2026).

Zhengzhu Liu, Zeming Gao, Haoyuan Qin, Jiawei Hu, Junhao Wu, Miao Zhu, Haipeng Zhang, Chennan Ma, Siqi Shen, and Cheng Wang. Stagnant neuron: Towards understanding the plasticity loss in multi-agent reinforcement learning value factorization methods, 2026. arXiv:2606.25335.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, 2023. arXiv:2303.01486.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado van Hasselt, Razvan Pascanu, and Will Dabney. Normalization and effective learning rates in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2407.01800.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. In *Proceedings of the 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 750–783, 2025.

Guozheng Ma, Lu Li, Sen Zhang, Zixuan Liu, Zhen Wang, Yixin Chen, Li Shen, Xueqian Wang, and Dacheng Tao. Revisiting plasticity in visual reinforcement learning: Data, modules and training stages. In *International Conference on Learning Representations*, 2024.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems 28*, 2015. arXiv:1506.02617.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. Deep reinforcement learning with plasticity injection. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.15555.

Mary Phuong and Christoph H. Lampert. Functional vs. parametric equivalence of ReLU networks. In *International Conference on Learning Representations*, 2020.

Haoyuan Qin, Chennan Ma, Mian Deng, Zhengzhu Liu, Songzhu Mei, Xinwang Liu, Cheng Wang, and Siqi Shen. The dormant neuron phenomenon in multi-agent reinforcement learning value factorization. In *Advances in Neural Information Processing Systems*, 2024.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, 2023.

Pierre Stock, Benjamin Graham, Rémi Gribonval, and Hervé Jégou. Equi-normalization of neural networks. In *International Conference on Learning Representations*, 2019. arXiv:1902.10416.

Jiuqi Wang, Jayanth Srinivasa, Claire Chen, Shuze Daniel Liu, Ali Payani, and Shangtong Zhang. Predicting plasticity in deep continual learning: A theoretical perspective. *arXiv preprint arXiv:2605.09044*, 2026.

Qinyou Wang. Fiber fingerprints of hidden learning-state dynamics. *arXiv preprint arXiv:2608.15976*, 2026a.

Tongxi Wang. Hidden gauge controls feature specialization in ReLU networks. *arXiv preprint arXiv:2608.06766*, 2026b.

Guowei Xu, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, Shuzhen Li, Yanjie Ze, Hal Daumé III, Furong Huang, and Huazhe Xu. DrM: Mastering visual reinforcement learning through dormant ratio minimization. In *International Conference on Learning Representations*, 2024. arXiv:2310.19668.

Kenny Young and Tian Tian. MinAtar: An Atari-inspired testbed for thorough and reproducible reinforcement learning experiments, 2019. arXiv:1903.03176.

Mingqi Yuan, Qi Wang, Guozheng Ma, Caihao Sun, Bo Li, Xin Jin, Yunbo Wang, Xiaokang Yang, Wenjun Zeng, Dacheng Tao, and Jiayu Chen. Plasticine: Accelerating research in plasticity-motivated deep reinforcement learning, 2025. arXiv:2504.17490.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. In *Advances in Neural Information Processing Systems 35*, 2022. arXiv:2205.10637.

A Claims and scope; protocol deviations and definition corrections

A deviation discovered by its own authors and left unwritten is indistinguishable from one that was hidden, so each is recorded here rather than absorbed. Tables 4 and 5 compress the record; the columns are the ones a reader needs to decide whether a deviation could have changed a conclusion. None of the seven changes a number reported from the MLP response experiment: D1, D3, D4, D6 and D7 concern the convolutional response collection, which was stopped incomplete, resumed

claim	evidence	scope
a layer's common scale is lost at the normalisation link; masks and fractions equal alike	Prop. 1; the audit on the pairs (Table 1; E17–E18)	twelve formulas under stated clamp conditions, 40 pairs; GraMa at $\tau = 0.01$ at a floor; reference-code score equal only as a fraction
the same diagnostic values, a different adaptation score	paired experiment (Figure 5; E9), re-measured on new adaptation seeds (E16)	ten checkpoints, one task, SGD; re-measured on the same checkpoints, *common* only
not the identity's trajectory re-expressed; remains after per-arm rate selection	float64 per-block control (Section 6.3; E21); post-hoc analysis of E16 (E19)	*common* only; the control passes its gate; a continuous scalar shift is not excluded
persists under greedy evaluation in separate episodes	greedy episodes beside the online score (Section 5; E23)	original seeds and fixed grids; greedy area over the horizon; exploration-mediated data differences not separated
detectable at a rescaling comparable in size to the observed change of one blind coordinate, in either direction	dose–response at $m \in \{\pm 0.1, \ldots, \pm 0.5\}$ (Section 6.4; E25)	*common* only; one coordinate's size, not the drift's direction or frequency; both arms re-collected in one acquisition after the cross-environment gate failed (90/90 within it); no equivalence margin at $\lvert m \rvert = 0.1$; no model
an optimisation difference on fixed data with no environment interaction	fixed transition sets, identical minibatch sequences, frozen or per-arm targets (Section 6.4; E26)	random-policy data; a TD objective, not policy quality; consistent with, not the pathway of, the online difference
small and negative at initialisation, largest at the intermediate snapshot, positive at the final	the stage collection (Appendix D; E22)	three snapshot–task settings on the same runs; snapshot and task change together; low initial returns
a predictor given the scale descriptor chooses no better than a fixed rule	the decision table (Table 14; E14)	constructed representatives; no gain over "always *common*"
between two representatives with equal diagnostics (E16), the noise-corrected response difference is 0.326 return units; evaluation noise is an estimated 11.4% of the observed mean square	noise separation on the E16 record (Section 5; E27, retrospective)	two points of each orbit, rates held fixed; supports that the contrast is not evaluation noise; not the orbit, not natural states
with the predictors tested, adding diagnostics did not improve prediction over the intervention's identity alone; the per-layer scale's incremental improvement is borderline (ridge) or undetected (k-NN)	predictor baselines on the E9 rows (Section 7; E28, retrospective)	constructed representatives; regret unchanged; a worse fit is not shown to be an absence of information
an initial first-order score (training-gradient energy at θ_0) has limited predictive power for the fixed-data and dose–response differences	first-order approximation at the selected rates (Section 6.4; E29, retrospective)	does not beat the constant predictor (9/10 vs 10/10); sign of the point estimates at 1/8 of eight non-independent magnitudes; no mechanism excluded
same-diagnostic pairs respond differently on a second architecture; the MLP's positive level effect is not reproduced	convolutional experiment (Section 6.2, Table 18; E11)	ten ConvNet checkpoints, SGD, a 20-rate grid; uncorrected per-pattern tests, no interaction test

Table 3: Each claim, its evidence and the scope inside which it is made.

as a continuation and read once, and now enters Section 6.2 with its continuation recorded in D7; D2 removes evidence rather than adding it; D5 is a viewing breach on pilot rows that never enter confirmatory evidence.

Definition corrections, all before the corresponding numbers were reported. Five corrections changed a metric definition or the statement of an estimand rather than a protocol, and are listed because a reader comparing this paper with an earlier draft or with the source papers will meet them. (a) The gradient-magnitude score had been implemented with the gradient with respect to the incoming weights; the source's Eq. 2 uses the gradient with respect to the hidden activation, and the audit was recomputed under it (the anisotropic range moved from 4.04 to 8.64; the layer-uniform verdict is 0.000 under both). (b) What an earlier draft called the feature rank of Lyle et al. (2022) was a relative-threshold rank; the published estimator uses an absolute cutoff on $\Phi/\sqrt{n}$, was added as its own row, and is the row that moves on the layer-uniform subgroup. (c) The convolutional update-activity audit had included the bias in the per-unit weight norm, which scales with the layer's own factor but not the previous layer's, so a quantity that must be invariant was not; corrected, it reads 0.000. A plausible explanation for the wrong value—one unit crossing the threshold at floating-point resolution—had been available and was wrong. (d) An earlier statement of Definition 1 wrote the in-sample maximum $\max_\eta \bar{A}(\eta)$, which is not what the analysis computes and would not reproduce the reported numbers; the definition now states the leave-one-out selection the code performs. In the same pass the aggregation bridge of Section 7 was found to have been fed the interquartile mean where it requires the arithmetic mean, and its two figures were recomputed and the earlier ones withdrawn; and a matched-selection sensitivity analysis for the equivariant reference was found to have transferred a rate selected on one curve to another, and its reported sensitivity was withdrawn (all three family-level verdicts are unchanged under the correct matched selection). Under (d) no confirmatory acquisition number changed; the two derived bound figures were recomputed from the arithmetic mean and the earlier ones withdrawn, and the text describing the estimand changed. A results note had also carried the ReBorn share's interval as $[67\%, 73\%]$ without recording its estimator; it is superseded by the recomputation of Appendix F, whose narrower interval is the output of a stated estimator and not stronger evidence. (e) An external review of the compiled draft found four further mismatches between text and computation, all fixed: the selection-aware interquartile interval had been printed from the leaky one-copy re-run ($[0.168, 0.328]$) while the prose described the every-copy re-run, whose interval $[0.1565, 0.3233]$ is now printed; the dormant ratio's clamp had been listed in the formula table but omitted from the conditional list, and its inactivity on the paired rows is now verified and stated; the weight-difference row had been described as relative to an identity value that is zero by construction, and is now labelled as the absolute range the audit actually computes; and the response figure's top-panel marker had been labelled P^{loo} although it is the pooled curve's in-sample maximum, and is now labelled as descriptive.

	registered	changed	visible at the time	what replaced it	effect on confirmatory evidence
D1	Architecture selection for the convolutional collection screens on a TD-loss trend over the final 100k training steps, plus a `NaN`/`Inf` clause; adequacy on a trailing-10 episode median.	2026-09-01	Both candidate runs' logs. No loss value had ever been logged, for any run.	Adequacy evaluated under all four available readings (both candidates pass every one); the `NaN`/`Inf` clause applied (both pass); the trend clause recorded as not evaluable, and a probe-based surrogate computed and *rejected* because it flags both candidates. The cheaper candidate retained under the pre-specified tie-break. Logging added from the confirmatory seeds on.	None: the only branch reaching the other candidate required the retained one to fail, and no available reading fails it. The selection is on adequacy and numerical stability, not on divergence.
D2	Horizon sensitivity: $T \in \{2000, 3000, 4000\}$ primary, $T = 1000$ degenerate.	2026-09-01	All horizon results.	2000 and 3000 re-designated exploratory; 4000 remains the primary, pre-registered reading; 1000 found structurally non-estimable.	Claims less: two horizons leave the confirmatory record, nothing enters it. Every detected cell stands at 4000 as read.
D3	Grid-extension rule for the convolutional pilot: one step in *the* failing direction, at most twice, within $[10^{-8}, 0.1]$.	2026-09-02	Two of nine pilot rows; then a non-evidentiary probe (three gauge points, $K = 2$, argmax positions only, flagged non-evidence on every row).	Amended, before any confirmatory row existed, to require one guard point beyond the most extreme probe argmax in *each* direction, inside the unchanged hard bounds. The rule had been underspecified (two directions failed) and insufficient (extending *to* an argmax leaves it censored); a first attempt made the second error and was stopped before collection.	None yet reported. Moves the measurement domain so the estimand is observable; no hypothesis, threshold, adequacy bar or estimator moved. Pilot and probe rows never enter confirmatory evidence.
D4	Three verdict labels per endpoint (non-negligible, practically negligible, inconclusive); censoring handling stated for seeds, not endpoints.	2026-09-02	Nothing beyond D3's probe; fixed before the pilot rows existed.	A censored endpoint is a fourth label carrying no directional or null verdict; censoring is per endpoint, not per condition; the descriptive "lockstep" read-out is declared non-decisive for the convolutional verdict, which comes from the condition-wise rule alone.	Removes evidence only: no reading under which a censored cell helps the hypothesis.

Table 4: Protocol deviations D1–D4, compressed from the deviation record (D5–D7 in Table 5). Dates are when the deviation was written down, which in every case is the same day it was found. "Visible at the time" is what had been looked at when the replacement rule was chosen, because a rule chosen after seeing the outcome it governs is a different kind of object from one chosen before.

	registered	changed	visible at the time	what replaced it	effect on confirmatory evidence (cont.)
D5	Pilot viewing firewall: no endpoint values and no ordering between gauge conditions before the adequacy read.	2026-09-06	**Breached**: five of forty-five pilot rows seen at endpoint level, with their ordering, while checking an acquisition stamp printed on the same stream.	Stamp verification moved to a tool that reads the acquisition line only. The runner's print is not changed mid-collection because that source file is hashed into the freeze digest. No threshold, floor or rule touched.	None for any reported number: adequacy and censoring are computed by script against frozen thresholds, and pilot seeds never enter confirmatory evidence. The censoring step was the one genuinely at risk, and is stated as such.
D6	Pilot adequacy: every headline contrast has $\geq 4/5$ paired-usable seeds; on failure, extend the grid in the censoring direction.	2026-09-06	The completed pilot's adequacy read: $45/45$ rows, zero edge-censoring, one contrast at $3/5$ because of bands broken in the grid's interior.	The rule keys its response on a censoring direction and there was none. Amended, after seeing the counts, to evaluate usability per endpoint and to report a below-gate cell *without* a direction as non-estimable rather than extend a grid. The amendment does not rescue the failing contrast ($3/5$ under both readings, checked before adopting it).	Reduced claim for the collection, had it completed: level at every condition; bandwidth and location at every condition except one, which carries no verdict there, not even a null. No floor or threshold moved.
D7	The convolutional response acquisition runs to 90 rows and is read in the pre-registered order; it enters the paper only on completion.	2026-09-12	Row counts, stamps and process state only. No row's content was read: 20 of 90 rows on disk, all parseable, no duplicates, one acquisition stamp.	The acquisition was **stopped** on 2026-09-12 (the worker tree died twice without error output, 2026-09-03 and 2026-09-11, and the machine entered standby repeatedly) with no verdict formed from the partial rows, then **resumed as a continuation** on 2026-09-13 at the user's instruction: the same acquisition stamp, checkpoints and grid; batch A (seeds 106–110) from its 20 kept rows to 45, batch B (seeds 111–115) from 0 to 45; the resumed collection was throttled by the machine's power mode for part of its run (timing enters no result). Rows were read once, at 90, in the protocol's order (2026-09-14).	The reading is Section 6.2: the pre-registered structural target is met, the MLP's level direction does not carry. No row was read before all 90 existed; the pilot firewall (D5) was kept.

Table 5: Protocol deviations D5–D7, continued from Table 4; same columns and conventions.

B FULL AUDIT TABLE AND METRIC PROVENANCE

A diagnostic is a formula, not a name, and three things routinely differ: what a paper defines, what its reference code computes, and what a third party audits. Table 6 states, for each audited diagnostic, the formula *as we compute it*, so that an exact-invariance claim attaches to a specified object, and Table 7 gives the full audit that Table 1 excerpts. Throughout, h_i is the post-activation output of hidden unit i, expectations are over a fixed probe batch drawn once per checkpoint, $\Phi \in \mathbb{R}^{n \times H}$ is the post-activation feature matrix of the last hidden layer on that batch, uncentred, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots$, and $\mathcal{L}$ is the training loss on probe batches: the TD loss of the checkpoint's own stored probe transitions (its pre-switch training task) with a frozen copy of the checkpoint as the target network, so the metric imports no optimizer or target state from elsewhere in training. The two gradient scores aggregate in different orders: the gradient-magnitude score takes the per-sample magnitude $|\partial \mathcal{L}_{\text{batch}} / \partial h_i^{(b)}|$ and averages over samples, a $\mathbb{E}|\nabla|$; the update-activity ratio takes the norm of the batch-averaged weight gradient, a $|\mathbb{E}\nabla|$ that includes cancellation across samples. Both transform the same way under the gauge and mean different things; and a gradient-to-weight ratio is not the update-to-weight ratio an optimizer with momentum or per-parameter scaling produces. Within-layer means are over the units of one layer. Every implementation is pinned by a test on a fixed numerical fixture; the two transcribed from a library are additionally checked against it at a named commit.

Provenance chain. For each row the chain is: source and version $\rightarrow$ equation $\rightarrow$ the tensor that equation names $\rightarrow$ our function $\rightarrow$ what we changed $\rightarrow$ the fixture that pins it. The library rows are transcribed from `plasticine_metrics/rank.py` at commit `aa00b4bb18f7fe298a47e1ce36c32ba55ce064e8` and pinned by `tests/test_plasticine_parity.py` on fixed fixtures. The gradient-magnitude score is audited against v2 (2026) of its arXiv source, and needs three readings kept apart. (a) *As printed*, Eq. 2 takes $|\partial \mathcal{L} / \partial h_i|$ with h_i the neuron's activation; our implementation differentiates the post-ReLU tensor, literally that. (b) *As argued*, the source's Theorem 1 reasons through the ReLU gate, i.e. $\partial \mathcal{L} / \partial z_i = \partial \mathcal{L} / \partial h_i \cdot \mathbf{1}[z_i > 0]$, which vanishes on a dead unit where $\partial \mathcal{L} / \partial h_i$ need not. The two raw per-input gradients agree wherever the unit is active and differ where it is not; but the metric averages over inputs and normalises by the layer mean, so a unit inactive on some inputs has $g_i \neq \tilde{g}_i$, and a change in any unit's raw magnitude moves every unit's normalised score (raw $(1, 1)$ against $(1, 0)$ normalise to $(1, 1)$ against $(2, 0)$), so the two readings can classify *always-active* units differently as well. What the readings share is the layer-uniform invariance, since both scale by $1/c_i$; what they do not share is the metric's value or its inactive set. (c) *As implemented upstream*: the reference code's formula, transcribed and run on the same pairs (Table 8); the upstream package itself is not run. Reading (a) is the audited formula; reading (c) is reported beside it.

The audit on the confirmatory pairs themselves. Table 7 is read on gauge draws around one trained network. Table 8 repeats the audit on the 90 E9 representatives (the 40 layer-uniform and 40 anisotropic pairs of Section 5, rebuilt and asserted against their frozen rows; `PROTOCOL_E17_AUDIT_PAIRS.md`, `runs/e17_read.json`), with GraMa in both readings—Eq. 2 as printed and the reference code's formula transcribed—at the source's threshold and at the audit's, and with each diagnostic's identity values across the ten checkpoints so that a "40/40 equal" cannot be mistaken for a statistic pinned at a floor. Three things are to be read as found. (i) Every within-layer-normalised, thresholded diagnostic is bit-identical to the identity's value on all 40 layer-uniform pairs, at every threshold and in both GraMa readings, and moves on the anisotropic pairs. (ii) **GraMa's Eq. 2 reading at the source's own $\tau = 0.01$ is 0 on every identity checkpoint** (and stays 0 on 20 of the 40 anisotropic pairs): at that threshold, on these networks and this probe, the statistic sits at its floor, so its equality across a pair says less than the same equality at $\tau = 0.25$, where the identity values are 0.047–0.089; the reference-code reading at $\tau = 0.01$ is not at the floor (0.109–0.193). (iii) The reference-code reading's continuous score is *not* bit-identical on any layer-uniform pair (largest relative change 1.0×10^{-6}, the additive 10^{-9} in its normaliser); only its thresholded fraction is. We therefore do not say that the reference code's formula is exactly invariant: on the pairs evaluated, its reported fraction was equal (the formula was transcribed and compared on fixtures; the upstream package was not run). The feature rank (absolute cut) moves on 10 of the 40 layer-uniform pairs and the whole-parameter gradient norm on all 40.

diagnostic	formula as audited	normalisation	threshold	bias	ε	variant
fraction of active units	$\frac{1}{H}\sum_i \mathbf{1}[\mathbb{E}\lvert h_i\rvert > 0]$, pooled over hidden layers	none (sign only)	0	–	–	ours
numerical rank	$\#\{i : \sigma_i > \max(n, H)\,\epsilon_{64}\,\sigma_1\}$, the `torch.linalg.matrix_rank` default (torch 2.13, float64), i.e. a tolerance relative to σ_1	relative tol.	see formula	–	–	ours
dormant ratio (RDU)	$s_i = \mathbb{E}\lvert h_i\rvert / \max(\mathrm{mean}_k\, \mathbb{E}\lvert h_k\rvert,\ \varepsilon)$; dormant iff $s_i \le \tau$; fraction pooled over layers	within-layer mean	$\tau = 0.025$	–	10^{-12} clamp on the mean	paper (Sokar et al., 2023); the clamp is ours, and is inactive on every paired row (7.8×10^{10} margin)
GraMa dormant ratio	$g_i = \mathbb{E}\lvert\partial\mathcal{L}/\partial h_i\rvert$; $G_i = g_i / \max(\mathrm{mean}_k\, g_k,\ \varepsilon)$; inactive iff $G_i \le \tau$	within-layer mean	$\tau = 0.25$	–	10^{-12} clamp	paper, Eq. 2 as printed (v2, 2026) (Liu et al., 2025): h is the post-ReLU activation; the reference code differentiates the weights instead (Table 7, ‡); $\tau = 0.25$ is the audit's setting, the source's experiments use 0.01
stagnant ratio (RUA)	$\mathrm{UA}_i = \lVert\nabla_{w_i}\mathcal{L}\rVert / (\lVert w_i\rVert + \varepsilon)$ averaged over probe batches; $\mathrm{RUA}_i = \mathrm{UA}_i / \mathrm{mean}_k \mathrm{UA}_k$; stagnant iff $\mathrm{RUA}_i < \alpha$	within-layer mean	$\alpha = 0.25$	excluded	10^{-8} additive	analogue (Liu et al., 2026): snapshot over probe batches, not a training-time window
cumulative spectral rank	$\min\{k : \sum_{i\le k}\sigma_i / \sum_i \sigma_i \ge 1-\delta\}$	ratio of singular mass	$\delta = 0.01$	–	–	paper (Kumar et al., 2021); implementation transcribed from Plasticine (Yuan et al., 2025), which calls it "stable rank"
stable rank	$\lVert\Phi\rVert_F^2 / \sigma_1^2$	ratio to σ_1^2	–	–	–	ours (the linear-algebra quantity)
relative threshold rank	$\#\{i : \sigma_i/\sigma_1 > r\}$	ratio to σ_1	$r = 0.01$	–	–	ours
entropy effective rank	$\exp\big(-\sum_i p_i \log p_i\big)$, $p_i = \sigma_i / \sum_j \sigma_j$	ratios	–	–	–	paper (Roy & Vetterli, 2007); implementation transcribed from Plasticine (Yuan et al., 2025), its "effective rank"
feature rank	$\#\{i : \sigma_i(\Phi/\sqrt{n}) > \epsilon\}$	$1/\sqrt{n}$ only; **absolute** cut	$\epsilon = 0.01$	–	–	paper, Def. 2 (Lyle et al., 2022)
gradient norm	$\lVert\nabla_\theta\mathcal{L}\rVert_2$ over all parameters, averaged over probe batches	none	–	included	–	ours
weight difference	$\lVert\theta - \theta_{\mathrm{ref}}\rVert_2$ over all parameters, θ_{ref} the identity representative	none	–	included	–	ours; two-state, reference held fixed

Table 6: The twelve audited formulas as computed: five published diagnostics as printed (dormant ratio, GraMa Eq. 2, cumulative spectral rank, entropy effective rank, feature rank), one snapshot analogue of a published training-window quantity (stagnant ratio), and six quantities of our own included as controls or to disambiguate colliding names (fraction of active units, numerical rank, stable rank, relative threshold rank, gradient norm, weight difference); the reference-code GraMa reading is a second reading of one of the five. "Bias" says whether bias parameters enter a per-unit or whole-parameter norm; "ε" whether a small constant enters a denominator, and how. Exact invariance on the layer-uniform subgroup is claimed for the formula in the second column and, where an ε or clamp appears, under the condition stated in Section 3.2.

Table 7: The full audit on a trained MLP with the realised function held to 1.2×10^{-15} (Table 1 shows eight of these rows). "Formula / code": where the formula comes from and what was run. "Layer-uniform": the invariance *class* of the formula as run, then the *agreement observed* over the 15 layer-uniform draws (exact: the stored values are equal; $\leq$: largest relative difference; "paired": additionally bit-identical on all 40 confirmatory pairs). "Anisotropic": relative range over the anisotropic draws, read only as zero against non-zero. The weight difference is absolute and sits in its own panel.

diagnostic	signal / normalisation / threshold	formula / code	layer-uniform: class; agreement	anisotropic
fraction of active units	activation, sign only	own	ideal; exact	**0.000**
numerical rank $\mathrm{rank}\,\Phi$	spectral, default tolerance	own; `torch` default	numerical rank at tolerance; exact	**0.000**
dormant ratio (RDU)[¶]	activation, within-layer mean	paper formula (Sokar et al., 2023); own clamp	conditional (clamp inactive); exact, paired	2.44
GraMa, Eq. 2 reading[‡¶]	gradient, within-layer mean	Eq. 2 as printed (v2, 2026) (Liu et al., 2025), post-ReLU; own τ, own clamp; upstream code uses the weight gradient (see ‡)	conditional (clamp inactive); exact	8.64
stagnant ratio, snapshot analogue[†]	gradient, within-layer relative	own analogue of Liu et al. (2026), time average removed	conditional ($\varepsilon = 0$); exact at $\varepsilon = 0$	4.61
cumulative spectral rank	spectral, ratio of singular mass	Kumar et al. (2021); library (Yuan et al., 2025), pinned commit	numerical rank at tolerance; exact, paired	0.56
stable rank	spectral, ratio $\|\Phi\|_F^2/\sigma_{\max}^2$	own	ideal; $\leq 1.5 \times 10^{-15}$	0.68
relative threshold rank	spectral, ratio to $\sigma_{\max}$	own	numerical rank at tolerance; exact	0.73
entropy effective rank	spectral, ratios	Roy & Vetterli (2007); library (Yuan et al., 2025), pinned	ideal; $\leq 8.2 \times 10^{-16}$	0.75
feature rank	spectral, **absolute** cut on $\Phi/\sqrt{n}$	paper, Def. 2 (Lyle et al., 2022)	not invariant; moves 0.362 (responds when a scale change moves a singular value across the cut: 10 of the 30 reached confirmatory pairs, Table 10)	0.12
gradient norm (all parameters)	gradient, none	own	not invariant; moves 6.79	487
absolute range (identity value 0 by construction), not comparable with the rows above				
weight difference[§]	parameter, none	own	not invariant; moves 12,277	78,178

[†] The published estimator carries an ε in its denominator and averages over a training window; ours is computed at one checkpoint over probe batches. The value shown is the reported thresholded ratio, which was unchanged here; the underlying score is exactly invariant only at $\varepsilon = 0$. [‡] Audited against the source's Eq. 2 as printed (v2, 2026), whose signal is the gradient with respect to the hidden activation. The source's reference code differs: in `torressliu/grad-based-plasticity-metrics`, `utils/ReDo.py` at commit `e4c169dc`, `GradientReDo._get_layer_mask` scores a unit by `layer.weight.grad.abs().mean(dim=1)`—the *weight* gradient, averaged over inputs—normalised by `grad_magnitude.mean() + 1e-9` and thresholded at `<= tau` (read 2026-09-13; no activation hook). That reading is also a within-layer-normalised score; its normaliser is additive, so the score is not exactly invariant on H_{unif}, and on the confirmatory pairs its thresholded fraction was equal while its continuous score changed by up to 1.0×10^{-6} (Table 8; an earlier draft of this audit ran it: anisotropic range 4.04 against 8.64 for Eq. 2; Appendix A, correction (a)). The audit threshold $\tau = 0.25$ is ours; the source's experiments use $\tau = 0.01$ (and 0 for the activation-function study), at which Eq. 2 reads 0 on every identity checkpoint here (Table 8). [¶] Divides by a layer mean clamped at 10^{-12} (Section 3.2); the clamp is inactive on the audit's draws and on all 90 confirmatory representatives for both rows (smallest layer mean of $|h|$ 7.8×10^{-2}, of g 3.0×10^{-4}; `scripts/e9_clamp_check.py`, record E18). [§] A two-state quantity $\|\theta - \theta_{\mathrm{ref}}\|_2$. The convention here transforms one state and holds the reference fixed at the identity representative, so the row reads the distance from a fixed frame and its identity value is 0 by construction. It therefore cannot be normalised like the other rows: the audit divides by the identity value when that is non-zero and by 1 otherwise, so **this row's entries are absolute ranges** ($\max - \min$ over draws, in parameter ℓ_2 units: the draws span $[16.6,\ 1.229 \times 10^4]$ on the layer-uniform subgroup and $[39.4,\ 7.822 \times 10^4]$ on the anisotropic family) and are not comparable with the relative entries above them. The row is listed because monitoring collections report the quantity, and its non-invariance is a property of the fixed-reference convention.

Table 8: The audit on the 90 E9 representatives (record E17). "Ideal inv.": whether the *ideal* formula ($\varepsilon = 0$, no clamp) is invariant on H_{unif}, as in Table 1; a row's implemented stabiliser is in the "clamp / ε" column and the residual it leaves in "continuous score", so a "yes" here is not a claim that the implemented formula is exactly invariant (it is not, for the reference-code and additive-ε rows). "Reported integer/fraction equal": pairs on which the diagnostic's reported value equals the identity representative's bit-for-bit, out of 40 layer-uniform and 40 anisotropic pairs (with the largest relative change where not all are equal); left blank for the continuous diagnostics, whose bit equalities count floating-point agreement only. "Continuous score": the largest relative change of the underlying score vector's sum of squares, or of the value itself for a continuous diagnostic, so that exact invariance of the normalisation is told apart from floating-point agreement. Thresholds marked "source" are the source paper's defaults or the setting of its experiments. The realised function is held to 1.2×10^{-15} on every pair.

diagnostic	threshold	clamp / ε	ideal inv.	identity values (10 ckpts)	reported integer/fraction equal		continuous score, max rel. change	
					layer-uniform	anisotropic	layer-uniform	anisotropic
dormant ratio (RDU)	$\tau = 0.025$ (source)	clamp 10^{-12}, inactive	yes	0.109–0.219	40/40	3/40	—	—
GraMa, Eq. 2 reading	$\tau = 0.01$ (source's experiments)	clamp 10^{-12}, inactive	yes	0 (all)	40/40	20/40	4.2×10^{-16}	9.623
GraMa, Eq. 2 reading	$\tau = 0.25$ (audit's earlier setting)	clamp 10^{-12}, inactive	yes	0.0469–0.0885	40/40	0/40	4.2×10^{-16}	9.623
GraMa, reference-code formula	$\tau = 0.01$ (source's experiments)	additive 10^{-9} in the normaliser	yes	0.109–0.193	40/40	14/40	1.0×10^{-6}	8.932
GraMa, reference-code formula	$\tau = 0.25$	additive 10^{-9}	yes	0.12–0.219	40/40	0/40	1.0×10^{-6}	8.932
stagnant ratio, snapshot analogue	$\alpha = 0.25$ (source)	$\varepsilon = 10^{-8}$ additive	yes	0.115–0.224	40/40	0/40	—	—
stagnant ratio, snapshot analogue	$\alpha = 0.25$	$\varepsilon = 0$	yes	0.115–0.224	40/40	0/40	4.0×10^{-16}	30.0
cumulative spectral rank	$\delta = 0.01$ (source)	—	yes	49–54	40/40	0/40	—	—
relative threshold rank	$r = 0.01$ (ours)	—	yes	37–54	40/40	0/40	—	—
numerical rank	`torch` default	—	yes	57–61	40/40	40/40	—	—
fraction of active units	0 (sign)	—	yes	0.807–0.891	40/40	40/40	—	—
entropy effective rank	none (continuous)	—	yes	15.1–23.4	—	—	2.7×10^{-15}	0.876
stable rank	none (continuous)	—	yes	1.27–1.92	—	—	1.1×10^{-15}	0.340
feature rank (absolute cut)	$\epsilon = 0.01$ (source)	—	no	57–61	30/40 (max 0.018)	12/40	—	—
gradient norm, all parameters	none (continuous)	—	no	1.72–4.22	—	—	0.579	178.5

Names that collide. Three formulas circulate under two names. The library's "stable rank" is the cumulative singular-mass rank; the stable rank of linear algebra is $\|\Phi\|_F^2/\sigma_1^2$; the library's "effective rank" is the entropy form. No bare rank name appears in this paper, and each row above carries its formula.

The clamp and ε conventions in full. Two within-layer normalisations are clamped, and a clamp is not invariant. The dormant ratio divides each unit's mean $|h_i|$ by $\max(\overline{|h|}, \varepsilon)$ and the gradient-magnitude score divides g_i, the gradient with respect to the hidden activation as in the source's Eq. 2, by $\max(\overline{g}, \varepsilon)$, both with $\varepsilon = 10^{-12}$. The unclamped scores are invariant under layer-uniform rescaling; the clamped ones are not (with $a = (1, 3)$ and common factor $\varepsilon/8$ the clamp binds and the score moves from $(0.5, 1.5)$ to $(0.125, 0.375)$), and a small enough factor moves the thresholded ratio too). Each implementation therefore agrees with its invariant score on comparisons for which the clamp is inactive at both endpoints. For the dormant ratio that condition is verified on every pair the argument of Section 5 uses; for the gradient-magnitude score it is verified on the same 90 representatives (smallest layer mean of g: 3.0×10^{-4}, against a clamp of 10^{-12}; Table 11). **The update-activity ratio is a snapshot analogue.** The published $\mathrm{UA}_i = \|g_i\|/(\|w_i\| + \varepsilon)$ is averaged over a training window; ours is computed at one checkpoint over probe batches with $\varepsilon = 10^{-8}$. At $\varepsilon = 0$ the raw quantity is homogeneous of degree -2 and the layer mean cancels it exactly; at $\varepsilon > 0$ this generally fails ($\max_i |\Delta \mathrm{RUA}_i|$ over six layer-uniform draws: 8.9×10^{-16} at $\varepsilon = 0$, 1.9×10^{-7} as shipped; the thresholded ratio was unchanged in both). Proposition 1 is applied to the $\varepsilon = 0$ analogue only. **The rank rows are numerical ranks** at a stated tolerance (Appendix B); the mathematical rank is invariant as a quantity, but neither estimator is it.

Table 9: How the raw signals transform under a layer-uniform rescaling with $u_\ell = \log c_\ell$ ($u_0 = u_{m+1} = 0$), for a fixed probe and the same scalar loss, and what the within-layer normalisation then removes. The laws differ from signal to signal; the shared blindness arises only after each is divided by its own layer's common factor (derivations in Appendix B).

raw quantity (layer ℓ, unit i)	factor after rescaling	what the layer normalisation removes
activation magnitude $\mathbb{E}\|h_{\ell i}\|$	e^{u_ℓ}	the layer's common activation scale
activation-gradient magnitude $\mathbb{E}\|\partial\mathcal{L}/\partial h_{\ell i}\|$	e^{-u_ℓ}	the layer's common gradient scale (opposite sign)
incoming-weight-gradient magnitude (row of $\nabla_{W_\ell}\mathcal{L}$)	$e^{u_{\ell-1}-u_\ell}$	a scale set by two adjacent layers
gradient-to-weight ratio $\|\nabla_{w_{\ell i}}\mathcal{L}\|/\|w_{\ell i}\|$ ($\varepsilon = 0$, bias excluded)	$e^{2(u_{\ell-1}-u_\ell)}$	the same, squared
singular values of Φ_ℓ	e^{u_ℓ}	the spectrum's overall size

Which gauge reaches which quantity. Under a diagonal gauge $W'_\ell = C_\ell W_\ell C_{\ell-1}^{-1}$ ($C_0 = C_{L+1} = I$; biases $b'_\ell = C_\ell b_\ell$), for the same scalar loss, $h'_i = c_{\ell,i} h_i$ and $\partial\mathcal{L}/\partial h'_i = (\partial\mathcal{L}/\partial h_i)/c_{\ell,i}$ for i in layer ℓ, and $\nabla_{W'_\ell}\mathcal{L} = C_\ell^{-1}(\nabla_{W_\ell}\mathcal{L})C_{\ell-1}$. The activation signal therefore carries weight $+1$ and the activation gradient -1 in the unit's own factor alone. The incoming row w_i and its gradient row are different: $w'_i = c_{\ell,i}\, w_i C_{\ell-1}^{-1}$ and $\nabla_{w'_i} = c_{\ell,i}^{-1}\,(\nabla_{w_i})C_{\ell-1}$, so their norms scale by one factor only when the previous layer is not rescaled anisotropically—the single-layer case the earlier draft of this paragraph described. On H_{unif} exactly, $\|w'_{\ell i}\| = (c_\ell/c_{\ell-1})\|w_{\ell i}\|$, $\|\nabla_{w'_{\ell i}}\| = (c_{\ell-1}/c_\ell)\|\nabla_{w_{\ell i}}\|$, and with $\varepsilon = 0$ and the bias excluded $\mathrm{UA}'_{\ell i} = (c_{\ell-1}/c_\ell)^2\,\mathrm{UA}_{\ell i}$: one factor per layer, which the layer mean removes. Under an anisotropic $C_{\ell-1}$ the row norms mix the previous layer's factors and no single per-unit exponent describes them; the anisotropic entries of the tables are measurements, not instances of a weight rule. Only the gauge of the last hidden layer reaches Φ, as a column scaling $\Phi \mapsto \Phi C_L$; a rescaling confined to an earlier layer leaves Φ unchanged (Table 10). Within-layer normalisation turns each per-layer factor into a relative reweighting that cancels, which is the content of Proposition 1 for these rows.

Why the values were equal (record E18). Tables 10 and 11 decompose the equalities of Table 8 on the same 90 representatives (`PROTOCOL_E18_AUDIT_STAGES.md`; nothing chosen after reading). *Reach.* Each layer-uniform pattern multiplies the post-ReLU activation matrix of each hidden layer by one factor or leaves it bit-unchanged: *common* scales all three layers by $e^{0.7} = 2.01$, *first* only the first and *last* only the last by $e^{0.7\sqrt{3}} = 3.36$ (the width-weighted budget on one layer), *alternating* the three by $e^{0.7}, e^{-0.7}, e^{0.7}$. The spectral rows read the last layer, which the ten *first* pairs leave unchanged: on them the measured feature did not change, so they do not test whether a normalisation discards scale and are set aside; on the 30 pairs that do rescale it the cumulative

spectral rank and the relative threshold rank are equal on all 30, the feature rank on 20 (it moved on 10, all of them reached pairs), and the continuous entropy and stable ranks agree to 10^{-15}. Per layer, the same holds for every hidden layer of the *common* pairs. *Stages.* For every unit-based diagnostic the raw signal changes by one common factor within each layer on all 40 layer-uniform pairs (RDU by up to $2.4\times$, GraMa by 1–3×, the update activity by $15\times$; the $\varepsilon = 10^{-8}$ variant of the latter is not one exact factor, as its additive term predicts), the normalised score changes by at most 2×10^{-15} (5×10^{-6} for the reference-code reading, 3×10^{-8} at $\varepsilon = 10^{-8}$), the masks are identical unit for unit on 40/40, and so the fractions are equal: the scale was discarded at normalisation, not hidden by the threshold or the pooling. *Margins, per diagnostic and in the same normalised-score units as the changes.* On the identity checkpoints the smallest distance of any unit's score to its threshold, against that diagnostic's largest layer-uniform score change over all units, layers and pairs: dormant ratio 9.8×10^{-5} against 2.2×10^{-15}; GraMa Eq. 2 2.1×10^{-4} ($\tau = 0.25$) and 5.2×10^{-2} (0.01) against 1.8×10^{-15}; reference-code GraMa 5.7×10^{-4} (0.25) and 1.2×10^{-3} (0.01) against 5.1×10^{-6}; stagnant ratio 1.4×10^{-3} against 9.8×10^{-15} ($\varepsilon = 0$) and 3.0×10^{-8} (10^{-8}). No unit sits on a boundary; for every row the margin exceeds the change by at least two orders of magnitude, so no unit could have crossed, and none did (Table 11). *Anisotropic pairs.* There the score changes are of order 1–60 and the masks flip: median flip rates over the 40 pairs 0.07 (dormant ratio), 0.32 (GraMa, $\tau = 0.25$), 0.23 (reference code, 0.25) and 0.52 (stagnant ratio) of the units, maxima 0.23, 0.61, 0.48 and 0.73. An equal pooled fraction with a different mask—the case in which the aggregation step, not the normalisation, erases a difference—occurred on exactly one of the 40 pairs (dormant ratio; the GraMa readings at $\tau = 0.01$ that are equal on 20 and 14 pairs are equal as masks as well, most of those masks being empty), and on the other pairs the flip rate exceeds the change of the fraction only slightly (medians 0.073 against 0.070 for the dormant ratio, 0.32 against 0.29 for GraMa). So on these pairs the audit's equalities on H_{unif} hold for the scalar and for the mask alike, and the pooling erased little that the mask retained; whether the aggregation step hides unit-level differences in other settings is not established here. *Probe dependence.* Recomputing the gradient diagnostics on the identity checkpoints with the 512 probe transitions split into 1, 4 (as used) or 8 batches leaves the Eq. 2 values at $\tau = 0.01$ (0 throughout) and at 0.25 unchanged, moves the reference-code fraction by up to 0.02 and the stagnant ratio by up to 0.03 (its update activity is a norm of a per-batch gradient, so the batching enters), and leaves every minimum margin above 3×10^{-4}; the floor at $\tau = 0.01$ is not a batching artefact.

Table 10: Reach (record E18): which hidden-layer activation matrices each layer-uniform pattern changes, by what factor, and the last-layer spectral values that equal the identity's. "Unchanged" is bit-for-bit or within 10^{-12}; the entropy and stable ranks and $\|\Phi\|_F$ are continuous, so their integer equalities count only floating-point agreement and are not evidence either way.

pattern (10 pairs each)	Φ_1	Φ_2	Φ_3 (read by the spectral rows)	last-layer spectral values equal to identity's
common	$\times 2.014$	$\times 2.014$	$\times 2.014$	cum. spectral 10/10, rel. threshold 10/10, feature rank 7/10, entropy 3/10, stable 1/10, $\|\Phi\|_F$ 0/10
first	$\times 3.362$	unchanged	unchanged	cum. spectral 10/10, rel. threshold 10/10, feature rank 10/10, entropy 3/10, stable 6/10, $\|\Phi\|_F$ 8/10
last	unchanged	unchanged	$\times 3.362$	cum. spectral 10/10, rel. threshold 10/10, feature rank 6/10, entropy 4/10, stable 1/10, $\|\Phi\|_F$ 0/10
alternating	$\times 2.014$	$\times 0.497$	$\times 2.014$	cum. spectral 10/10, rel. threshold 10/10, feature rank 7/10, entropy 1/10, stable 0/10, $\|\Phi\|_F$ 0/10

by last-layer reach, all 40 layer-uniform pairs: reached ($\Phi_3' = c\,\Phi_3$, 30 pairs): cum. spectral 30, rel. threshold 30, feature rank 20, entropy 8, stable 2; not reached (10 pairs): every integer-valued rank 10, and nothing is tested.

Domain. The invariance statements are made on the set where each formula is defined by its code, and the degenerate inputs return fixed values: the cumulative spectral rank and the entropy rank return 0 when $\sum_i \sigma_i = 0$; the stable rank and the relative threshold rank return 0 when $\sigma_1 = 0$; the dormant ratio and the Eq. 2 score divide by a layer mean clamped at 10^{-12}; the update activity adds ε to $\|w_i\|$ and its layer normaliser is clamped at 10^{-12} (a layer of zero raw activity scores 0 everywhere and is counted wholly stagnant); the scale descriptor clamps $\|w_i^{\mathrm{in}}\|$ and the ratio at 10^{-12} before the logarithm. None of these cases occurred: over all 90 representatives the minima are $\sum_i \sigma_i \geq 124$, $\sigma_1 \geq 22.6$, layer-mean $|h| \geq 0.078$, layer-mean $g \geq 3.0 \times 10^{-4}$, layer-mean UA $\geq 8.5 \times 10^{-4}$, $\|w_i^{\mathrm{in}}\| \geq 4.8 \times 10^{-3}$ and $\|w_i^{\mathrm{out}}\| \geq 2.0 \times 10^{-3}$ (record E18).

Table 11: Four stages (record E18) on the 40 layer-uniform and 40 anisotropic pairs. "One factor per layer": pairs on which every layer's raw signal changed by a single common factor (to 10^{-9}); "max change": the largest absolute change of any unit's normalised score on the layer-uniform pairs; "mask identical": pairs on which the flagged set is the same unit for unit; "fraction equal": pairs on which the pooled fraction is equal; "margin": the smallest $|s_i - \tau|$ on any identity checkpoint. The stagnant ratio's mask uses the source's strict inequality.

	raw signal (layer-uniform)		score	mask identical		fraction equal		margin
diagnostic	one factor per layer	max rel. change	max change (l.-u.)	l.-u.	aniso.	l.-u.	aniso.	min identity m_i
dormant ratio, $\tau = 0.025$	40/40	2.36	2.2×10^{-15}	40/40	2/40	40/40	3/40	9.8×10^{-5}
GraMa Eq. 2, $\tau = 0.01$	40/40	1.01	1.8×10^{-15}	40/40	20/40	40/40	20/40	5.2×10^{-2}
GraMa Eq. 2, $\tau = 0.25$	40/40	1.01	1.8×10^{-15}	40/40	0/40	40/40	0/40	2.1×10^{-4}
GraMa ref. code, $\tau = 0.01$	40/40	3.06	5.1×10^{-6}	40/40	14/40	40/40	14/40	1.2×10^{-3}
GraMa ref. code, $\tau = 0.25$	40/40	3.06	5.1×10^{-6}	40/40	0/40	40/40	0/40	5.7×10^{-4}
stagnant, $\varepsilon = 10^{-8}$	8/40	15.44	3.0×10^{-8}	40/40	0/40	40/40	0/40	1.4×10^{-3}
stagnant, $\varepsilon = 0$	40/40	15.44	9.8×10^{-15}	40/40	0/40	40/40	0/40	1.4×10^{-3}

The scale descriptor S, in full. For hidden unit i of layer ℓ, $q_{\ell i} = \log\big(\|w^{\text{out}}_{\ell i}\| / \max(\|w^{\text{in}}_{\ell i}\|, 10^{-12})\big)$, the incoming norm over the row of W_ℓ and the outgoing norm over the column of $W_{\ell+1}$, biases excluded, computed from the current parameters alone (no reference to the original checkpoint). $S = (\mu_q, \sigma_q)$ is the mean and the population standard deviation of q pooled over all 3×64 hidden units with equal weight per unit, which here equals equal weight per layer. On H_{unif} with biases excluded, $q'_{\ell i} = q_{\ell i} + \log c_{\ell-1} + \log c_{\ell+1} - 2\log c_\ell$ ($c_0 = c_{L+1} = 1$): each unit reads its neighbours' factors as well as its own, the pooled spread changes even under a globally uniform factor, and different transformations can produce the same S. It is a pooled scale-sensitive descriptor, not a coordinate on the gauge group, and the manuscript calls it that.

Per-layer $\bar{q}$ against the pooled S. With $\bar{q}_\ell$ the mean of $q_{\ell i}$ over layer ℓ and $u_\ell = \log c_\ell$, a layer-uniform rescaling with positive norms and inactive clamps gives $\Delta\bar{q}_\ell = u_{\ell-1} + u_{\ell+1} - 2u_\ell$; for three hidden layers $\Delta\bar{\mathbf{q}} = M\mathbf{u}$ with M the tridiagonal matrix with -2 on the diagonal and 1 beside it, $\det M = -4$, so M is invertible and the per-layer vector $\bar{\mathbf{q}}$ distinguishes distinct layer-uniform rescalings within an orbit. The pooled $S = (\mu_q, \sigma_q)$ compresses that vector to two numbers and can assign the same value to different $\mathbf{u}$. This is a statement about discrimination on the orbit, not about prediction; S as used in E14 is kept as defined, and E14 fitted no predictor with $\bar{\mathbf{q}}$. The retrospective E28 does fit one (Table 19): added to the intervention's identity and the audited values it lowers the ridge error marginally and leaves the decision unchanged, so discriminating the orbit is not by itself predicting the response.

The additive stabiliser and the margin condition. For a score $s_i = x_i/(\mu + \varepsilon)$ and a common factor a on the layer, $s'_i = ax_i/(a\mu + \varepsilon)$, so exactly $s'_i - s_i = s_i\,\varepsilon(a-1)/(a\mu + \varepsilon)$: an additive stabiliser leaves a residual of scale information in the score. Its thresholded mask is unchanged whenever $\max_i |s'_i - s_i| < \min_i |s_i - \tau|$, a sufficient condition read directly from Table 11's largest score change and smallest margin per row; it holds on every audited row, by at least two orders of magnitude. The reconstruction claim of Section 7 is for the ideal normalisation ($\varepsilon = 0$); with a stabiliser, $x_i = (\mu + \varepsilon)s_i$ still holds when ε is known.

Probe, loss and aggregation, in full. The probe is 512 transitions collected once per training run under a uniformly random policy on the checkpoint's own task, stored with the run and never resampled; the activation and spectral diagnostics read the 512 observations, the gradient diagnostics the 512 transitions in four consecutive batches of 128 in stored order. The loss is the DQN objective, $\text{smooth-}\ell_1$ with mean reduction over the batch, of $Q(s,a)$ against $r + \gamma(1-d)\max_{a'} Q_{\text{tgt}}(s')$ with γ the run's, the terminal flag d as recorded, and the target network a frozen copy of the identity representative (the same target for every representative of a checkpoint). The per-batch signals are averaged over the four batches before normalisation. Fractions pool over all hidden units (192), not over per-layer fractions; for the MLP's equal widths the two coincide, for the ConvNet (16, 16 channels and 64 units) they do not and the pooled form is reported. In the ConvNet audit a channel's activation and activation-gradient signals are the mean of $|\cdot|$ over the batch and both spatial dimensions, and a channel's weight row is the flattened kernel.

The convolutional audit. The unit of the gauge is the channel; the flatten between the last convolution and the first linear layer makes the correction a block-column scaling, and the function was checked numerically first (layer-uniform 2.06×10^{-15}, channel-wise 1.60×10^{-15}). The partition carries: every within-layer-normalised diagnostic reads exactly 0.000 on the channel-uniform subgroup, the unnormalised gradient norm moves, and the absolute-threshold feature rank moves (1.283). That last row was a prediction committed before the audit ran, with the failure branch fixed in advance: had it read 0.000, the partition would have been narrowed to a statement about the MLP.

Position relative to the closest prior work. Table 12 states, for the six papers nearest to this one, what each established, on what object, what this paper adds to it, and what this paper leaves open.

	established	object	added here	still open here
Dinh et al. (2017)	a reparameterisation changes sharpness while the function and generalisation stay fixed	one trained net; sharpness	the same instrument turned on diagnostics that do *not* move, paired with a response that does	sharpness and other non-audited measures
Lebeurrier et al. (2026)	rescaled parameterisations of one function train differently across learning rates; a conditioning strategy	training from initialisation; dynamics	that the audited diagnostics jointly discard the conditioning information the effect depends on, on trained checkpoints	whether a continuous scalar rate closes the gap; Adam
Lyle et al. (2024)	parameter norm and normalisation set the effective learning rate in RL	training runs; the norm as a variable	a paired orbit on which that variable moves while every ratio-normalised diagnostic is fixed	the pathway from conditioning to the online return
Lyle et al. (2022)	feature rank correlates with, and is not sufficient for, learning progress	training runs; rank	a constructed pair with equal rank (and equal dormancy, GraMa, stagnancy) and unequal response	rank's value on natural states
Wang (2026b) (v2)	a hidden per-neuron gauge at an identical initial function selects which neuron learns a feature and when; no scalar rate or clock change matches (flow and small-step GD)	two-layer teacher–student; initialisation	trained checkpoints, layer-uniform gauges, audited diagnostics, finite-budget protocol response	their theorem's analogue on a finite grid
Wang et al. (2026)	counterexamples to representation- and NTK-rank readiness; an optimisation-readiness score with a one-step guarantee and multi-step evaluation	constructed states sharing $f = 0$ but not gauge-related; rank moves	pairs on one gauge orbit where several audited diagnostics are simultaneously fixed; no replacement score	a predictor that works on natural states

Table 12: The closest prior work, by what it established and what this paper adds; Wang (2026b) compared in full text (v2 of 2026-09-13), Wang (2026a) (Section 2) by abstract and listing. The last column is answered for this paper as well: its decision-table evidence is on constructed representatives only (Table 14).

C PROOFS

Definition 2 (Blind subgroup)**.** $\mathrm{Blind}(D) = \{\, g \in G : D(g\theta) = D(\theta) \text{ for every } \theta \,\}$. The quantifier makes this a subgroup; the pointwise set at a fixed θ is a level set of $g \mapsto D(g\theta)$, not in general closed under composition.

Corollary 1 (Irreducible prediction error)**.** *If $D(\theta_1) = D(\theta_2)$ and $|T(\theta_1) - T(\theta_2)| = \Delta$, then every predictor φ observing only D satisfies $\max_i |\varphi(D(\theta_i)) - T(\theta_i)| \geq \Delta/2$.*

Proof. φ emits one value v at both points, so $\Delta \leq |v - T(\theta_1)| + |v - T(\theta_2)| \leq 2\max_i |v - T(\theta_i)|$. □

Corollary 2 (Approximate equality)**.** *Let $\|\cdot\|$ be a norm on the space of diagnostic vectors and φ be L-Lipschitz with respect to* that *norm. If $\|D(\theta_1) - D(\theta_2)\| \leq \varepsilon_D$ and $|T(\theta_1) - T(\theta_2)| = \Delta$, then $\max_i |\varphi(D(\theta_i)) - T(\theta_i)| \geq (\Delta - L\varepsilon_D)_+/2$. The residuals of Section 4 are componentwise maxima, i.e. ℓ_∞ distances, so L must be the ℓ_∞ Lipschitz constant; in another norm ε_D must be re-expressed.*

Proof. With $v_i = \varphi(D(\theta_i))$, $|v_1 - v_2| \leq L\varepsilon_D$ and $\Delta \leq |v_1 - T(\theta_1)| + |v_1 - v_2| + |v_2 - T(\theta_2)| \leq 2\max_i |v_i - T(\theta_i)| + L\varepsilon_D$. □

Corollary 1 is the case $\varepsilon_D = 0$, which the bit-identical fractions realise; for the reference-code GraMa score, whose continuous value moves by up to 5×10^{-6} across a pair (Section 4), the bound applies with that ε_D and the predictor's Lipschitz constant in the chosen norm, which is a property of the predictor and is not estimated here. This is a statement about the best any L-Lipschitz predictor can do on the pair, not about the predictors of Appendix D.

From a pairwise bound to a measured aggregate. Corollary 1 bounds a pair; what we measure is a mean over seeds and intervention patterns. Writing $\delta_{sj} = T(h_j\theta_s) - T(\theta_s)$ with $D(h_j\theta_s) = D(\theta_s)$, and e_s for the largest absolute error at seed s of any predictor seeing only D,

$$e_s \;\geq\; \tfrac{1}{2}\,\mathrm{mean}_j|\delta_{sj}| \;\geq\; \tfrac{1}{2}\left|\mathrm{mean}_j\delta_{sj}\right| \quad\Longrightarrow\quad \mathrm{mean}_s\, e_s \;\geq\; \tfrac{1}{2}\left|\mathrm{mean}_s\mathrm{mean}_j\delta_{sj}\right|. \tag{3}$$

Each application of the corollary stays within one seed, where the function is fixed; only the bounds are averaged across seeds.

Proposition 1. An element $h \in H_{\mathrm{unif}}$ is a tuple $(c_1, \ldots, c_{L-1})$ of positive scalars, one per hidden layer, acting by (1) with $C_\ell = c_\ell I$. The group G is the direct product of the per-layer groups of positive diagonal matrices and (1) is an action of it, so h is the composition of the single-layer elements h_ℓ, each multiplying the units of layer ℓ by the common factor c_ℓ and leaving every other layer's units unchanged. Degree-zero homogeneity under a common positive scaling of the units of any single layer is exactly $D(h_\ell\theta) = D(\theta)$ for every ℓ, every $c_\ell > 0$ and every θ. Applying it once per factor, $D(h\theta) = D(h_{L-1} \cdots h_1\theta) = D(\theta)$. □

Proposition 3 and Corollary 1 are proved in the main text; their content is in the premises, which Sections 4 and 5 supply.

Proposition 2, in full. Let G be the group of positive diagonal rescalings (1) (hidden layers $1, \ldots, L-1$; $C_0 = C_L = I$), and for $g \in G$ write $c_\ell \in \mathbb{R}^{H_\ell}_{>0}$ for its factor on hidden layer ℓ; $\mathrm{Blind}_G(D) = \mathrm{Blind}(D) \cap G$. (i) Let $x_\ell(\theta) \in \mathbb{R}^{H_\ell}_{\geq 0}$ be a per-unit signal of layer ℓ that transforms unit by unit, $x_{\ell i}(g\theta) = c_{\ell i}^{\,p}\, x_{\ell i}(\theta)$ for a fixed $p \neq 0$ (the activation magnitude, $p = 1$, by the equivariance $h_\ell(g\theta) = C_\ell h_\ell(\theta)$; the activation-gradient magnitude, $p = -1$, by the chain rule; Table 9), such that for every pair $i \neq j$ some state in the domain has $x_{\ell i}, x_{\ell j} > 0$, and let $s_\ell = x_\ell/\mathrm{mean}_k\, x_{\ell k}$ be its full within-layer-normalised score vector. Then $\mathrm{Blind}_G(s_\ell) = \{g : c_\ell \text{ is constant}\}$, the other layers' factors unconstrained. (ii) For $D = (s_1, \ldots, s_{L-1})$, one such score vector per hidden layer, $\mathrm{Blind}_G(D) = H_{\mathrm{unif}}$; the same remains true after adding to D any components whose blind sets contain H_{unif}. (iii) Let $r(\Phi) = (\sigma_i/\sum_j \sigma_j)_i$ be the full normalised singular-value spectrum of the last hidden layer's feature matrix. If the realisable feature matrices include one with orthonormal columns, then $\mathrm{Blind}_G(r) = \{g : c_{L-1} \text{ is constant}\}$, the earlier layers' factors unconstrained.

Proof. (i) Sufficiency is Proposition 1, and the other layers' factors do not enter by the assumed transformation law. Necessity: if $c_{\ell i} \neq c_{\ell j}$, take a state with $x_{\ell i}, x_{\ell j} > 0$; there $s_{\ell i}/s_{\ell j} = x_{\ell i}/x_{\ell j}$ becomes $(c_{\ell i}/c_{\ell j})^p \, x_{\ell i}/x_{\ell j} \neq x_{\ell i}/x_{\ell j}$, so s_ℓ changes and $g \notin \mathrm{Blind}_G(s_\ell)$. (ii) $\mathrm{Blind}_G(D) = \bigcap_\ell \mathrm{Blind}_G(s_\ell)$, and the intersection of the sets in (i) over all hidden layers is H_{unif}; intersecting further with a set containing H_{unif} leaves it unchanged. (iii) Under g, $\Phi \mapsto \Phi\, C_{L-1}$ (Appendix B, where the last hidden layer is indexed L); a constant c_{L-1} multiplies every singular value by it and fixes r; for Φ with orthonormal columns every singular value is 1 and those of ΦC_{L-1} are the entries of c_{L-1}, so a non-constant c_{L-1} changes r. □

Scalar summaries of these vectors—thresholded masks, pooled fractions, integer ranks, or a scalar function of r—have blind sets within G that *contain* the sets above; equality needs a diagnostic-specific argument. On the orthonormal construction of (iii) the entropy effective rank and the stable rank change under any non-constant c_{L-1} (their value is H at r uniform and strictly less otherwise), whereas the cumulative and relative-threshold ranks need a singular value to cross their cutoff. For the dormant ratio a hidden unit with zero incoming weights and bias b has $\mathbb{E}|h| = b$, so the score vectors range over every positive vector, a unit can be placed on either side of τ, and the thresholded fraction's blind set within G is that of (i)–(ii) as well; for the gradient-based scores this realisability is not established here. The proposition characterises blindness within G only, and says nothing about transformations outside it (permutations, or the non-diagonal reparameterisations of Dinh et al., 2017).

D ESTIMAND SPECIFICATION AND THE BOOTSTRAP

	P^{loo}	W_{90}, ridge	r_{cal}	D_{equiv}
selection	LOO over seeds, within the condition being scored	none	*two* rates: the condition's and the identity's	LOO for the scalar arm; **none** for the reference arm
evaluation	selected rate, on the held-out checkpoint	whole curve of the row	both on the held-out checkpoint's curve	scalar at its selected rate; reference at its own maximum
reference	that seed's identity, same selection rule	$c + 0.9(m - c)$, m the row's **in-sample** max, c its pre-update mean	none; it is already a difference, and is 0 on identity	median over points of $\max$ of the oracle curve: **in-sample**
aggregation	within-seed mean over the fixed four patterns, then IQM over seeds	as P^{loo}	as P^{loo}	as P^{loo}; relative form divides by $P^{\mathrm{loo}}(\mathrm{id}) - c(\mathrm{id})$ on the same seed
missing	incomplete seed dropped whole, and counted	W_{90}: as P^{loo} (the largest run is defined); ridge: disconnected or edge-touching rows, at either arm, excluded	as P^{loo}	rows with no oracle curve carry none
resampling	seeds; selection done once beforehand	seeds; no selection to re-run	seeds; selection fixed beforehand	seeds; the reference moves with its seed

Table 13: The six choices behind each reported estimand. Boldface marks the two places where a reference level is *not* the held-out quantity, which is a deliberate choice discussed where each estimand is defined. The specification is checked against the frozen implementation by `scripts/e9_estimand_spec.py`, which also asserts the two r_{cal} identities of Section 3.1.

The band, its target and its details. The definition, moved here from Section 3.1:

Definition 3 (Near-optimal band, W_{90} and the ridge)**.** For one checkpoint and condition let c be the mean return over the evaluation samples drawn *before the first gradient update* of that same record, so that no rate can have acted on them. With $m = \max_{\eta \in \mathcal{G}} \bar{A}(\eta)$ over the same record, the band is $\{\eta : \bar{A}(\eta) \geq c + 0.9\,(m - c)\}$. Its largest connected run of grid points is taken; W_{90} is that run's width in $\log_{10} \eta$ and the ridge is its midpoint, with each endpoint placed by linear interpolation of the normalised curve between the last in-band point and its out-of-band neighbour. An endpoint at the first or last grid rate is not extrapolated: the grid value is used and the row is flagged as edge-contacting. c is a *pre-update reference*, not a no-learning counterfactual: ε decays during adaptation, so the return can change with no update, and $m - c$ is the protocol's performance above that reference, not an isolated learning gain.

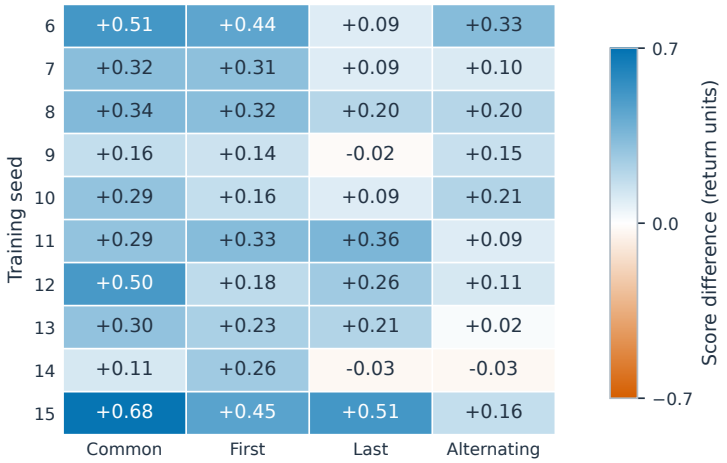

Figure 5: The pre-registered E9 paired contrasts (SGD): each cell is a pattern's LOO-selected held-out score minus the same checkpoint's identity score, at the record's two-decimal precision; the dormant ratio and cumulative spectral rank are bit-identical on all 40 pairs; 3 of 40 negative. The aggregates are Table 2; ten independent checkpoints, not forty observations.

Target. For the band the estimand is $T_W = \mathbb{E}[W_{90}(\bar{A})]$, not $W_{90}(\mathbb{E}[A])$: the maximum, the threshold crossing and the connected run are non-linear, so two conditions with the same expected curve but a different variance of the curve estimate, or a different correlation across rates, can have different T_W. A difference in T_W is therefore a difference in *expected protocol width*, which the shape of the expected curve, the distribution of the adaptation randomness, or both can produce; we do not upgrade it to a statement about the near-optimal range of an underlying curve, one reason the width is a descriptive complement.

Three details. Each is recorded per row. **The normaliser is the in-sample maximum m, not P^{loo}** (m exceeds it on 26 of the 50 layer-uniform and identity rows): W_{90} is a *relative* near-optimality width of one measured curve—curves at different levels can share it—not a bandwidth "around P^{loo}" nor the range of rates meeting a fixed standard. **The threshold is measured above c**: against m alone a level difference would leak into width and location; that c is identical across rates and conditions within a seed is an acquisition gate checked to exact equality on every row. **Exclusion is per endpoint.** P^{loo} is estimable on every row; W_{90} is computed on every row with disconnection and edge contact flagged; the ridge is admitted only if both rows of a contrast have a single connected, edge-free band, a failing contrast is excluded from the ridge alone and reported (under SGD 3 of 90 rows, all anisotropic, are disconnected; none touch an edge), and the ridge inherits the pre-registered 80% complete-case gate per condition.

Undefined outputs and seed coupling. The protocol returns no band when $m \leq c$ and no ridge when the band is disconnected or edge-contacting on either row of a contrast, and an expectation over all realisations must say what it does with them. The width and ridge *contrasts* are defined conditional on *paired* admissibility, $\mathbb{E}[R_h - R_{\mathrm{id}} \mid E_h \cap E_{\mathrm{id}}]$ with E the event that the row's band exists (width) or is single, connected and edge-free (ridge)—not $\mathbb{E}[R_h \mid E_h] - \mathbb{E}[R_{\mathrm{id}} \mid E_{\mathrm{id}}]$, which compares different realisation sets—and the admissible set can differ between interventions, a further reason both are descriptive. The observed frequency of the paired event: the band existed on every row of both optimizers (smallest $m - c$: 1.09 SGD, 1.98 Adam), the ridge pair on 77 of 80 contrasts under SGD and 70 under Adam—frequencies, not a guarantee that the protocol always returns a value; P^{loo} is defined on every realisation. *Seed coupling.* One fixed list of 6 adaptation seeds is used at every rate, every representative and the identity; seed k fixes the environment, exploration, replay-sampling and probe streams as functions of the training seed and k, so the curves entering $\bar{A}$ are paired on their streams across rates and across the two arms. That couples the streams, not

the trajectories, which diverge once the policies differ; the joint distribution across rates that T_W depends on is the one it induces.

In-sample selection disagrees. Holding the aggregation fixed, selecting on the scored checkpoint gives a layer-uniform contrast of $+0.251$ $[0.191, 0.330]$ where (2) gives $+0.230$ $[0.161, 0.299]$; the gap is a difference between two selection-and-evaluation procedures measured on a difference of two conditions, not an estimate of either one's bias.

chooser / predictor reads	held-out MAE (9 rows)	regret of its choice
nothing (constant; picks at random)	0.270 $[0.233, 0.302]$	0.381 $[0.336, 0.435]$
always *common* (best fixed pattern in every fold)	—	0.032 $[0.007, 0.065]$
S: pooled scale descriptor only	0.141 $[0.121, 0.161]$	0.052 $[0.014, 0.099]$
D: dormant ratio, spectral rank	0.173 $[0.150, 0.197]$	0.237 $[0.191, 0.288]$
D + S	0.151 $[0.130, 0.172]$	0.032 $[0.007, 0.065]$

Table 14: A predictor's error falls with the scale descriptor; its choice does not beat a fixed rule. 90 E9 rows, k-NN ($k = 5$), leave-one-checkpoint-out with training labels re-selected without the held-out checkpoint; return units; 95% cluster bootstraps over 10 checkpoints. "Regret": the highest *estimated* score among the checkpoint's nine representatives minus that of the one chosen (ties averaged; the maximum's selection effect enters every row alike). Constructed representatives; S is the pooled scale-sensitive descriptor of Appendix B, which on these representatives separates the interventions by construction. Ridge, the identity rule, same-D rows: Appendix D.

Where the level contrast sits: decomposition and within-checkpoint selection (post-hoc). Designed on 2026-09-13 after every confirmatory result had been read, with the rules below fixed before the script (`scripts/e9_selection_decomposition.py`) was first run on the frozen record and unchanged by its output; the frozen pipeline is imported, not modified. Three quantities, all on the layer-uniform pairs and aggregated as the pre-registered estimators are (per-checkpoint mean over the four patterns, then arithmetic mean and interquartile mean with the frozen 10,000-resample cluster bootstrap in the frozen seed order; the selection is *not* re-run inside resamples, so these intervals are of the selection-fixed kind). *Decomposition.* With M the row's grid maximum of the K-averaged curve, $\Delta P^{\mathrm{loo}} = \Delta M - \Delta(M - P^{\mathrm{loo}})$ exactly; M is a noisy in-sample maximum and not an intrinsic quantity, so the split says only where the observed contrast sits. *Within-checkpoint selection.* The six adaptation seeds are split 3:3 in all 20 ways; on each split the grid rate is the argmax of the mean curve over the three selection seeds and the score is the mean over the other three at that rate; both arms use the same splits; the 20 split contrasts are averaged to one value per checkpoint and pattern and never counted as independent. This removes the cross-checkpoint transfer of η^{-s} at the price of a three-seed selection; it does not refine the grid. *Peak refinement (E13).* Pre-specified on 2026-09-13 in `PROTOCOL_E13_PEAK_REFINEMENT.md` after the decomposition was read, then collected: for the identity and *common* arms of all 10 checkpoints, two rates at $\eta_{\mathrm{peak}}/\sqrt{2}$ and $\eta_{\mathrm{peak}}\sqrt{2}$ around each arm's own $K = 6$ grid maximum (no arm peaked at a grid edge, the pre-specified exclusion), E9's optimizer, horizon, seeds and streams (the pre-update samples are identical to E9's on every row), 240 adaptation runs under acquisition stamp `2e4d003dfbe91ad5`, function change at most 1.1×10^{-15}; the primary statistic is the within-checkpoint 3:3 contrast on each arm's 17-rate grid, read once, with the rule fixed in advance: retained if the interval excludes zero and the point is at least half the 15-grid value (0.188), reduced if below it, inconclusive otherwise. Read: $+0.359$ $[0.251, 0.476]$ (interquartile mean $+0.324$ $[0.237, 0.487]$), **retained**; an added rate was selected on 35% of identity splits and 43% of common splits. Two limits are intrinsic to the design. The refinement positions were chosen from each arm's $K = 6$ mean curve, so the evaluation seeds of every 3:3 split had already taken part in placing the candidate rates: the split separates the final rate selection from its evaluation, not the construction of the grid, and E13 is a robustness analysis on a data-chosen local grid rather than a fully independent evaluation (that would need a fixed grid scored on new seeds, or grid placement done inside the selection seeds). And "retained" is the pre-specified rule, not an equivalence test on the effect's size: that the point estimate kept about 95% of its 15-grid value is descriptive. It is a local-refinement statement for one pattern; it does not test curve shape, the other patterns, or a continuous shift, and the recalibration question of Appendix E is narrowed, not closed.

Independent evaluation seeds (E16). Pre-specified in `PROTOCOL_E16_INDEPENDENT_EVAL.md` to remove the first E13 limit, with its scope fixed before any row existed: the identity and *common* representatives of the ten checkpoints (asserted bit-identical to E9's on the recorded diagnostics), each arm's 17-rate grid taken from the E13 plan and **fixed**, E9's optimizer, horizon and warm-up, and adaptation seeds 6–11, never used in E9, E13 or E14 (E9's stream rule, so environment, exploration, replay and probe streams are all new); 2,040 adaptation runs under acquisition stamp `4d103930ea30d6ec`, function change at most 1.1×10^{-15}. Analysis, fixed: per arm, the rate is the argmax of the $K = 6$ mean curve on seeds 0–5 *only* (E9 and E13 records); the score is the arm's mean over seeds 6–11 at that rate; the contrast is *common* − identity per checkpoint, aggregated by the arithmetic mean with the frozen cluster bootstrap and by the interquartile mean; retained if the interval excludes zero and the point is at least 0.188. Read once: $+0.317$ $[0.228, 0.402]$, interquartile mean $+0.325$ $[0.204, 0.420]$, **retained**; the ten per-checkpoint contrasts, in frozen seed order 6–15, are $0.391, 0.543, 0.296, 0.243, 0.154, 0.091, 0.415, 0.427, 0.177, 0.429$, all positive (a post-hoc addendum to the protocol, fixed before its script ran: the exact two-sided paired sign test and the exact sign-flip randomisation test of the mean over all 2^{10} assignments both give p =0.002, `scripts/e16_sign_test.py`; a check on the direction that does not depend on the percentile bootstrap, reported beside it); the selected rate was one of the two E13-added rates on 30% of identity arms and 60% of common arms. A positive contrast is thus re-observed on the same ten checkpoints under new adaptation randomness; the estimate $+0.317$ is reported alongside the earlier within-checkpoint values $+0.376$ (15 rates) and $+0.359$ (17 rates), and no agreement of size is claimed (overlapping intervals are not an equivalence test); the E9 cross-checkpoint selection contrast ($+0.2359$) is a different estimand and is not compared. Two descriptive companions, both pre-specified as estimators in the E16 protocol: the 3:3 within-checkpoint contrast computed on the new seeds alone (grid placement and evaluation fully disjoint from the split) is $+0.379$ $[0.324, 0.429]$ (its reading as a per-arm recalibration check, with the coverage gate, is the post-hoc analysis E19 of Appendix E); and the decision regret of Section 7's two fixed rules re-scored on the new seeds is 0.000 for "always *common*" (it is the better arm on every checkpoint) and 0.317 $[0.228, 0.402]$ for "always the identity". Scope: the same ten trained checkpoints under new adaptation randomness, two arms, one pattern; not a replication on new checkpoints, a new task or a new architecture, and not a test of any predictor.

Training stage (E22; moved here from the body). Pre-specified in `PROTOCOL_E22_E23_STAGE_AND_EVAL.md` after E21 was read: the identity and *common* arms of the ten runs at their untrained snapshot (step 0, task 1) and their mid-training snapshot (step 150,000, after three switches, task 4), E9's grid, seeds, horizon and selection rule, float32; 3,600 runs, stamps `ad6a104b97fc127f` and `c86acd4937810412`. Read once: at initialisation *common* − identity is -0.021 $[-0.042, -0.002]$ (interquartile mean -0.022 $[-0.041, 0.002]$), the identity's selected level 0.48 above a reference of 0.10; at mid-training $+0.601$ $[0.525, 0.671]$ (interquartile mean $+0.619$ $[0.511, 0.686]$), identity level 2.03; at the final checkpoint (E9) $+0.351$ $[0.255, 0.449]$, level 2.12. Paired by training run: mid − untrained $+0.622$ $[0.540, 0.700]$, final − untrained $+0.372$ $[0.268, 0.481]$, final − mid -0.250 $[-0.349, -0.149]$. Read: the contrast was small and negative at initialisation (the arithmetic mean's interval excludes zero; the interquartile mean's does not), larger and positive at the intermediate snapshot, and smaller but still positive at the final snapshot; the paired comparisons establish a non-monotonic pattern across the three evaluated snapshot–task settings. Because the adaptation task changes with the snapshot (tasks 1, 4 and 7), the pattern does not isolate training stage as its cause; the low returns at initialisation further limit interpretation of the small initial contrast, and c is a pre-update reference, not a no-learning counterfactual, so it does not measure how much was learned. Two trained stages do not make a curve, the snapshots of one run are not independent samples, and no diagnostic change was paired with the sequence. The protocol's provisional sentence for this outcome ("the effect grows with training") was not adopted, for these reasons.

Greedy evaluation (E23). Pre-specified with E22: the identity and *common* arms of the ten final checkpoints on the fixed 17-rate E16 grids, adaptation seeds 0–5, float32, with 5 greedy episodes ($\epsilon = 0$, a separate environment copy seeded per adaptation seed, never written to the replay buffer) at each of the 20 evaluation points, their mean recorded beside the online return; the online curve is pinned byte-identical (`tests/test_dtype_path.py`); 2,040 runs; the acquisition stamp hashes each checkpoint's own grid and is therefore one per checkpoint. Read once, leave-one-checkpoint-out with E9's selection rule: (i) online score at the online-selected rate, *common* − identity $+0.296$ $[0.182, 0.414]$; (ii) greedy area at that same rate $+0.519$ $[0.405, 0.635]$ (interquartile mean $+0.516$

$[0.375, 0.653]$); (iii) greedy area at the rate selected on the greedy curve $+0.495$ $[0.362, 0.630]$; the identity's greedy level 2.58 against its online 2.18. All ten per-checkpoint greedy contrasts are positive. Read: the contrast persists when the learned policies are evaluated greedily in separate evaluation episodes, including at rates selected on the online score, so it is not confined to the exploration-inclusive return used during adaptation. This evaluation does not separate optimisation effects from exploration-mediated differences in the training data (exploration still shapes the policy, the collected data and the replay buffer), and the outcome is greedy performance integrated over the adaptation horizon rather than a terminal return, which was not recorded. The collection uses the original adaptation seeds 0–5 and grids whose two added rates were placed with those seeds, so its leave-one-checkpoint-out reading does not make grid placement independent of the evaluation seeds (E16 does); it is not a third independent-seed re-verification. The five episodes reduce per-run measurement error, not the number of independent checkpoints. The protocol's provisional sentence ("a difference in what was learned") was narrowed to this on review.

Dose–response (E25). Pre-specified in `PROTOCOL_E25_E26_DOSE_AND_FIXED_DATA.md` after the 26th review, when the no-new-experiments rule was lifted for this and the next collection: the *common* rescaling at signed log-magnitudes $m \in \{\pm 0.1, \pm 0.2, \pm 0.35, \pm 0.5\}$ on the ten checkpoints, E9's grid, adaptation seeds 0–5, horizon and code path (7,200 runs, stamp `fe32ebeb7e3a3c72`); the natural scale fixed from Appendix F before the rows existed ($|m| = 0.35$: the first hidden layer's mean shift of q over a run, 0.33; $|m| = 0.1$: the second half, at most 0.11), and the sentences for each outcome fixed with it. *Fidelity gate, read first.* The identity was re-collected on one checkpoint and compared with E9's frozen curves: 48/90 per-seed values reproduce E9 exactly and the rest diverge (worst difference 0.89 return units; pre-update samples and streams identical; $K = 6$ mean curves within 0.05 except at the largest rates)—the float32 divergence of Appendix E, between the E9 acquisition environment and the present one; not function equality (change 0 on the identity), not the pre-update return, not the diagnostics. Under the pre-fixed failure branch the identity was re-collected on all ten checkpoints (900 runs, stamp `25588a8126391590`) and is the reference for every magnitude. Accounted for afterwards (`scripts/e25_gate_final.py`): the present environment reproduces itself exactly (the seed-6 identity collected twice, 90/90 per-seed values bit for bit), so both arms of every contrast share one acquisition; against E9 it reproduces 530/900 per-seed values exactly, worst difference 1.09, $K = 6$ mean curves within 0.30 (0.17 excluding the two largest rates), pre-update samples identical throughout—the cross-environment gate failed and is reported; the within-acquisition gate passed. E9's own $+0.7$ contrast is quoted within its own acquisition (the E9 *common* row against the re-collected identity gives $+0.308$ $[0.226, 0.383]$, not pooled). Read once: $m = -0.5$: -0.321 $[-0.391, -0.250]$; -0.35: -0.251 $[-0.300, -0.201]$; -0.2: -0.190 $[-0.280, -0.099]$; -0.1: -0.093 $[-0.168, -0.016]$; $+0.1$: -0.000 $[-0.069, 0.074]$; $+0.2$: $+0.128$ $[0.075, 0.193]$; $+0.35$: $+0.173$ $[0.104, 0.242]$ (interquartile mean $+0.164$ $[0.095, 0.248]$; 9 of ten positive); $+0.5$: $+0.267$ $[0.182, 0.357]$; E9's $+0.7$: $+0.351$ $[0.255, 0.449]$. Diagnostics equal to the identity's on 81 of 81 rows; function change at most 1.3×10^{-15}. Under the rules: both intervals at $|m| = 0.35$ exclude zero; at $|m| = 0.1$ the negative direction is detected and the positive is not (no equivalence margin); at the evaluated negative magnitudes the mean contrast is negative and more negative at larger $|m|$; the eight point estimates increase with m except between 0 and $+0.1$ ($m = 0$ a definition, $+0.7$ a separate collection), so "ordered in m" is not written, no model is fitted and no optimum or slope of a continuous response is inferred; the natural scale is a size reference for one coordinate, not the drift's direction or its frequency in training; the per-block SGD rates e^{-2m} on W_1 and the hidden biases and e^{2m} on W_4 (0.50 and 2.01 at $m = 0.35$) are a reading.

Adaptation on fixed data (E26). Pre-specified with E25: for each checkpoint and data seed $s \in \{0, \ldots, 5\}$, 4,000 training and 1,000 held-out transitions on task 7 collected once under a uniformly random policy; the identity and *common* representatives (asserted bit-identical to E9's on the recorded diagnostics) fitted by plain SGD (momentum 0.9, fresh state) for 3,000 updates at each rate of E9's grid, batch 32 drawn by one index sequence shared by both arms and all rates, float32, no environment stepped; the held-out TD loss recorded at 20 points and a run's score its negated mean, so E9's selection rule applies unchanged (the rate selected on the other checkpoints' curves, scored on the held-out checkpoint's own held-out set; the reader's rule reproduces every frozen E9 P^{loo} exactly).

Amendment, before any row existed (2026-09-15, with E25 at 0 of 81 rows): a fixed-target condition—the target network frozen at the shared initial function, so both arms minimise one

objective—collected in the same runs beside the per-arm-target condition (each arm's target synced every 1,000 updates as in the online protocol), because a per-arm target lets the two objectives diverge after the first sync; 3,600 runs, stamp `c8d5991b1743f5da`; the amendment does not have the status of the original design. Read once, fixed target first: pre-update held-out loss equal across rates and arms (1.2×10^{-7}); fixed target $+0.108$ $[0.085, 0.132]$ (interquartile mean $+0.103$ $[0.079, 0.130]$), ten of ten positive, identity's selected-rate held-out loss 0.363 against *common*'s 0.255, final held-out loss contrast $+0.060$ $[0.041, 0.081]$; per-arm target $+0.123$ $[0.095, 0.154]$, final $+0.105$ $[0.077, 0.138]$, ten of ten. Under the rules: the conditioning difference is present without online data feedback. Limits: random-policy data; a TD objective with bootstrapped targets (a regression to fixed targets in the fixed-target condition), not a policy-quality measure; the same ten checkpoints; an optimisation difference in the absence of data-collection differences, not by itself an explanation of the online return difference.

Retrospective analyses of existing records (E27–E30). The analysis specifications were documented in `PROTOCOL_E27_E29_POSTHOC_ANALYSES.md` and `PROTOCOL_E30_BLIND_RATE_REGRET.md` after the underlying experimental records had already been examined and before the corresponding analysis scripts were executed. These analyses are post-hoc; fixing their specifications before execution does not make them confirmatory.

E30, shared against state-specific rate selection. The action set is the *global* support: the 15 learning rates present as the same floating-point values in both states' records at *every* checkpoint, which turns out to be exactly E9's grid. Each E16 arm carries its own two refinement rates, so the arms' 17-rate grids differ on eight of the ten checkpoints; on the other two the arms agree and that checkpoint's pairwise intersection is 17, but those two extra rates are *excluded* from the primary so that one action set serves every checkpoint (they are a labelled sensitivity, not merged in). States count as indistinguishable only through readouts asserted equal bit-for-bit on every analysed pair (11 of the audited formulas; the rule admits a readout iff record E17 reports it identical on 40 of 40 layer-uniform pairs, which excludes the absolute-cut feature rank and the gradient norm). Within each 3:3 split of the six adaptation seeds, each state's rate is selected on three seeds and a shared rate on the same three from the states' mean curve; both are scored on the other three. Splits are averaged within a checkpoint before anything else, so the independent unit is the checkpoint and the twenty overlapping splits are never counted as replicates. *Primary* (E16, seeds 6–11): $+0.004$ $[-0.016, 0.027]$. *Secondaries:* the two states choose the same rate on 0.325 $[0.155, 0.530]$ of splits, and carrying the other's rate costs $+0.023$ $[-0.016, 0.070]$. On E9's five layer-uniform states the same quantities are $+0.014$ $[-0.016, 0.045]$, 0.272 $[0.205, 0.336]$ and $+0.093$ $[0.059, 0.139]$; E9's six seeds placed the refinement and selected the frozen P^{loo}, and the protocol fixed before the run that this record is read after the primary and does not overturn it. What is computed is the held-out advantage of one selection procedure over another; it is *not* a lower bound on the regret of a rule that reads these diagnostics, because the two selectors draw on different numbers of curve observations and their selection losses enter the expectation with opposite signs. No argmax fell on a grid endpoint.

E27, noise separation on E16: with each arm's rate as E16 selected it on seeds 0–5 and held fixed, the $K = 6$ new-seed paired differences d_{sk} give $\hat{\delta}_s = \frac{1}{K}\sum_k d_{sk}$ (E16's value, reproduced exactly) and its evaluation-noise variance $\hat{\sigma}_s^2 = s_d^2/K$ with s_d^2 the sample variance with denominator $K - 1$; under independent evaluation seeds $\hat{\delta}_s^2 - \hat{\sigma}_s^2$ is unbiased for δ_s^2. The per-checkpoint corrected values are averaged without truncation; the RMS $\sqrt{\max(0,\ \mathrm{mean}_s\hat{\delta}_s^2 - \mathrm{mean}_s\hat{\sigma}_s^2)}$ is a transformed estimator, not itself unbiased. Its interval resamples the ten checkpoints with replacement (10,000 draws, seed 0), each carrying its $(\hat{\delta}_s, \hat{\sigma}_s^2)$, and recomputes the correction and the square root in every draw (percentile 2.5–97.5). Mean observed squared contrast 0.120, mean noise variance 0.014 (an estimated 11.4% of the observed mean square—not a share of adaptation performance, of between-checkpoint differences, or of anything about natural training), noise-corrected RMS 0.326 $[0.232, 0.404]$, mean noise sd 0.111, smallest $|\hat{\delta}_s|/\hat{\sigma}_s$ 1.04.

E28, predictor baselines: E14's nested leave-one-checkpoint-out, k-NN and ridge, with the intervention's identity (one-hot of the nine representatives), every audited formula's reported value on the representative (the 15 values of record E17) and the per-layer mean of q (Appendix B) as features. Held-out MAE: identity alone 0.129 $[0.112, 0.146]$ (k-NN) and 0.130 $[0.114, 0.144]$ (ridge); with the 15 diagnostics 0.140 $[0.128, 0.152]$ and 0.157 $[0.133, 0.187]$; with the per-layer scale added 0.139 $[0.129, 0.150]$ and 0.148 $[0.130, 0.170]$; the 15 diagnostics alone 0.167 $[0.147, 0.187]$ and

0.200 $[0.159, 0.258]$ with regret 0.222 $[0.167, 0.283]$ and 0.119 $[0.048, 0.200]$; every set containing the identity, and the best fixed rule chosen inside each fold (*common* in all ten), leaves regret 0.032 $[0.007, 0.065]$. The primary paired difference, MAE with the per-layer scale minus without, is -0.000 $[-0.003, 0.002]$ (k-NN) and -0.009 $[-0.019, -0.000]$ (ridge). Read: with ridge the MAE decreases by 0.009 and the interval's upper endpoint is only 4×10^{-5} below zero, so the pre-fixed criterion is met at the boundary; with k-NN no improvement is detected; regret is unchanged. The comparison is incremental within the set ID + D_{full}, whose ridge error (0.157) the scale partly recovers (0.148) without reaching the identity alone (0.130): an incremental predictive improvement for that feature set and predictor, not evidence for or against checkpoint-specific information in itself, and not a practical selection gain; a worse fit with more features may also reflect the distance metric of k-NN or ridge's shrinkage (Table 19). The rebuild of the representatives was first committed in lexicographic run order; the per-representative equality assertion against E9 failed on the first anisotropic one before any metric was computed, the order was corrected to E17's, and the script ran once.

E29, first-order prediction: at θ_0, on each E26 training set regenerated from its seeds, the fixed-target objective's full-training-set gradient g is split by parameter block; the first-order approximation to the decrease of that training objective in one step of an arm with per-block factors p_b at rate η is $\eta \sum_b p_b \|g_b\|^2$ ($p_b = e^{-2m}$ on W_1 and the hidden biases, e^{2m} on W_4, 1 elsewhere; checked against the gradient at the rescaled point, relative error 5×10^{-7}). It is a training-gradient quantity used as a proxy for the held-out and online responses: the first-order held-out decrease would be $\eta\, g_{\text{held}}^{\top} P g_{\text{train}}$, and the two finite-sample gradients are not equal. The identity's energy lies 72% on W_1. Against E26's fixed-target contrast, each arm at its E26-selected rate (identity 2×10^{-3} on every checkpoint, *common* 4×10^{-3} on nine and 8×10^{-3} on one): sign agreement 9/10, Spearman 0.09 $[-0.57, 0.72]$; the constant predictor 10/10, the selected-rate ratio alone 10/10 with Spearman 0.06 $[-0.35, 0.53]$. Against the E25 signs at the E25-selected rates, the predicted sign is $+$ at every negative m (the effective first-layer rate $e^{-2m}\eta$ rises) and $-$ at every positive m; the observed point estimate has the opposite sign at 7 of 8 magnitudes (1/8; "sign of m" 7/8). This count is descriptive: signs are those of point estimates, the point at $m = +0.1$ has an interval including zero, and the eight magnitudes share the same checkpoints and evaluation conditions, so they are not eight independent observations. The pre-fixed sentence "a first-order per-block reading predicts the direction" is not adopted. What the record shows is that this initial, local score predicts the multi-update held-out loss and the online return no better than simple baselines; curvature, gradient change along the trajectory, momentum and the data distribution all lie between the two, so the failure of this predictor does not exclude the per-block conditioning reading.

Predictive-utility check (E14). Pre-specified in `PROTOCOL_E14_PREDICTION.md` after everything above was read, then amended once (section 6 of the protocol, rules fixed before the amended script ran) on a reviewer's three points; on the 90 E9 rows and nothing new. The row's P^{loo} is predicted from the two recorded diagnostics (D: dormant ratio, cumulative spectral rank) or from D plus the pooled scale descriptor (S: the mean and population standard deviation of $\log\|w_i^{\text{out}}\|/\|w_i^{\text{in}}\|$ over all hidden units, Appendix B), by k-nearest-neighbour ($k = 5$) and ridge ($\lambda = 1$) regression on features standardised with the training fold's mean and standard deviation, all nine rows of a checkpoint held out together. *Labels:* because a row's frozen P^{loo} is selected on the other nine checkpoints, the held-out checkpoint's curves enter the training labels; the primary analysis therefore re-selects every training row's rate over the eight other training checkpoints only (nested labels), the held-out rows keeping their frozen value; the frozen-label variant is `runs/e14_prediction.json` and differs by at most 0.004 in any figure. Metrics are averaged over the ten folds (95% cluster bootstraps). Held-out MAE in return units, nested labels: k-NN on D 0.173 $[0.150, 0.197]$, on D+S 0.151 $[0.130, 0.172]$ (difference -0.022 $[-0.038, -0.006]$); ridge on D 0.168 $[0.152, 0.186]$, on D+S 0.154 $[0.140, 0.168]$ (difference -0.014 $[-0.024, -0.000]$); the constant predictor 0.270 $[0.233, 0.302]$. *Where the change is:* on the five same-D rows of each checkpoint the k-NN error falls from 0.172 $[0.147, 0.198]$ to 0.122 $[0.093, 0.154]$ (difference -0.050 $[-0.079, -0.023]$; ridge -0.020 $[-0.034, -0.007]$), on the four anisotropic rows it does not ($+0.013$ $[-0.005, 0.032]$; ridge -0.006 $[-0.020, 0.011]$); the pairwise ordering error on the five same-D rows is 0.5 by construction for D (every pair tied) and 0.21 $[0.14, 0.28]$ for D+S, over all 36 pairs 0.27 $[0.22, 0.33]$ against 0.19 $[0.14, 0.23]$. *Tie-forced error:* the MAE of the best single value on the five same-D rows is 0.116 $[0.092, 0.143]$, which contributes $5/9$ of it, 0.064 $[0.051, 0.079]$, to the nine-row MAE; it is the empirical constant-predictor error on observed scores, not the irreducible error with respect to the expected response, since evaluation noise alone spreads the five observed scores. Read as fixed

in advance: the scale descriptor removes part of the same-D error and nothing on the anisotropic rows; this is a prediction-and-ranking evaluation, not a regret after a chosen intervention; the descriptors include the intervention coordinates and the split is over ten checkpoints, so nothing here estimates predictive utility on training-produced representatives; and the diagnostics are read on the current-task probe while the target is the next task's response, so the predictive shortfall of D mixes missing task information with the discarded scale (the invariance itself holds under any probe; a next-task-probe reading of D was not fitted). *Second amendment (protocol section 7).* Two further readings of the same rows, rules fixed first. *Leave-one-pattern-out:* with every checkpoint's row of one pattern held out and the predictor fitted on identity and the other seven patterns (frozen labels, not nested), the k-NN error on held-out layer-uniform patterns is 0.144 $[0.121, 0.167]$ (D) against 0.112 $[0.093, 0.131]$ (D+S), difference -0.032 $[-0.050, -0.013]$ (ridge -0.025 $[-0.037, -0.014]$), and on held-out anisotropic patterns 0.227 $[0.191, 0.262]$ against 0.234 $[0.197, 0.275]$ (difference $+0.007$ $[-0.021, 0.037]$; ridge -0.014 $[-0.030, 0.007]$): the descriptor's value is not specific to the layer-uniform pattern it was fitted on, and it does not carry to the anisotropic points. *Decision regret:* in the nested leave-one-checkpoint-out folds each predictor chooses the representative with the highest predicted score among the nine, and regret is the checkpoint's highest estimated score minus the chosen one's (ties: the mean over tied candidates). Choosing at random (the constant predictor) leaves 0.381 $[0.336, 0.435]$; D leaves 0.237 $[0.191, 0.288]$ (k-NN) and 0.194 $[0.152, 0.237]$ (ridge); D+S leaves 0.032 $[0.007, 0.065]$ with either, a paired difference of -0.205 $[-0.271, -0.141]$ (k-NN) and -0.162 $[-0.216, -0.110]$ (ridge). A decision among constructed representatives of one checkpoint, not an intervention chosen in training. *Third amendment (protocol section 8): baselines.* On the same folds, always choosing the identity leaves 0.382 $[0.300, 0.469]$; always choosing *common*, 0.032 $[0.007, 0.065]$; the best fixed pattern chosen inside each training fold (it is *common* in all ten folds), the same; a predictor reading S alone, 0.052 $[0.014, 0.099]$ (k-NN; ridge 0.032 $[0.007, 0.065]$), with held-out MAE 0.141 $[0.121, 0.161]$ (k-NN; ridge 0.156 $[0.141, 0.173]$). The paired difference between D+S and the always-common rule is exactly zero in every fold. Read as fixed in advance: on these rows the scale descriptor's regret gain is the knowledge of which intervention is generally better, not a checkpoint-specific reading; S is a pooled scale-sensitive descriptor, so on constructed representatives it identifies the intervention by construction, and whether it reads anything about training-produced state is the untested part. *Addition (post-hoc, after these numbers were read):* the numerical gain is S's as well—the paired held-out MAE of D+S minus S alone is $+0.010$ $[-0.007, 0.031]$ (k-NN) and -0.002 $[-0.015, 0.012]$ (ridge): on these rows and with these two predictors no additional predictive gain from D was detected; no equivalence margin was fixed, so this is an absence of detected gain, not a finding of none. *Threshold.* The width contrast at 0.8, 0.9 and 0.95 of $m - c$, all three reported, none adopted after the fact. Table 15 gives the results with the absolute scores that set the scale.

layer-uniform contrast, return units	arithmetic mean	interquartile mean
ΔP^{loo} (pre-registered)	$+0.2359$ $[0.1740, 0.2999]$	$+0.230$ $[0.161, 0.299]$
ΔM, grid maxima	$+0.260$ $[0.202, 0.320]$	$+0.251$ $[0.191, 0.330]$
$\Delta(M - P^{\mathrm{loo}})$, selection transfer	$+0.024$ $[0.002, 0.046]$	$+0.025$ $[-0.003, 0.054]$
within-checkpoint 3:3 selection	$+0.262$ $[0.194, 0.327]$	$+0.264$ $[0.185, 0.338]$
common only, 15-rate grid	$+0.376$ $[0.277, 0.468]$	$+0.394$ $[0.260, 0.499]$
common only, 17-rate grid (E13)	$+0.359$ $[0.251, 0.476]$	$+0.324$ $[0.237, 0.487]$
ΔW at 0.8 (dex)	$+0.19$ $[0.06, 0.29]$	$+0.23$ $[0.10, 0.30]$
ΔW_{90} (pre-registered)	$+0.1541$ $[-0.0120, +0.2835]$	$+0.190$ $[0.04, 0.30]$
ΔW at 0.95 (dex)	$+0.17$ $[0.09, 0.25]$	$+0.16$ $[0.08, 0.27]$

Table 15: Post-hoc decomposition of the layer-uniform level contrast and the threshold sensitivity of the width, 10 checkpoints, 95% selection-fixed cluster-bootstrap intervals. Scale: the identity's selected score averages 2.12 (its grid maximum 2.15, pre-update reference 0.08) and the layer-uniform rows' 2.35; the within-checkpoint identity score averages 2.09. In the within-checkpoint analysis the two arms' selected rates differ on 67% of splits.

What the return is. The returns entering A are those of the training episodes run during adaptation, with the exploration policy in force; at each of the 20 evaluation points the mean of the last ten finished episodes is recorded (fewer if fewer have finished, 0 if none), so a trailing window can span a different number of environment steps under conditions with different episode lengths. The four evaluations before the first update are identical across a checkpoint's conditions, since no rate

has acted; they dilute every contrast equally and bias none, and their equality is the acquisition gate of Section 3.1. A no-update baseline run under the same exploration schedule would be common to both arms of a checkpoint and cancel in the paired contrast; its absence therefore does not explain a contrast, but it does leave the reading of the score as "amount learned" unsupported, which is why the greedy evaluation (E23) and the fixed-data collection (E26) are reported beside it.

The default-rate gap. With η_{id}^{-s} the rate selected on the identity condition of the other checkpoints, $r_{\mathrm{cal}}(\theta_s, p) = P^{\mathrm{loo}}(\theta_{s,p}) - \bar{A}(\eta_{\mathrm{id}}^{-s}; \theta_{s,p})$: the held-out cost of using the identity's rate instead of the condition's own. It is not an independent component: on any gauge condition $r_{\mathrm{cal}} = \Delta P^{\mathrm{loo}} - \Delta\bar{A}(\eta_{\mathrm{id}}^{-s})$ identically, and it is exactly zero on the identity row because both terms are then the same number; both identities are verified to machine precision by `scripts/e9_estimand_spec.py`. Since out-of-sample selection is not guaranteed optimal on the held-out checkpoint it can be negative, and it is on 19 of 80 gauge rows.

Reading the number, with the matching functional. The right-hand side of (3) is an unweighted mean over seeds of the within-seed mean over patterns. The headline $+0.230$ $[0.161, 0.299]$ is an *interquartile* mean, a different functional, so half of it bounds nothing; being non-linear, the IQM of noisy per-seed contrasts does not estimate the IQM of the expected contrasts even under mean-zero evaluation noise: it is the pre-registered robust aggregate, not the bridge's input, which is $\mathrm{mean}_s \mathrm{mean}_j \delta_{sj} = +0.2359$ ($[0.1740, 0.2999]$ selection fixed, $[0.1707, 0.3173]$ selection re-run; Table 2). **The plug-in estimate of the lower bound is about** 0.118. **Mapping the post-hoc, ID-excluding selection-refit percentile interval through the bound gives a lower endpoint of about** 0.085 (0.1179 and 0.0853 in Appendix D); that interval is used because the pre-registered bootstrap holds the selection fixed, it carries no claim of nominal coverage, and its third decimal is sensitive to the bootstrap's Monte Carlo variation (Appendix D). Neither is an error nor a bound known to hold: the first is a plug-in estimate of a lower bound on the mean, over seeds and patterns, of the worst-case absolute error of any predictor reading only D; the second is the lower endpoint an interval gives that bound; neither is read off a fitted predictor—the post-hoc predictor evaluation is E14 below—and a given predictor's actual error can be far larger. The bound concerns the expected response T under Definition 1's protocol, not best attainable adaptation, a correlation in ordinary training or a population guarantee; more data could move it toward zero.

The interquartile mean and the expected contrast. Writing an observed per-seed contrast as $\widehat{\delta}_s = \delta_s + \epsilon_s$ with δ_s the expected contrast under the protocol and ϵ_s the evaluation noise, $\mathbb{E}[\mathrm{IQM}(\widehat{\delta})] \neq \mathrm{IQM}(\delta)$ in general even when $\mathbb{E}[\epsilon_s] = 0$, because the interquartile mean trims by rank. The pre-registered IQM and its cluster bootstrap are therefore a robust aggregate of the observed contrasts, not a precise inference about the IQM of the expected contrasts; the arithmetic mean, which is linear in the noise, is the estimate of the mean expected contrast and the quantity the bridge (3) consumes, with the selection-refit interval as its uncertainty under the limits stated below.

Why there are two intervals on the mean, and which one the conservative figure is read from. The pre-registered cluster bootstrap resamples per-seed contrasts computed once with η^{-s} selected on the full seed set, so it propagates none of the selection step's variability and leaves the per-seed values dependent, each having been scored at a rate chosen on the others. Re-running the leave-one-out selection inside every resample gives $[0.1707, 0.3173]$ for the mean against $[0.1740, 0.2999]$ with the selection held fixed. One detail of that re-run decides its answer. A resample draws checkpoints with replacement, so it usually contains the scored checkpoint more than once; holding out only one copy leaves the others inside the data its own rate is selected on, and here that raised the lower end to 0.1816—an observation on these data, not a general property of resampling. Holding out every copy of the checkpoint's identity gives the interval quoted, and 0.0853 is half its lower end, truncated rather than rounded, because a lower bound rounded upward asserts a floor the premise does not support. The interval remains a post-hoc, selection-refit percentile interval: removing a leak does not establish nominal coverage.

Bootstrap specification, and the Monte Carlo spread of the endpoints. Every interval on the E9 and E10 condition contrasts and on the two bound figures is a percentile interval from $B = 10{,}000$ resamples of the 10 training checkpoints with replacement, drawn as index draws

from `numpy.random.RandomState(0)` (`randint(0, n, n)` per resample), the statistic recomputed on each resample and the interval read at the 2.5th and 97.5th percentiles under numpy's linear-interpolation rule; point estimates are computed on the original ten values. With index draws the finite-B endpoints depend on the order in which the ten checkpoints are listed—the same draw sequence picks different checkpoints when the list is permuted—so the frozen order (run directories sorted as strings) is part of the specification; drawing checkpoint *identities* from a canonical order instead reproduces the frozen endpoints and is invariant to the input order (`scripts/e9_bootstrap_mc_check.py`, a post-hoc check that leaves the frozen numbers untouched). The same check runs 21 independent realisations of the $B = 10{,}000$ bootstrap: the pre-registered interquartile-mean interval's endpoints span 0.159–0.163 and 0.299–0.307, the fixed-selection mean's 0.172–0.176 and 0.300–0.305, the ID-excluding selection-refit mean's 0.168–0.173 and 0.316–0.321, and the mapped bound figure 0.0841–0.0864 (0.0856 at $B = 100{,}000$). The reported values are the frozen realisation; their third decimal is sensitive to this variation, and the spread is about the Monte Carlo approximation only—it says nothing about coverage or about there being ten independent seeds.

Mapping an interval through the bound. A bound on $|\cdot|$ is read off the whole interval, not one endpoint: it is $\min(|\ell|, |u|)/2$ when $[\ell, u]$ excludes zero and 0 when it does not. Halving $|\ell|$ alone would report a positive floor from an interval that admits zero, which is exactly the case for W_{90}'s signed mean over seeds, $+0.1541$ $[-0.0120, +0.2835]$ (its interquartile mean $+0.190$ $[0.04, 0.30]$ excludes zero; the patterns partly cancel in the signed mean), and is why no aggregate bound is claimed from the width; the unsigned alternative $\mathrm{mean}_{s,j}|\delta_{sj}|$ would not cancel and is not promoted after seeing the signed one lose its interval.

Four questions, not one (moved from Section 7). Whether a diagnostic responds to a rescaling (the audit), distinguishes different rescalings (the same-value pairs; the per-layer q vector of Appendix B), predicts the expected response (the limited predictor evaluation above) or improves a choice (no gain over "always *common*") are separate questions; the feature rank's moving answers the first only. The scores of Wang et al. (2026) and Baveja et al. (2025) were not audited and are not assumed to follow Table 9.

E THE SECOND OPTIMIZER, THE RECALIBRATION CHECK, AND THE EQUIVARIANT REFERENCE

condition	ΔP^{loo}	ΔW_{90} (dex)	Δridge (dex)
layer-uniform (family)	$+0.230$ $[0.161, 0.299]$	$+0.19$ $[0.04, 0.30]$	-0.00 $[-0.07, 0.07]$
alternating	$+0.136$ $[0.067, 0.194]$	$+0.35$ $[0.16, 0.44]$	-0.15 $[-0.23, -0.05]$
common	$+0.339$ $[0.238, 0.449]$	$+0.11$ $[-0.16, 0.24]$	$+0.18$ $[0.13, 0.28]$
first	$+0.270$ $[0.198, 0.361]$	$+0.37$ $[0.17, 0.53]$	$+0.19$ $[0.09, 0.34]$
last	$+0.158$ $[0.055, 0.284]$	$+0.02$ $[-0.16, 0.12]$	-0.25 $[-0.32, -0.19]$
anisotropic (family)	-0.214 $[-0.382, -0.109]$	$+0.14$ $[-0.02, 0.28]$	no estimate (7/10)
$z = 0.70$, dir. 0	-0.002 $[-0.141, 0.105]$	$+0.34$ $[0.13, 0.46]$	-0.15 $[-0.24, -0.05]$
$z = 0.70$, dir. 1	$+0.163$ $[-0.061, 0.333]$	$+0.23$ $[-0.10, 0.47]$	-0.18 $[-0.24, -0.11]^{n=9}$
$z = 1.80$, dir. 0	-0.683 $[-0.867, -0.452]$	-0.17 $[-0.34, 0.14]$	-2.31 $[-2.61, -1.86]^{n=9}$
$z = 1.80$, dir. 1	-0.390 $[-0.583, -0.185]$	$+0.08$ $[-0.22, 0.37]$	-2.22 $[-2.39, -2.03]^{n=9}$

Table 16: Per-condition contrasts under SGD against the same checkpoint's identity representative: interquartile mean over 10 seeds with a 95% cluster-bootstrap interval, the pre-registered estimator, every cell generated from the recorded summary. An interval covering zero is printed, never "unchanged"; superscript n marks fewer usable seeds; a family row below the 80% complete-case gate carries no estimate. The family rows are the pre-registered contrasts; the per-pattern rows are prespecified point-wise contrasts with no multiplicity correction. The width and location columns are descriptive (Appendix D); the three columns are drawn per condition in Figure 8. Adam: Table 17.

The direct optimizer contrast. $D(s, g) = \Delta_{\mathrm{SGD}}(s, g) - \Delta_{\mathrm{Adam}}(s, g)$ on the same checkpoint and gauge point, so only the adaptation optimizer differs; a cell enters only if usable under both

condition	ΔP^{loo}	ΔW_{90} (dex)	Δridge (dex)
layer-uniform (family)	$+0.012\ [-0.064, 0.051]$	$+0.13\ [-0.02, 0.31]$	$+0.03\ [-0.03, 0.08]^{\,n=9}$
alternating	$-0.020\ [-0.110, 0.023]$	$+0.20\ [0.08, 0.32]$	$+0.01\ [-0.05, 0.08]^{\,n=9}$
common	$+0.050\ [-0.011, 0.142]$	$+0.16\ [-0.03, 0.34]$	$+0.07\ [0.02, 0.13]^{\,n=9}$
first	$+0.040\ [-0.034, 0.096]$	$+0.19\ [-0.04, 0.46]$	$-0.03\ [-0.11, 0.05]^{\,n=9}$
last	$-0.072\ [-0.138, -0.007]$	$-0.03\ [-0.22, 0.19]$	$+0.07\ [-0.01, 0.12]^{\,n=9}$
anisotropic (family)	$-0.096\ [-0.156, -0.016]$	$+0.04\ [-0.07, 0.24]$	no estimate (7/10)
$z = 0.70$, dir. 0	$-0.021\ [-0.107, 0.062]$	$+0.07\ [-0.07, 0.26]$	$-0.09\ [-0.18, 0.00]^{\,n=9}$
$z = 0.70$, dir. 1	$-0.042\ [-0.128, 0.028]$	$+0.12\ [-0.06, 0.28]$	$-0.14\ [-0.22, -0.09]^{\,n=8}$
$z = 1.80$, dir. 0	$-0.140\ [-0.231, -0.022]$	$-0.02\ [-0.18, 0.25]$	$-0.79\ [-0.93, -0.66]^{\,n=8}$
$z = 1.80$, dir. 1	$-0.121\ [-0.245, -0.016]$	$+0.06\ [-0.14, 0.31]$	$-0.80\ [-0.95, -0.70]^{\,n=9}$

Table 17: Per-condition contrasts under Adam from the same checkpoints as Table 16, same estimator, same conventions; generated from the summary read in the pre-registered order. Read at the pre-registered horizon on all 10 checkpoints. Per-cell stability: of the fifteen verdicts, five of fifteen hold across adaptation horizons of 2,000, 3,000 and 4,000 steps, and four (level at $z = 1.80$ dir. 1, level and width at uniform common, ridge at uniform last) flip when a single training checkpoint is left out; no effect reverses sign in any of them. The ridge at $z = 1.80$ dir. 0 sits exactly at the 80% complete-case floor, so any deletion sends it to "inconclusive" by rule rather than by estimate, and the leave-one-out exercise is uninformative about it in both directions.

condition	ΔP^{loo}	ΔW_{90} (dex)	Δridge (dex)
layer-uniform (family)	$-0.039\ [-0.152, 0.097]$	$+0.30\ [0.07, 0.52]$	no estimate (7/10)
alternating	$-0.082\ [-0.246, 0.046]$	$+0.23\ [-0.07, 0.59]$	$-0.58\ [-0.75, -0.30]^{\,n=8}$
common	$+0.083\ [-0.024, 0.258]$	$+0.24\ [-0.09, 0.52]$	$-0.09\ [-0.20, 0.01]^{\,n=8}$
first	$-0.173\ [-0.248, -0.096]$	$+0.25\ [-0.00, 0.56]$	$-0.28\ [-0.44, -0.19]^{\,n=9}$
last	$+0.053\ [-0.099, 0.177]$	$+0.38\ [0.16, 0.62]$	$-0.31\ [-0.45, -0.21]^{\,n=9}$
anisotropic (family)	$-0.603\ [-0.690, -0.477]$	$-0.10\ [-0.36, 0.15]$	no estimate (6/10)
$z = 0.70$, dir. 0	$-0.137\ [-0.239, 0.009]$	$+0.14\ [-0.16, 0.41]$	$-0.29\ [-0.42, -0.16]^{\,n=8}$
$z = 0.70$, dir. 1	$-0.207\ [-0.410, -0.041]$	$-0.07\ [-0.36, 0.19]$	$-0.24\ [-0.43, -0.04]^{\,n=8}$
$z = 1.80$, dir. 0	$-0.975\ [-1.162, -0.761]$	$-0.26\ [-0.52, 0.06]$	$-2.63\ [-2.89, -2.05]^{\,n=8}$
$z = 1.80$, dir. 1	$-0.957\ [-1.197, -0.754]$	$-0.22\ [-0.59, 0.07]$	$-2.51\ [-2.85, -2.06]^{\,n=8}$

Table 18: Per-condition contrasts on the convolutional network under SGD (record E11; the same estimators as Table 16, a 20-rate grid, channel-uniform and per-channel rescalings at the same magnitudes): interquartile mean over 10 seeds with a 95% cluster-bootstrap interval; superscript n marks fewer usable seeds; a family row below the 80% complete-case gate carries no estimate. Moved from the body: the family band widens by $+0.30\ [0.07, 0.52]$ (*last* $+0.38\ [0.16, 0.62]$) and the ridge shifts by -0.28 to -0.58 dex on three patterns. The protocol's thresholds (0.10 return units, 0.15 dex, half a grid step $= 0.151$ dex) and its three-way classification are in `RESULTS_E11.md`.

optimizers, and the 80% completeness rule is applied to that intersection. On the level the layer-uniform family gives $+0.234$ $[+0.168, +0.310]$ with all four patterns excluding zero in the same direction, while the anisotropic contrasts change sign with magnitude; on the width all ten intervals cover zero, which is "difference not detected" and not "preserved", since no equivalence margin was registered; on the ridge the four layer-uniform patterns all exclude zero with split signs and the family lands at a value no member supports, and three anisotropic cells are withheld under the completeness rule, including the numerically largest. No multiplicity correction was pre-registered, so isolated point-wise detections are pre-specified point-wise contrasts, not independent discoveries.

The recalibration check in full. A continuous shift-only model $F_{h\theta}(x) = F_\theta(x + b)$ was fitted to each layer-uniform pair—one free parameter, no vertical rescaling, no warping, on a support held fixed across candidate shifts—and its residual compared to a noise floor obtained by splitting the 6 adaptation seeds in half and measuring how far two independent estimates of the *same* curve fall apart.

continuous best shift, median	-0.383 dex (-1.27 grid steps)
residual after best shift, median	0.041 of curve range
same-curve noise floor, on a comparable basis	0.051
residual / noise floor	$0.81\times$

The fixed support retains 7 of the 15 rates and largely excludes the peaks: it contains the identity's observed peak for 2 of 10 checkpoints and none of the 40 shifted comparison curves', so the small residual describes an alignment over the rising flank, not the region the endpoints are read from. The reference scale is not a calibrated null (it compares identity against identity and does not reproduce the shift fitting), and refitting the shift on three adaptation seeds and evaluating on the other three moves the residual to 0.074 against a matched reference of 0.056. Post-hoc and exploratory; the 40 pairs are repeated comparisons within 10 checkpoints, and the ratio is a descriptive summary with no calibrated rejection or equivalence threshold. Figure 6 shows where the peaks sit relative to the fitted window and how the in-sample residual relates to the fitted shift.

Per-arm rate selection on independent seeds (E19). A post-hoc analysis of the E16 record, not a new collection: specified in `PROTOCOL_E19_SHIFT_TEST.md` after E16 had been read (its within-new-seeds 3:3 value was known; no peak position or coverage count was), with no new runs. Four facts fix what it computes. *Selection data*: within the new seeds 6–11, each arm's rate is the argmax of the mean curve over three of them; *evaluation data*: the other three, never used for that selection (all 20 splits averaged per checkpoint; the original seeds 0–5 enter only through the grid, whose two added rates came from E13). *Coverage*: the gate, read first, asks whether the new-seed $K = 6$ argmax is interior to the fixed 17-rate grid and a strict peak; it holds for both arms on all 10 checkpoints (positions 12–15 of 17). It is a check that the peaks are sampled, not that a shifted comparison region is covered at every rate. *Hypothesis and error*: a pure shift $F_{h\theta}(x) = F_\theta(x + b)$ preserves the maximum, so under it the two arms' selected scores differ only by their search losses on the grid and by evaluation noise. Read: the split estimator is $+0.379$ $[0.324, 0.429]$ (interquartile mean $+0.398$ $[0.316, 0.443]$); the descriptive in-sample maxima give $+0.378$ $[0.318, 0.429]$. For scale, on the new-seed mean curves the local spacing at the peak is 0.15–0.30 dex and the drop to the better neighbouring rate averages 0.07 (identity, maximum 0.15) and 0.11 (*common*, maximum 0.24); to the worse neighbour up to 0.63 (`runs/e19_grid_loss.json`, computed after the read). These are descriptive: a sampled peak's neighbour drops do not bound the distance to a continuous optimum lying between grid points, and no smoothness assumption is made here. So the sentence licensed is: *in the common pattern, a difference of* $+0.379$ *remains after each arm's learning rate is selected on its own from a fixed finite grid and evaluated on seeds the selection did not use*; a continuous scalar shift is not excluded, and nothing is said about the other patterns, Adam, or deformations that are not shifts. The pre-fixed rule's stronger sentence ("a single scalar recalibration does not account for it") was withdrawn on review for this reason, before any manuscript number changed.

The rescaling as a per-block learning rate. Write the gauge as $\theta' = A\theta$ with A diagonal: $A = c_\ell / c_{\ell-1}$ on the entries of W_ℓ and c_ℓ on b_ℓ ($c_0 = c_L = 1$). Since $\nabla_{\theta'}\mathcal{L} = A^{-1}\nabla_\theta\mathcal{L}$ for the same loss, one step of plain SGD in the rescaled coordinates, mapped back, is $\theta_{t+1} = \theta_t - \eta A^{-2}\nabla_\theta\mathcal{L}$:

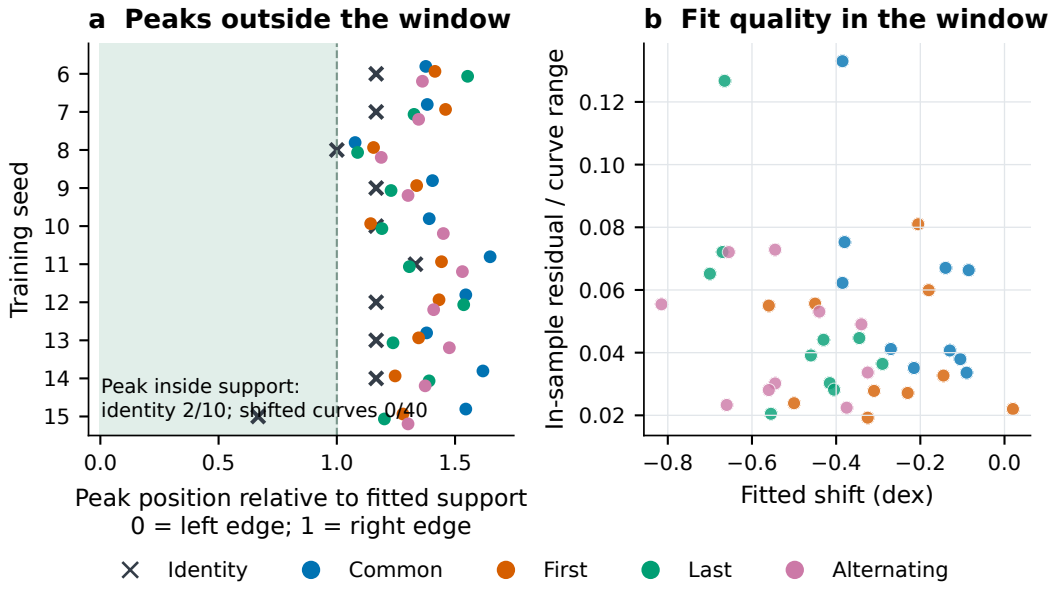

Figure 6: Why the shift-only fit does not settle the recalibration alternative. A shift-only fit on a fixed seven-rate support (401 candidate shifts in ± 1 dex). **(a)** Grid-maximum positions relative to the fitted window (edges 0 and 1): the identity's peak lies inside it for 2 of 10 checkpoints and none of the 40 shifted curves' peaks does, so the fit never sees the region the endpoints are read from. **(b)** Fitted shifts against the in-sample residual; a small residual here does not establish alignment near the peaks. Neither panel is a calibrated test; the 40 pairs are repeated measurements within 10 checkpoints. The test that does address the alternative, on independent seeds with the peaks covered, is the paragraph that follows.

feature set	held-out MAE, k-NN	held-out MAE, ridge
ID (intervention identity, one-hot)	$0.129\ [0.112, 0.146]$	$0.130\ [0.114, 0.144]$
ID + D_{full} (15 audited values)	$0.140\ [0.128, 0.152]$	$0.157\ [0.133, 0.187]$
ID + D_{full} + $\bar{\mathbf{q}}$ (per-layer scale)	$0.139\ [0.129, 0.150]$	$0.148\ [0.130, 0.170]$
D_{full} alone	$0.167\ [0.147, 0.187]$	$0.200\ [0.159, 0.258]$
paired: with $\bar{\mathbf{q}}$ − without	$-0.000\ [-0.003, +0.002]$	$-0.009\ [-0.019, -0.000]$
paired: with D_{full} − ID alone	$+0.011\ [-0.004, +0.026]$	$+0.027\ [+0.001, +0.068]$

Table 19: Predictor baselines with the intervention's identity (E28, retrospective): held-out MAE (return units) of P^{loo} on a held-out checkpoint's nine constructed representatives, nested leave-one-checkpoint-out, E14's k-NN ($k = 5$) and ridge ($\lambda = 1$), 95% cluster-bootstrap intervals over the ten checkpoints. ID is the intervention's identity (a one-hot of the nine representatives, not the identity network arm); D_{full} the 15 audited values of record E17 (E14's D is two of them); $\bar{\mathbf{q}}$ the per-layer mean scale. Decision regret is $0.032\ [0.007, 0.065]$ for every set containing ID and for the best fixed rule chosen in the fold (always *common*); for D_{full} alone $0.222\ [0.167, 0.283]$ (k-NN) and $0.119\ [0.048, 0.200]$ (ridge). The ridge paired difference's upper endpoint is 4×10^{-5} below zero (printed -0.000). An incremental improvement within a feature set is not a lower error than the identity alone, and no set chooses better than the fixed rule.

SGD with the per-block rate $\eta(c_{\ell-1}/c_\ell)^2$ on W_ℓ and ηc_ℓ^{-2} on b_ℓ. For *common* ($c_\ell = e^{0.7}$ on the three hidden layers): $\eta e^{-1.4} \approx 0.25\,\eta$ on W_1 and on b_1, b_2, b_3, η on W_2, W_3, $\eta e^{1.4} \approx 4.06\,\eta$ on W_4. For *first* ($c_1 = e^{0.7\sqrt{3}}$): $\eta e^{-2.42} \approx 0.089\,\eta$ on W_1, b_1, $\eta e^{2.42} \approx 11.3\,\eta$ on W_2, η elsewhere; *last* mirrors it on W_3, b_3 and W_4; *alternating* alternates the factors. With momentum the buffer transforms the same way and the statement holds for the momentum-buffered update. Adam divides each coordinate's update by a running scale of its own gradient, which removes the gradient's factor A^{-1} but not the coordinate map: the update mapped back is $-\eta A^{-1}\, m/(\sqrt{v} + \epsilon')$, a per-block rate ηA^{-1} in place of ηA^{-2} (factors 0.5 and 2 for *common* where SGD has 0.25 and 4), half the effect in log scale, up to ϵ and the running-average transient—one first-order reason the layer-uniform

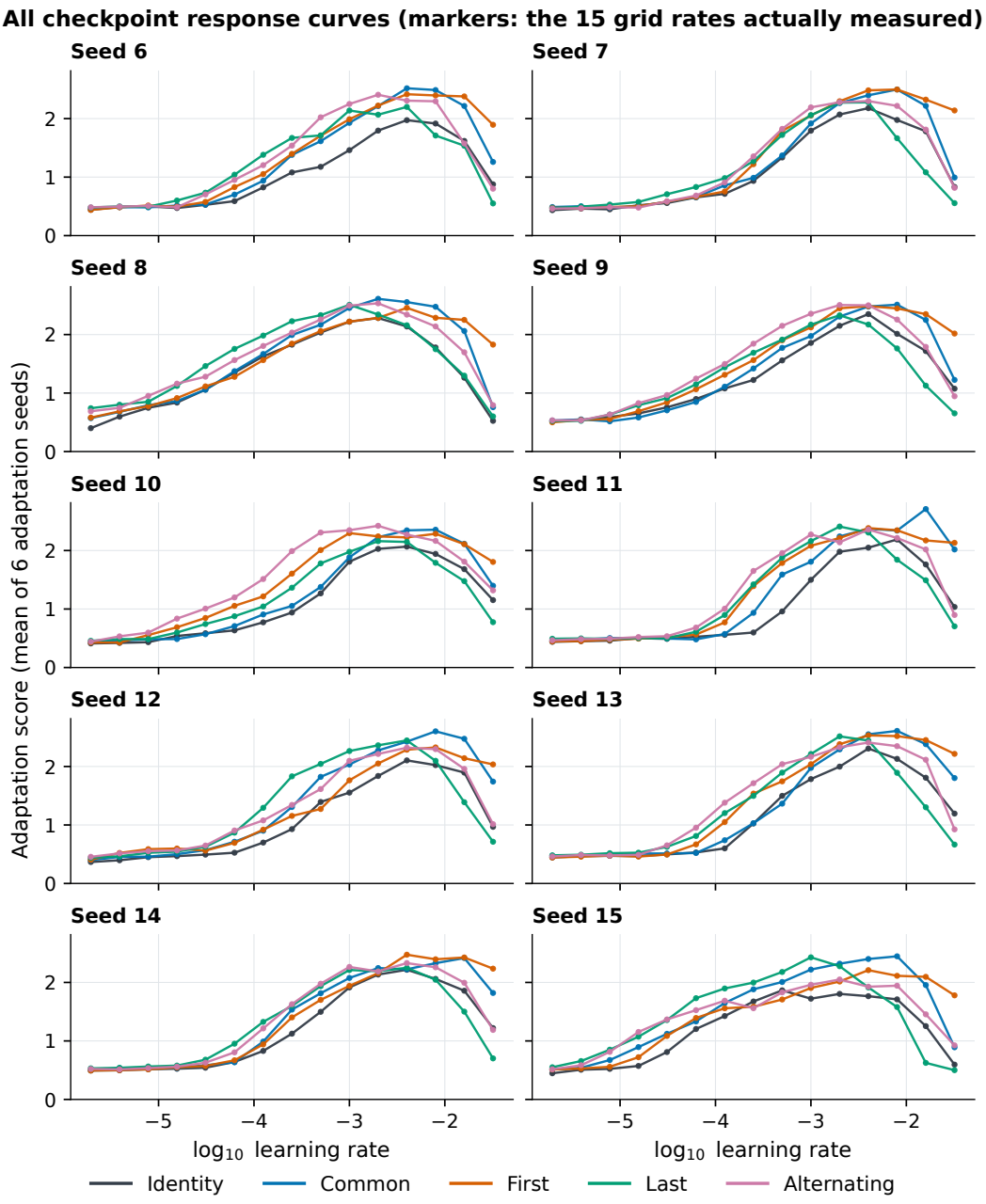

Figure 7: All ten checkpoints' response curves for the identity and the four layer-uniform conditions on common axes; every curve is the mean of the 6 stored adaptation-seed curves, and the markers are the 15 grid rates actually measured—the lines between them are interpolation, and whether this grid separates a shift from a change of shape is unresolved (Section 5). Checkpoints are in seed order, none excluded or selected for effect size. Descriptive; the paired, checkpoint-level intervals are in Table 16.

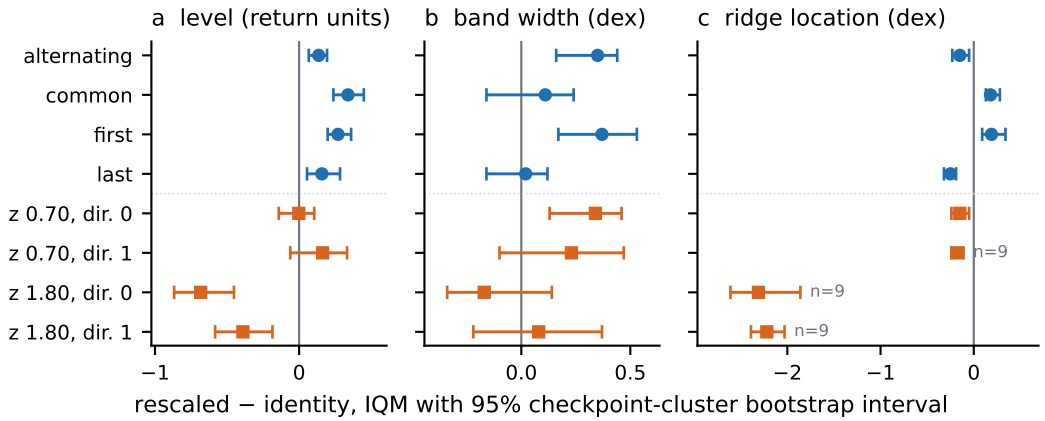

Figure 8: Per-condition contrasts under SGD against the same checkpoint's identity representative (Table 16 drawn), each quantity on its own axis and in its own unit: circles the four layer-uniform patterns, squares the four anisotropic points. **(a)** The level contrast in return units; **(b)** the band-width contrast in dex; **(c)** the ridge-location contrast in dex, whose axis spans about three dex because the two $z = 1.80$ shifts are large. Each panel's limits come from the intervals it draws, so *every* interval is inside its axis: nothing is clipped, and no value is replaced by an off-scale marker. Interquartile means with 95% checkpoint-cluster bootstrap intervals, ten checkpoints per contrast except the cells marked $n = 9$; width and ridge are descriptive, and both are defined conditional on *paired* admissibility, so a ridge contrast is an expectation over the checkpoints on which both rows have a single connected edge-free band and the admissible set can differ between interventions (Appendix D). The two recorded diagnostics are bit-identical to the identity on every layer-uniform row and move on the anisotropic rows (Table 1).

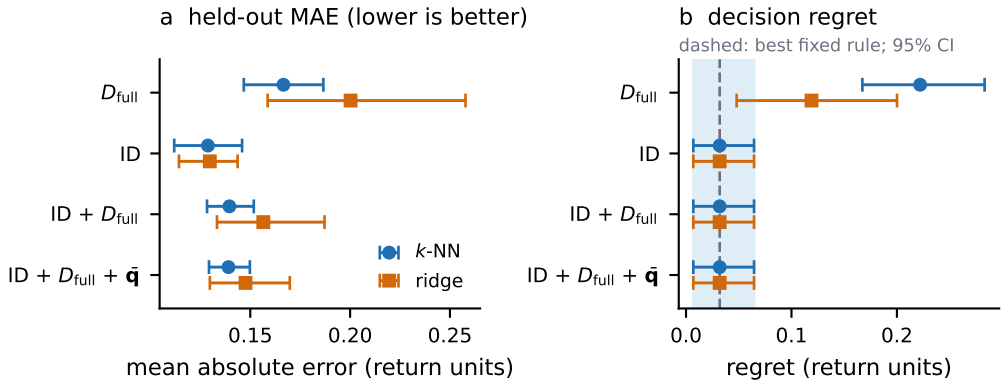

Figure 9: Table 19 drawn from the same record (`scripts/figure_e28_decision.py` reads `runs/e28_read.json`; no value is transcribed). **(a)** held-out MAE and **(b)** decision regret on the 90 constructed representatives, nested leave-one-checkpoint-out, 95% cluster bootstraps over 10 checkpoints, return units. The two panels answer different questions and their sizes are not compared. "ID" is a one-hot of the nine constructed representatives — the transformation's identity, *not* the identity-network arm; D_{full} the 15 audited values of record E17; $\bar{\mathbf{q}}$ the per-layer mean scale. Every set containing ID carries the same regret as the best fixed rule chosen in the fold (always *common*), which is one reported value and not a claim that the fitted predictions coincide. A higher MAE with more features is not itself evidence that the added features carry no information (k-NN distances, ridge shrinkage). Constructed representatives only: nothing here bears on training-produced states.

effect may be smaller under Adam (Section 6.1); a reading, not a test. The gauge-matched oracle below is exactly SGD at these per-block rates on the identity representative, applied in the rescaled coordinates.

The equivariant reference, in full. The oracle is `GaugeMatchedSGD` (`gauge_plasticity/optim/gauge_matched.py`): plain SGD on the rescaled representative with an elementwise rate $\eta\, s_{ij}^2$, $s_{ij} = c_i^\ell / c_j^{\ell-1}$ for weights and $(c_i^\ell)^2$ for biases, which makes the update commute with (1) exactly (`tests/test_gauge_matched.py`, float64 to 10^{-16}); it carries *no momentum*, whereas the scalar arm is SGD with momentum 0.9, a difference the pre-registered test did not net out. The oracle is run on the gauged points of Table 16 over the same grid and adaptation seeds; because it reproduces one trajectory for every gauge up to float32 divergence, the reference per checkpoint is the *median over oracle points* of the maximum of the oracle curve, and the spread across those points is a fidelity check: worst spread 0.162 return units, 7.9% of the identity's post-warmup range, below the 10% margin (admissible). $D_{\text{equiv}} = P^{\text{loo}} - P_{\text{equiv}}$ per row, divided by the same checkpoint's identity range $P^{\text{loo}} - c$; the pre-registered rule is equivalence if the 90% cluster-bootstrap interval of the family's per-seed mean lies inside $\pm 10\%$ (two one-sided tests), meaningful difference if the 95% interval lies entirely outside, inconclusive otherwise. As read: identity 0.101 $[0.059, 0.138]$ inconclusive (the reference's own offset, the momentum and precision differences included); layer-uniform 0.218 $[0.184, 0.252]$ meaningfully different; anisotropic -0.023 $[-0.056, 0.021]$ equivalent, a cancellation of $z = 0.70$ ($+0.088$, $+0.162$) against $z = 1.80$ (-0.232, -0.111), so point-wise equivalence is not claimed. D_{equiv} scores the scalar arm at a held-out rate while reading its reference off the maximum of the oracle curve, so the two arms are selected differently by construction. In a post-hoc sensitivity analysis, selecting the oracle rate from its own leave-one-checkpoint-out curves and applying the original aggregation and three-state equivalence rule left all three family-level verdicts unchanged. This assesses sensitivity to the specified reference-selection procedure rather than estimating or removing bias; a statistic shifted in one direction is not automatically conservative against a symmetric equivalence margin—that caution stands—and on these data the verdicts did not turn on it. The anisotropic family's "equivalent" verdict is itself a cancellation of opposite-signed components at the two magnitudes, and point-wise equivalence was not pre-registered and is not claimed.

The momentum-matched reference (E15): inadmissible at the pre-registered fidelity. Because the E9 oracle lacks the scalar arm's momentum, a version with momentum 0.9 on the elementwise-scaled update was pre-specified (`PROTOCOL_E15_MOMENTUM_ORACLE.md`; it commutes with (1) exactly, `tests/test_gauge_matched.py`) and run on the same three oracle points, checkpoints, grid, seeds and horizon (2,700 adaptation runs, stamp `68ffee2a9e32dffa`, function change at most 1.2×10^{-15}). Its scope was fixed before any row existed: it is a *coordinate-wise* update control matched to the transformation, and it does not address whether one scalar learning-rate recalibration explains the difference. The control did not meet its pre-registered fidelity gate (worst spread of the three oracle points' curve maxima 0.224 return units, 11.0% of the median identity range against the 10% margin; E9's no-momentum oracle 7.9%) and is not used as evidence for the response comparison; the margin was not moved, and the family values it would have produced stay in `runs/e15_read.json`. One descriptive fact, allowed by the protocol: the identity family's own offset is -0.030 $[-0.043, -0.008]$ with momentum matched against 0.101 without. This is a failure of a control, not a test of a hypothesis, and its cause was located after the read (`scripts/e20_fidelity_trace.py`, record E20): on one checkpoint, with the same fixed minibatch sequence, the same targets and no environment, the identity under plain momentum SGD and each rescaled oracle point mapped back through its gauge agree to 6×10^{-8} relative after the first update in float32—one unit in the last place, so the update commutes as implemented—and to 10^{-16} in float64; in float32 the relative parameter difference stays below 10^{-6} for about 100 updates, then grows to 5×10^{-4} by update 200 and 3×10^{-2} by update 2,000, and the first greedy-action disagreement on the 512 probe states occurs at update 176 (common and $z = 0.70$) and 100 ($z = 1.80$); in float64 no action disagrees and the difference stays at 10^{-15} through 2,000 updates. On the fixed-data path the trace therefore supports numerical amplification of float32 rounding, not a mismatch of the update rule; that the online failure has the same cause is consistent with the first action flip but is not established by this trace, and the float64 acquisition of the control (E21, next paragraph) is the test of it.

The control in float64 (E21). Pre-specified in `PROTOCOL_E21_FLOAT64_ORACLE.md` with its reading table fixed before any row existed: the identity and *common* scalar arms under plain momentum SGD and the three oracle points under the momentum-matched gauge-matched opti-

mizer, all in double precision, on E9's checkpoints, grid, seeds and horizon (4,500 runs, stamp `b3c73c9857db6683`); the float32 default path of the adaptation code is pinned byte-identical (`tests/test_dtype_path.py`). Read in the fixed order. *Fidelity*: the worst spread of the three oracle points' curve maxima is 0.0125 return units, 0.6% of the median identity range, against the unchanged 10% margin—admissible (11.0% in float32, E15). *Precision*: the float64 identity's P^{loo} minus the float32 E9 identity's is $+0.004$ $[-0.046, 0.053]$, the *common* arm's -0.001 $[-0.038, 0.039]$; the observed precision contrasts were small (no equivalence test was registered for them). *The control*: the identity family's own offset from the reference is -0.005 (90% interval $[-0.021, 0.000]$), **equivalent** within $\pm 10\%$—the reference reproduces the identity's selected level; the *common* arm's D_{equiv} is $+0.162$ (90% $[0.112, 0.204]$, 95% $[0.103, 0.212]$), **meaningfully different**. *Direct contrast*: *common* − identity in float64 is $+0.346$ $[0.248, 0.435]$ (interquartile mean $+0.365$ $[0.237, 0.446]$), beside E9's float32 $+0.351$ $[0.255, 0.449]$ for the same pattern. Under the rules fixed in advance (row 1 of the decision table and row 2 of the reading table): the layer-uniform response difference is not the identity's own trajectory re-expressed in the rescaled coordinates—plain SGD on the rescaled representative and the exact per-block-rate control differ, in double precision, with momentum matched—and the numerical-precision concern about the main comparison is reduced. The E9 no-momentum oracle verdicts above are superseded by this reading for the *common* pattern; the anisotropic scalar arms were not re-collected, so their E9 verdicts stand as disclosed. The scalar recalibration question (E19) is not addressed here.

Horizon cells. The fifteen cells of Table 17's caption are the three summaries (level, width, ridge) at the five conditions named in the horizon protocol—*last*, *alternating*, *common*, and $z = 1.80$ in both directions—each read at 2,000, 3,000 and 4,000 steps under both optimizers; a cell is stable if its detected/not-detected verdict is the same at all three.

Diagnostic specificity, in full. The pre-registered test pools the Spearman rank correlation between a diagnostic (the dormant ratio, and the pooled spread σ_q of the scale descriptor read directly) and each response endpoint over the usable seed-by-intervention rows (7–9 rows per seed, 10 seeds; the point estimate is not reweighted for unequal seeds), with a cluster-bootstrap interval over seeds; the control removes intervention magnitude as a *fixed effect*, centring ranks within each magnitude level, and the rule fixed before reading is that an association not surviving the control is magnitude-confounded. Under SGD, in the pre-declared endpoint order (D_{equiv}, r_{cal}, W_{90}, P^{loo}, ridge as descriptive), the dormant ratio's pooled correlations are -0.42, $+0.58$, -0.20, -0.45, -0.65 and its controlled ones $+0.14$ $[-0.22, 0.39]$, $+0.12$ $[-0.17, 0.41]$, $+0.01$ $[-0.14, 0.17]$, $+0.11$ $[-0.17, 0.24]$, -0.23 $[-0.48, 0.08]$, every interval covering zero; σ_q survives the control on P^{loo} (-0.19 $[-0.34, -0.05]$) and the ridge (-0.37 $[-0.53, -0.21]$), which is consistent with—not confirmatory of—the dormant ratio being a lossy reading of scale geometry. Under Adam no diagnostic is individually detected on any endpoint, and the pre-registered contrast asking whether a rank-based diagnostic retains more than the dormant ratio covers zero on every endpoint. The matched-diagnostic counterexample (EM) draws, per checkpoint, two lognormal gauge points ($\sigma = 0.7$) with the dormant ratio exactly equal to the identity's at $\tau = 0.025$ and intervention magnitudes $\|\log c\|_{\text{RMS}}$ matched within 2%, taking the pair that maximises a pre-specified imbalance statistic, all before any adaptation was run; its primary contrast $P^{\text{loo}}(A) - P^{\text{loo}}(B)$ over the ten checkpoints returned -0.0687 $[-0.1972, +0.0654]$ against a practical threshold of 0.10 (established: $|\text{point}| \geq 0.10$ and the interval excludes zero; null: the interval within ± 0.10): inconclusive, and neither the effect nor the null wording is available.

F ECOLOGICAL EVIDENCE AND THE TRANSPOSED INTERVENTION

Natural drift, in full. Over every snapshot of 16 training runs, the median displacement of the per-layer gauge coordinate—in the basis the theory fixes, not the directions our effects were largest on—is 0.260 in the layer-uniform mode against 0.192 in the anisotropic mode: 0.11–0.81 of what the layer-uniform patterns and the $z = 0.70$ points impose in that mode, and 0.06–0.17 of the anisotropic points in theirs. So the *projection* of ordinary training onto the blind coordinates moves by the order we impose—not that training stays on an orbit, since it also changes the function; whether the natural motion is *aligned* with our directions is a separate question, and at this dimension a displacement of the same size can be nearly orthogonal. Post-hoc, designed after E9 was read; a magnitude observation, not evidence of how often the blind directions are traversed in a way that matters. *Subspace projection and direction (record E24, post-hoc, descriptive).* Two things are separated: how much of the natural drift lies in the blind *subspace*, and whether it points along the *direction* the effect was measured on. The projection is made explicitly: for each of the ten E9 training runs, the change of the per-unit coordinate $q_i = \log(\|w_i^{\text{out}}\|/\|w_i^{\text{in}}\|)$ between two snapshots is split into its per-layer-constant part—the image of the layer-uniform gauge, which the ratio-normalised diagnostics cannot see—and the within-layer residual. Over the whole run (untrained to final), 37% of the squared change lies in the blind subspace (range 28–44%), with per-layer mean shifts of $(-0.33, +0.22, +0.15)$, about a fifth to a half of the *common* intervention's shift of $(-0.7, 0, -0.7)$; over the second half of training only 7% does (per-layer shifts at most 0.11). Direction is a different matter: for each training run the cosine between that run's three-entry per-layer mean shift and the *common* intervention's shift $(-0.7, 0, -0.7)$ is computed, and the mean over the ten runs is $+0.31$ over the whole run (range over runs $+0.03$ to $+0.49$) and $+0.21$ over the second half; these are descriptive (no interval is formed; the unit is the training run), and the last layer's mean natural shift has the opposite sign to the intervention's. So ordinary training does move within the blind subspace, mostly early and at a fraction of the magnitude we impose, only weakly along the direction measured; what a shift of that size does to the response is measured only as an imposed intervention (E25, Section 6.4), not along a natural drift, and the states training produces are not gauge images of one another, so this is alignment of one coordinate's drift, not evidence that the counterexample occurs in training.

The transposed intervention. In a redistribute-and-reset intervention transposed from ReBorn, the median ratio of dormant-ratio reductions under a function-preserving, exactly-dead-only redistribution arm and under the full operation was 68% $[67\%, 71\%]$ (10 adaptation seeds, one checkpoint). That arm includes a unit split, not an invertible gauge motion, so it is not evidence about the blind subgroup, and a ratio of two arms does not isolate a gauge effect or an additive contribution of redistribution within the full operation.

The intervention whose share is reported above is our transposition of ReBorn (Qin et al., 2024) to a single-agent MinAtar DQN, and is named as such. Operation by operation:

1. **Classification.** Units are scored by the dormant ratio of Table 6 on the checkpoint's probe batch. Dormant: $s_i \le \tau$ with the published $\tau = 0.025$. Over-active: $s_i \ge 2.0$; this cutoff is *ours*, since the source's definition is stated for its own setting and we did not transfer its value.
2. **Redistribution.** For each over-active unit x, draw $M \sim \mathcal{U}\{2, \dots, 5\}$ dormant recipients from the layer's pool (fewer if the pool is short; none if fewer than two remain). Over the $M+1$ units consisting of x and its recipients, draw $\beta_j \sim \mathcal{U}(0.5, 1.5)$ and $\alpha = \mathrm{softmax}(z)$ with z standard normal, so $\sum_j \alpha_j = 1$, and set for each $w_j^{\text{in}} \leftarrow \beta_j w_x^{\text{in}}$, $b_j \leftarrow \beta_j b_x$, $w_j^{\text{out}} \leftarrow (\alpha_j/\beta_j)\, w_x^{\text{out}}$. By ReLU homogeneity the $M+1$ units' combined contribution equals x's previous contribution, so this step is the rescaling (1) composed with a partition-of-unity split of one unit into several. It preserves the function exactly only when the recipients contributed nothing; at $\tau = 0.025$ they are quiet rather than dead, and their contribution is destroyed. The arm whose share is reported above therefore applies this step to exactly-dead recipients only ($\tau = 0$), where the measured function change is 6.7×10^{-16}; with the published τ the same step moves the function by up to 10^{-1}.
3. **Reset.** The dormant units not consumed by redistribution are re-initialised (Xavier), which changes the function; the full transposed method moves it by 6.2×10^{-1}.

The ratio quoted, 68% $[67\%, 71\%]$, is the median over 10 adaptation seeds *from one trained checkpoint* of the per-seed ratio of the dormant-ratio reduction under the $\tau = 0$ redistribution arm to that under the full transposed method, with a percentile interval from a bootstrap over those seeds (`scripts/e6_reborn_share.py`, which recomputes it from the recorded run). Three limits on what it means. It is a ratio between two separately applied arms, not a decomposition of the full operation into an additive redistribution contribution and a reset contribution, since the arms do not share a prefix of the same operation. The function-preserving arm composes a rescaling with a unit split that changes the tangent space and is not an invertible gauge motion, so it is not an element of H_{unif} and says nothing about the blind subgroup; whatever it shows is attributed to the composite. And it is one checkpoint, so it is an illustration and not evidence of generalisation across checkpoints.