SHARED COMPUTATION, DIFFERENT READOUTS: DEPTH COMPATIBILITY IN LOOPED TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-depth training can improve a looped transformer's accuracy under a shared output head, yet this advantage can reverse when both models receive a small adaptive interface. We study depth compatibility through the excess prediction risk of sharing one readout across depths instead of fitting equally expressive heads separately. In controlled in-context linear regression, we compare fixed-depth and multi-depth training with matched initializations, examples, optimizer updates, and cumulative recurrent computation. Multi-depth training lowers shared-head error but raises the error of a separately fitted head at the reference depth. After we freeze the cores, two shared adaptation directions remove over 90% of the all-depth sharing cost in each of six models. Four directions reverse the training comparison in all three initialization pairs on two separate 4,096-episode evaluations. Depth-specific gain and offset recover substantially less accuracy, even when we jointly optimize the common prediction direction. We characterize capacity-dependent risk under different, possibly singular feature geometries across depths. The task objective separates separately decodable risk from compatibility cost. An exact information-preserving recurrence shows that a large sharing cost can require only two adaptation directions. The comparison reverses with far less flexibility than fully independent heads provide. These results distinguish the prediction accuracy available in recurrent states from their compatibility with a shared output interface.

1 INTRODUCTION

Looped transformers reuse a learned computational block at different inference budgets. The same architectural principle supports iterative function learning and longer latent computation at test time (Dehghani et al., 2018; Yang et al., 2024; Geiping et al., 2025). This flexibility is useful only if the model can produce an answer when execution stops after different numbers of recurrent steps. Multi-depth training addresses this requirement by supervising several stopping depths through one output head. Its success raises a basic question: does it improve the prediction accuracy available in the recurrent states, or make those states easier to read through a common interface?

Shared computation does not make these objectives identical. A recurrent transition can preserve task information while changing which state coordinates a predictor must combine. A head that reads an answer accurately at one depth can then fail at another. Training several depths through the same head asks the core to support accurate predictions and to express them in a mutually compatible form. An improvement under that constraint can reflect either change. This distinction matters when we use output accuracy to compare depth-training strategies or decide whether additional recurrent computation helps.

Existing probes provide one part of the answer. Affine readouts can recover intermediate predictions that a model's native head misses (Alain & Bengio, 2016; Belrose et al., 2023). However, fully independent heads remove the cross-depth constraint almost entirely. They do not reveal how much readout freedom is needed to change a comparison between learned cores. We instead ask a capacity-controlled question: *does a training advantage survive when both frozen cores receive the same small increase in output-interface flexibility?* This intervention preserves every recurrent state and changes only how the answer is extracted.

We formalize the question through *depth compatibility*, the excess prediction risk imposed by sharing one affine readout across depths. At a fixed core, the population loss separates into the risk

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.

achievable with independent heads, the cost of sharing a head, and the current head's gap from the best common head. Once we optimize the head, the first two terms remain. Single-depth supervision has zero sharing cost by definition; supervision across depths adds a compatibility requirement. We characterize the cost of relaxing this requirement through a few shared adaptation directions, allowing feature distributions to differ across depths and to be rank deficient. An information-preserving recurrence shows why the size of a sharing penalty and the dimension needed to remove it are different properties.

We test this distinction in in-context linear regression, a controlled setting with known task targets and learned recurrent computation (Garg et al., 2022; Yang et al., 2024). Fixed-depth and multi-depth training use matched initializations, training examples, optimizer updates, and total recurrent-block applications. We freeze the resulting cores and evaluate sixteen depths under the same readout families. Multi-depth training wins with a completely shared head. Yet two learned directions remove more than 90% of the sharing cost in all six matched models, and four directions reverse the comparison in all three initialization pairs. A separate test confirms the reversal. Gain-and-offset controls recover substantially less accuracy, including when we jointly optimize their common prediction direction.

Our contributions connect the training objective to this intervention:

- We separate independently decodable accuracy from compatibility with a shared answer interface, and identify how supervision across depths adds the latter constraint.
- We characterize interface cost under depth-dependent feature geometry and show, in an exact recurrent example, that a large cost can require only two adaptation directions.
- We demonstrate an early reversal of a matched training comparison under equal readout capacity, supported by calibration controls, random directions, and per-pair results.

2 Related work

Shared computation and adaptive depth. Universal Transformers apply a shared attention-based transition across depth (Dehghani et al., 2018), while ALBERT shares layer parameters to reduce model size (Lan et al., 2019). Deep equilibrium models solve for a fixed point of a weight-tied network (Bai et al., 2019). Adaptive Computation Time and PonderNet learn how much recurrent computation to allocate (Graves, 2016; Banino et al., 2021). Recurrent-depth language models extend latent computation at inference (Geiping et al., 2025). These architectures motivate evaluation at several depths; we study their shared output-interface constraint through interventions on frozen cores.

Reading and comparing representations. Linear probes assess intermediate-state predictability (Alain & Bengio, 2016), and the tuned lens uses affine maps to decode latent language-model predictions (Belrose et al., 2023). Control tasks (Hewitt & Liang, 2019), information-theoretic probing (Pimentel et al., 2020), and minimum-description-length probing (Voita & Titov, 2020) show why interpretation depends on the predictor and its learning resources. We specify both readout capacity and fitting supervision. Representation similarity measures compare feature spaces (Kornblith et al., 2019), while model stitching tests task performance through learned connectors (Bansal et al., 2021). Task-loss matching and direct state matching can yield different similarity judgments (Balogh & Jelasity, 2024). Our intervention measures answer-prediction risk and tests whether equal readout flexibility changes the ordering of training regimes.

Learning algorithms and compatible predictors. Synthetic function-learning tasks (Garg et al., 2022) support comparisons with explicit learning algorithms. Transformers can implement or learn regression updates (Akyürek et al., 2023; Von Oswald et al., 2023), and looped architectures favor iterative computation (Yang et al., 2024). We use this setting to distinguish prediction quality from readout compatibility. Invariant risk minimization seeks a common optimal predictor across environments (Arjovsky et al., 2019); here, environments correspond to recurrence depths. Multi-task feature learning provides a precedent for shared low-dimensional predictors (Argyriou et al., 2006). Our common-metric and two-depth formulas use classical low-rank approximation (Eckart & Young, 1936) and matrix parallel sums (Anderson & Duffin, 1969). These tools explain the risk decomposition; the empirical finding is the small capacity that reverses the matched comparison.

Recent recurrent-state diagnostics. Recent studies link recurrent stability to test-time scaling (Yang et al., 2026), identify scale variables hidden by normalized readouts (Sharma & Vu, 2026), and study information transport across recurrences (Wang & Reid, 2026). Pressure to improve readout accuracy can also alter internal geometry (Oskin, 2026). Popescu et al. (2026) distinguish trajectory formation from exit-depth selection by fitting stopping gates to frozen trajectories. Their readout chooses when to exit; ours predicts an answer at a given depth. We compare matched training regimes under equal-capacity answer interfaces and find a reversal with only a few learned directions. The orthogonal example isolates the corresponding compatibility problem in an information-preserving system.

3 SETUP: THREE PREDICTION RISKS ON ONE FROZEN CORE

To distinguish available prediction accuracy from readout compatibility, we compare three readouts while keeping the recurrent states and targets fixed. Let $H_{t+1} = F_\theta(H_t)$ denote the recurrent state and $Z_t = \psi(H_t) \in \mathbb{R}^m$ the feature available to the readout. In our experiments, ψ applies the trained layer normalization (LayerNorm) to the query-token state. All depths in an episode share the scalar target Y. We measure risk by mean squared error (MSE), the expected squared prediction error. For affine heads, define $X_t = (Z_t^\top, 1)^\top$ and $R_t(v) = \mathbb{E}[(v^\top X_t - Y)^2]$, with $\mathbb{E}\|X_t\|^2 < \infty$ and $\mathbb{E}Y^2 < \infty$.

Fix a finite nonempty set of nonnegative integer depths S and positive weights π_t summing to one. Define

$$L_{\text{sep}} = \sum_{t\in S} \pi_t \min_v R_t(v), \qquad L_{\text{shared}} = \min_v \sum_{t\in S} \pi_t R_t(v), \tag{1}$$

$$\Gamma_\pi = L_{\text{shared}} - L_{\text{sep}}, \qquad E_{\text{native}} = \sum_t \pi_t R_t(v_{\text{native}}) - L_{\text{shared}}.$$

Both Γ_π and E_{native} are nonnegative population quantities. Native prediction risk therefore has the exact decomposition

$$L_{\text{native}} = L_{\text{sep}} + \Gamma_\pi + E_{\text{native}}. \tag{2}$$

The first term measures the best prediction risk at each depth within the chosen head family. The second measures the cost of using one head across depths, and the third measures the native head's gap from that common optimum. Separate prediction risk and sharing cost can vary independently.

3.1 WHAT SUPERVISION ACROSS DEPTHS OPTIMIZES

The same decomposition clarifies what depth supervision asks the core to learn. We use weights q for training and π for evaluation, since their depth distributions can differ. For core parameters θ, let $R_t^\theta(v)$ denote population task risk and define

$$J_q(\theta, v) = \sum_t q_t R_t^\theta(v), \qquad B_q(\theta) = \sum_t q_t \min_a R_t^\theta(a). \tag{3}$$

The training support is finite, and its weights are positive and sum to one. Let $\Gamma_q(\theta)$ be the sharing cost on this support and let $E_q(\theta, v)$ be the head's excess risk above its common optimum.

Proposition 1 (The compatibility requirement of depth supervision)**.** *Under the finite-second-moment assumptions, the population task objective satisfies*

$$J_q(\theta, v) = B_q(\theta) + \Gamma_q(\theta) + E_q(\theta, v), \qquad \min_v J_q(\theta, v) = B_q(\theta) + \Gamma_q(\theta). \tag{4}$$

Let $\mathcal{A}_t(\theta) = \arg\min_v R_t^\theta(v)$ be the set of optimal affine heads. Then

$$\Gamma_q(\theta) = 0 \iff \bigcap_{t\in \operatorname{supp} q} \mathcal{A}_t(\theta) \neq \varnothing. \tag{5}$$

In particular, supervision at one depth has zero sharing cost for every core.

After optimizing the head, the task objective still rewards both accurate depth-specific predictions and compatibility with a common predictor. Training across depths can change either term. The identity concerns unregularized population loss; our experiments measure the outcomes of finite training under evaluation weights π. Fixed training supervises depth 4, balanced training covers depths 1–7, and the capacity analysis weights depths 1–16 equally.

Estimation and head transfer. We estimate these quantities by fitting heads on one split and selecting ridge penalties on a calibration split. Their test-risk differences, $\widehat{\Gamma}_\pi$ and $\widehat{E}_{\text{native}}$, include estimation and regularization effects (Appendix B). To locate disagreement across depths, we also transfer fitted heads and measure

$$C_{s\to t} = \widehat{R}_t(\widehat{v}_s) - \widehat{R}_t(\widehat{v}_t), \tag{6}$$

where $\widehat{v}_s$ is fitted at source depth s. All risks in a comparison use identical test episodes.

4 Readout compatibility, capacity, and recurrence

We now express the sharing cost through differences among depth-specific predictors, then examine how adaptation reduces it.

4.1 Regression geometry and observable disagreement

A coefficient difference incurs error only through its effect on predictions, as the following regression identity shows. Let $\Sigma_t = \mathbb{E}[X_t X_t^\top]$, $b_t = \mathbb{E}[X_t Y]$, and choose any optimal head v_t satisfying $\Sigma_t v_t = b_t$. Then

$$R_t(v) = R_t(v_t) + (v - v_t)^\top \Sigma_t (v - v_t). \tag{7}$$

The sharing cost is therefore the minimum weighted disagreement with independently optimal heads, measured in each depth's feature geometry. This interpretation also covers singular Σ_t, as can arise from normalization: different coefficients can produce identical predictions.

A matrix parallel-sum formula gives the two-depth cost, including for singular feature distributions (Appendix A). Only coefficient differences that affect predictions contribute to the cost.

Head transfer provides another view of this decomposition. Define the population transfer cost $C^*_{s\to t} = R_t(v_s) - R_t(v_t)$. If every depth has the same second moment Σ, then

$$\Gamma_\pi = \frac{1}{2} \sum_{s,t} \pi_s \pi_t C^*_{s\to t}. \tag{8}$$

Under a common metric, sharing cost equals half the average pairwise transfer cost. Different depth metrics allow asymmetric transfer, so our experiments measure shared risk directly and use head transfer to locate disagreement.

4.2 Capacity under depth-dependent feature geometry

We next relax the common-head constraint with shared adaptation directions and measure how quickly their extra freedom reduces risk. Let $r \geq 0$ be an integer and $D = m + 1$ the augmented readout dimension. Define $\mathcal{R}_\pi(r)$ as the infimum of weighted risk for heads $w_t = w + U a_t$, where $U \in \mathbb{R}^{D\times r}$ contains shared adaptation directions. Rank zero shares the complete affine head, including its intercept. For independently optimal heads v_t, Equation (7) gives the exact capacity cost

$$C_\pi(r) := \mathcal{R}_\pi(r) - L_{\text{sep}} = \inf_{w, U, \{a_t\}} \sum_t \pi_t \|w + U a_t - v_t\|^2_{\Sigma_t}. \tag{9}$$

Here $\|q\|^2_{\Sigma_t} = q^\top \Sigma_t q$; depth-dependent and singular second moments are allowed.

What the capacity constraint controls. Every depth uses the same common component w and directions U, with its own coefficients a_t. Each direction changes feature weights together, so r counts patterns of change across heads. It measures neither their size nor the number of scalar parameters. A common invertible coordinate change preserves this family when the heads transform accordingly; depth-specific changes alter the sharing constraint. Our comparisons therefore hold feature extraction fixed and use one common coordinate transform.

Proposition 2 (Eliminating coefficients at a fixed shared subspace)**.** *Write † for the Moore–Penrose pseudoinverse. For fixed U, define*

$$P_t = \Sigma_t - \Sigma_t U (U^\top \Sigma_t U)^\dagger U^\top \Sigma_t, \quad A_U = \sum_t \pi_t P_t, \quad b_U = \sum_t \pi_t P_t v_t. \tag{10}$$

Then $P_t \succeq 0$, $b_U \in \mathrm{range}(A_U)$, and the best residual cost at that subspace is

$$C(U) = \sum_t \pi_t v_t^\top P_t v_t - b_U^\top A_U^\dagger b_U, \qquad C_\pi(r) = \inf_{U \in \mathbb{R}^{D\times r}} C(U). \tag{11}$$

The expression is unchanged when an optimal head is replaced by another representative that differs only within $\ker \Sigma_t$.

The matrix P_t measures the part of a coefficient mismatch that remains after adaptation along U. Once those directions are fixed, the common component solves a quadratic problem with a closed-form optimum. The outer search over directions remains nonconvex; Appendix B proves the result and treats possible nonattainment under singular metrics.

The same decomposition explains the training comparison. Let B_F, B_M denote two cores' separate risks, with $\delta = B_M - B_F > 0$. Then

$$\mathcal{R}_F(r) < \mathcal{R}_M(r) \quad \Longleftrightarrow \quad C_F(r) - C_M(r) < \delta. \tag{12}$$

At $r = \min(D, |S| - 1)$, the interface can represent arbitrary independent heads, so both costs vanish. The empirical question is how much smaller the required capacity is for learned cores. Our curves report test risks achieved by calibrated ridge fits. Appendix B distinguishes these risks from population infima and fixed-penalty fitting objectives.

4.3 The size and dimension of incompatibility

Does a large sharing cost require many adaptation directions? A common feature metric reveals why these quantities can differ. Under a common positive-definite second moment Σ, let $\overline{v} = \sum_t \pi_t v_t$ and form the matrix

$$M_\pi = \left[\sqrt{\pi_t}\,\Sigma^{1/2}(v_t - \overline{v})\right]_{t\in S}. \tag{13}$$

If σ_j are its singular values in decreasing order, classical low-rank approximation (Eckart & Young, 1936) gives

$$\Gamma_\pi = \|M_\pi\|_F^2 = \sum_j \sigma_j^2, \qquad C_\pi(r) = \sum_{j>r} \sigma_j^2. \tag{14}$$

The total squared spectrum determines the initial cost; its concentration determines the capacity needed to remove it. The same result holds on the identifiable range of a common singular metric. Appendix B proves the formula. For the learned cores, we use the depth-dependent formulation in Equation (9) and measure prediction risk directly.

A change in target scale makes this distinction explicit. Multiplying Y by a nonzero constant c multiplies each optimal head by c and each excess risk by c^2, while preserving the dimension needed to represent the heads. A large sharing cost alone therefore gives no lower bound on the required adaptation dimension. The normalized capacity curve instead describes how the cost is distributed across the available interface directions.

The following recurrence preserves all input information while concentrating its readout changes in just two directions.

Proposition 3 (Many depths, two adaptation directions)**.** *Let $K \geq 3$, $m \geq 2$, $Z_0 \sim \mathcal{N}(0, I_m)$, and $Y = e_1^\top Z_0$. Let A rotate the first two coordinates by $2\pi/K$ and act as the identity on the others. For $Z_t = A^t Z_0$ and uniform evaluation on $t = 1, \ldots, K$, affine readouts satisfy*

$$L_{\mathrm{sep}} = 0, \qquad C_\pi(0) = 1, \qquad C_\pi(1) = \tfrac{1}{2}, \qquad C_\pi(2) = 0. \tag{15}$$

Every depth preserves the entire input state, and its optimal slope is $A^t e_1$. These slopes trace a circle in one plane. Their mean is zero and their weighted second moment has two eigenvalues equal to $1/2$, which yields Equation (15). A completely shared head therefore loses all predictive accuracy, yet two directions recover every optimal head for any number of depths. Appendix A also characterizes how the distribution of depths interacts with recurrent frequencies.

5 EXPERIMENTS

The theory separates a core's prediction accuracy from the cost of its output interface. We test this distinction in trained looped transformers: does the advantage of multi-depth training survive a small increase in readout freedom? We match training budgets, freeze the cores, and compare them across readout capacities.

5.1 EXPERIMENTAL SETUP

This comparison needs a task whose target stays fixed as recurrent depth changes. We use procedurally generated in-context linear regression, following Garg et al. (2022) and its looped-transformer formulation (Yang et al., 2024). Each episode contains examples of a new linear function. We draw $\beta \sim \mathcal{N}(0, I_d/d)$ and inputs $x_i \sim \mathcal{N}(0, I_d)$, give the model k pairs $(x_i, \beta^\top x_i)$, and ask it to predict $Y = \beta^\top x_{k+1}$ for a query input. Training and test episodes use independently sampled functions and inputs. The model must infer the function from its context, and the same exact target lets us compare readouts throughout its recurrent trajectory.

The main models use four-dimensional inputs, eight context examples, and a width-32 transformer block that shares its parameters across recurrences. An affine head reads the normalized query state. We compare fixed-depth training, which supervises the output after four recurrences, with balanced multi-depth training, which supervises a terminal depth between one and seven. Each of three initialization pairs starts from identical weights and receives the same 96,000 training episodes. Both regimes use 1,500 optimizer updates and exactly the same cumulative number of recurrent-block applications. The comparison changes which depths receive supervision while holding the model, data, and training-compute budget fixed.

After training, we freeze all six cores and extract their query states at depths 1–16. We then fit the readout family $v_t = v + Ua_t$ from Section 4. A common head has $r = 0$; increasing r gives every depth access to the same r adaptation directions, with its own coefficients. We evaluate $r \in \{0, 1, 2, 4, 8, 15\}$, where $r = 15$ can express independent affine heads at all sixteen depths. Both regimes receive the same readout family and fitting data at each capacity. All evaluated depths participate in readout fitting, so these experiments measure which predictions can be extracted from the frozen trajectories with a specified interface budget.

We fit readouts on 2,048 episodes, select regularization on 512 calibration episodes, and evaluate the frozen heads on 4,096 fresh episodes. Mean squared error averages predictions across the evaluated depths, episodes, and three models in each regime. The capacity comparison, its calibration controls, and the two-depth analysis use separate test sets. Within each comparison, all models answer the same episodes. Reported intervals resample these episodes conditional on the six trained cores, and per-pair results track consistency across initializations. Appendix E gives the architecture, training schedule, and data splits; Appendices H and I give the readout objectives and optimization checks.

5.2 A SMALL ADAPTIVE READOUT REVERSES THE TRAINING COMPARISON

With the cores fixed, we start from a common head and ask how their ordering changes as adaptation becomes available. Multi-depth training leads at zero capacity: its mean error across sixteen depths is 0.0940, compared with 0.1461 for fixed-depth training. Figure 1 shows what happens when both cores receive the same additional readout freedom. One shared direction nearly closes the gap. Two directions give fixed-depth training lower mean error, and four directions make this ordering consistent across all three initialization pairs. At four directions, the mean errors are 0.0839 and 0.0887 for fixed-depth and multi-depth training, respectively. The paired difference of 0.0048 has a conditional 95% interval of $[0.0025, 0.0070]$.

The reversal occurs well before the interface reaches independent heads. Two directions remove 96.94% of the fixed-depth models' sharing cost and 91.45% of the multi-depth models' cost, measured relative to the independent-head reference. Each of the six models exceeds 90% recovery. The same shared directions serve all sixteen depths; only their coefficients change. This gives a compact account of the mismatch across depths: most of the error imposed by a common head can be removed by a small set of coordinated changes to its feature weights.

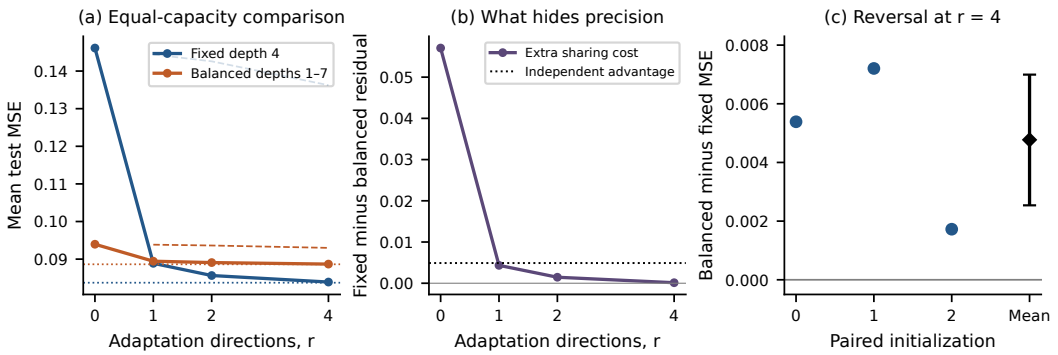

Figure 1: A few shared adaptation directions reverse the training comparison. Mean squared error averages sixteen depths and 4,096 test episodes. (a) Learned directions approach the independent-head reference; random directions leave much larger errors. (b) The extra sharing cost of fixed-depth training falls below its independent-head advantage. (c) Four directions give fixed-depth training lower error in all three initialization pairs. The interval resamples episodes conditional on these models.

Table 1: The reversal persists on a second set of 4,096 test episodes and requires more than scalar calibration. Both training regimes receive the same readout family. Errors average all sixteen depths and three paired models; lower is better. We reevaluate the existing capacity heads without refitting.

Interface	Fixed 4	Balanced 1–7
Common head	.140347	.089502
Frozen head + gain/offset	.125373	.089068
Joint slope direction + gain/offset	.113240	.088002
Two general adaptation directions	.081130	.084887
Four general adaptation directions	.079361	.084607
Independent-head reference	.079211	.084544

The absolute errors explain why this recovery changes the model comparison. Fixed-depth training reaches a lower independent-head error, but initially pays a much larger price for sharing the head. At zero adaptation capacity, its extra sharing cost is 0.0571, larger than its independent-head advantage of 0.0049. At four directions, that extra cost falls to 0.0001. The comparison therefore crosses the threshold in Equation (12): once the interface accommodates the main differences across depths, the fixed-depth core's more accurate fitted predictions determine the ranking. The recurrent states remain fixed throughout this intervention.

We use two controls to understand what the learned directions provide. Random directions test whether additional degrees of freedom are sufficient on their own. At rank two, random directions leave mean errors of 0.1426 and 0.0936, compared with 0.0856 and 0.0891 for learned directions. Most of the gain depends on choosing directions that suit the task and recurrent states. This agrees with the capacity decomposition: an interface succeeds when it captures differences among useful predictors across depths.

Scalar calibration tests a more specific explanation. Perhaps every depth already exposes the same useful scalar prediction, and only its scale and offset change. We therefore fit depth-specific gains and offsets to a frozen common prediction, and also allow the common prediction direction to be jointly relearned. These interfaces use $a_t \widehat{y}_t^{\text{shared}} + b_t$ and $s_t v^\top Z_t + b_t$, respectively. Both retain a single prediction direction, whereas two general adaptation directions allow the head to combine features in more varied ways. Within each family, the two training regimes receive identical fitting resources.

Table 1 evaluates these controls on a second set of 4,096 fresh episodes. For the fixed-depth cores, gain and offset reduce error from 0.1403 to 0.1254; relearning the common direction reduces it further to 0.1132. Two general directions reach 0.0811, close to the independent-head error of 0.0792. They recover 96.86% of the sharing cost, compared with 24.49% for frozen-output calibration and 44.34% for the joint-direction model. The multi-depth cores show the same ordering of these interfaces.

Table 2: Multi-depth training improves compatibility between depths 4 and 16. Errors average 4,096 test episodes and three paired models; lower is better. Sharing cost weights the two depths equally, and Δ denotes balanced minus fixed. Independent heads reveal the prediction accuracy available at each depth.

Readout / quantity	Fixed 4	Balanced 1–7	Δ
Native, depth 4	.037150	.064487	+.027337
Native, depth 16	.530162	.130497	-.399665
Separate, depth 4	.036215	.062651	+.026436
Separate, depth 16	.065144	.066791	+.001647
Separate, average ($\widehat{L}_{\mathrm{sep}}$)	.050679	.064721	+.014042
Shared, average ($\widehat{L}_{\mathrm{shared}}$)	.125813	.073274	-.052539
Sharing cost ($\widehat{\Gamma}_{1/2}$)	.075134	.008553	-.066581
Native-to-shared gap ($\widehat{E}_{\mathrm{native}}$)	.157843	.024218	-.133625

Thus the main gain comes from adapting which state features contribute to the prediction, beyond calibrating a common scalar output.

This second evaluation also confirms the model comparison. Multi-depth training retains lower error under both calibration controls, while fixed-depth training attains lower mean error with two general directions. Four directions again give fixed-depth training lower error in all three initialization pairs. The agreement between the two test sets ties the reversal to the readout family: a common or scalar-calibrated interface favors multi-depth training, while a small adaptive family exposes the fixed-depth core's advantage. Appendix I reports the full per-pair results.

5.3 Multi-depth training improves compatibility across depths

The capacity curves establish that the two training regimes differ in how easily one interface can read their states. We now examine where this difference enters the risk decomposition. We focus on depth 4, the supervised depth of fixed training, and depth 16, which lies beyond the supervision range of both regimes. Native, refitted common, and independent heads provide three views of the same pair of states: the model's original output, the accuracy available through one fitted interface, and the accuracy available when each depth has its own interface.

Table 2 shows a large improvement in the native depth-16 output under multi-depth training: error falls from 0.5302 to 0.1305. Independent heads read the same late states with similar errors, 0.0651 and 0.0668. Fixed-depth training also supports a more accurate independent prediction at depth 4, with error 0.0362 compared with 0.0627. The available prediction accuracy and the native output therefore give different views of these cores. The native advantage of multi-depth training becomes clear when we measure their compatibility with one common head.

At equal depth weights, the sharing cost falls from 0.0751 to 0.0086, a reduction of 88.6%. The reduction occurs in each initialization pair. A refitted common head consequently reaches mean error 0.0733 on the multi-depth cores, compared with 0.1258 on the fixed-depth cores. The decomposition also attributes part of the native improvement to the original head's fit: its gap from the best fitted common head falls from 0.1578 to 0.0242. Multi-depth supervision thus produces states that work better together under one output head and a native head that makes better use of that compatibility. This is the aspect of the training objective identified in Proposition 1.

The independent-head advantage of fixed-depth training and the shared-head advantage of multi-depth training are complementary parts of the result. They explain why the capacity curves cross. Multi-depth training places predictions from different stopping depths into a form that a common interface can use more accurately. Fixed-depth training supports more precise fitted predictions at its supervised depth, and a few shared adaptation directions make that precision useful across the evaluation depths. The choice of interface determines which advantage appears in the aggregate score.

We further test whether the common-head advantage depends on giving the two depths equal weight. For each frozen core, we fit common heads to $p\widehat{R}_4 + (1-p)\widehat{R}_{16}$, varying p from zero to one. This

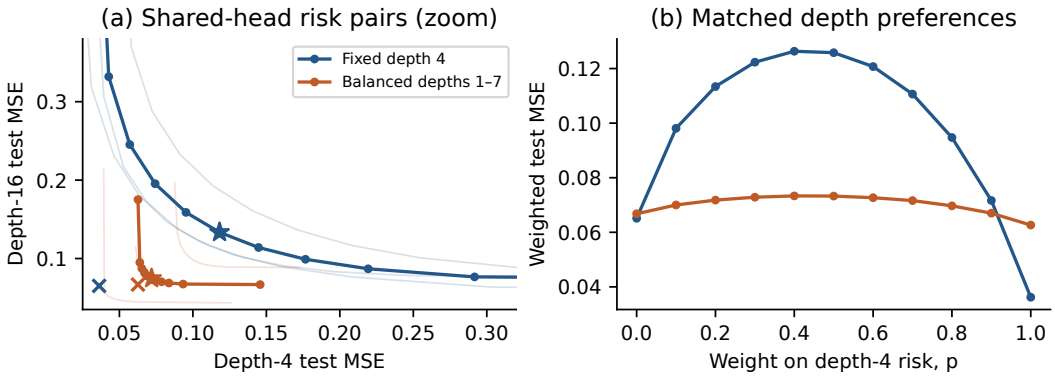

Figure 2: Multi-depth training supports a more accurate common readout across depth preferences. Left: depth-4 and depth-16 errors as the fitting weight p varies. Thin lines show individual models, thick lines show means, stars mark equal weights, and crosses mark independent-head references. The central range is shown; Appendix D gives all endpoints. Right: mean weighted error at the same depth preference p.

traces the achieved tradeoff between reference-depth and late-depth accuracy while preserving a fully shared output. The experiment asks whether a different readout preference can reproduce the effect of multi-depth training within the fixed-depth core.

Figure 2 shows that multi-depth training improves joint-depth prediction across the interior of this range. At equal weights, the depth-4 and depth-16 errors change from $(0.1181, 0.1335)$ to $(0.0719, 0.0746)$. When the fixed-depth core places more weight on depth 4, it can reach a similar reference error of 0.0743, but its depth-16 error is 0.1954. The multi-depth core therefore supports a substantially better late prediction at comparable reference-depth accuracy. It has lower mean weighted error at every interior grid weight $p \in \{0.1, \ldots, 0.9\}$, and its sampled curve improves both errors for seven of the fixed core's eleven sampled pairs.

At the endpoint that targets depth 4 alone, fixed-depth training retains its accuracy advantage. This endpoint links the tradeoff curves back to the capacity result: the fixed core contains accurate depth-specific predictors, while the multi-depth core makes predictors across depths easier to share. A new common head explores different compromises between those predictors; a small adaptive head can instead accommodate their differences. The two interventions answer distinct parts of the same question about depth compatibility.

These comparisons connect depth robustness to the interface used to measure it. Multi-depth training improves predictions through a common head, while a small adaptive family exposes the fixed-depth core's lower fitted risk. The capacity curve shows how much flexibility separates these two evaluations.

6 Conclusion

The risk decomposition and experiments show how depth supervision changes both available prediction accuracy and compatibility with a shared head. Multi-depth training improves the common interface, while four shared adaptation directions reverse the comparison in all three initialization pairs. Calibration controls confirm that this reversal requires more than rescaling one prediction.

The central implication is methodological: accuracy at a stopping depth is jointly determined by the information represented in the recurrent state and the constraints imposed by the output interface. Evaluations using only a native or fully shared head can therefore credit a training regime for making depth-specific predictors compatible, whereas fully independent probes remove that constraint altogether. Capacity-controlled readouts provide a middle ground by measuring how quickly useful depth-specific predictions become accessible as a common interface is relaxed.

AI USE STATEMENT

We used ChatGPT and OpenAI Codex to polish the writing, check the theoretical proofs, and assist with the design and implementation of the experimental code. AI assistance also supported proof improvement, analysis of the synthetic experiments, and literature search. The authors verified the proofs, code, numerical results, and references and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Appendix E describes the architecture, data generation, training, readout fitting, and evaluation splits. Appendices A and B provide proofs, assumptions, and counterexamples, while Appendix G maps the main analyses to the anonymous supplementary code. The supplement provides source code and execution instructions for synthetic-data generation, matched core training, frozen-state readout fitting, evaluation, and numerical checks. Appendices H and I describe the analysis protocols.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

W. N. Anderson, Jr. and R. J. Duffin. Series and parallel addition of matrices. *Journal of Mathematical Analysis and Applications*, 26(3):576–594, 1969.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-Task Feature Learning. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. *arXiv preprint arXiv:1909.01377*, 2019.

András Balogh and Márk Jelasity. How not to Stitch Representations to Measure Similarity: Task Loss Matching versus Direct Matching. *arXiv preprint arXiv:2412.11299*, 2024.

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to Ponder. *arXiv preprint arXiv:2107.05407*, 2021.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting Model Stitching to Compare Neural Representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal Transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Carl Eckart and Gale Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. *arXiv preprint arXiv:2208.01066*, 2022.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. *arXiv preprint arXiv:2502.05171*, 2025.

Alex Graves. Adaptive Computation Time for Recurrent Neural Networks. *arXiv preprint arXiv:1603.08983*, 2016.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529. PMLR, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*, 2019.

Mark Oskin. Off-Axis, On Purpose: Where a Transformer Computes Concepts and Why it Does So. *arXiv preprint arXiv:2608.10251*, 2026.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622. Association for Computational Linguistics, 2020.

Andrei Cristian Popescu, Haitz Sáez de Ocáriz Borde, and Pietro Liò. Adaptive Depth in Looped Transformers: Diagnosing Learned Halting Gates and Trajectory Readouts. *arXiv preprint arXiv:2607.20519*, 2026.

Rituraj Sharma and Tu Vu. Dense Supervision Is Not Enough: The Readout Blind Spot in Looped Language Models. *arXiv preprint arXiv:2606.24898*, 2026.

Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 35151–35174. PMLR, 2023.

Wenlong Wang and Fergal Reid. Looped Transformers under the Jacobian Lens: Does the Global Workspace Survive Recurrence? *arXiv preprint arXiv:2609.01924*, 2026.

Liu Yang, Kangwook Lee, Robert D. Nowak, and Dimitris Papailiopoulos. Looped Transformers are Better at Learning Learning Algorithms. In *International Conference on Learning Representations*, 2024.

Xiao-Wen Yang, Ziyu Han, Xi-Hua Zhang, Wen-Da Wei, Jie-Jing Shao, Lan-Zhe Guo, and Yu-Feng Li. Stabilizing Recurrent Dynamics for Test-Time Scalable Latent Reasoning in Looped Language Models. *arXiv preprint arXiv:2605.26733*, 2026.

A PROOFS AND RECURRENT EXAMPLES

We begin with head-transfer evidence that complements the capacity experiments, then establish the regression identities and recurrent examples used in the main text. The proofs make explicit when feature geometry and the choice of supervised depths affect sharing cost.

A.1 HEAD TRANSFER AND SUPPORTING STATE RECOVERY

Head transfer identifies which depths can already use one another's fitted predictors. It complements the capacity curves by showing where a common interface encounters disagreement.

Finite-sample risk differences can be negative; tables and saved arrays retain those values, while the heatmap caption specifies display clipping. We evaluate all 16×16 source–target head pairs (Figure 3). Native error in the fixed-depth core rises after its training depth, while separately fitted late heads remain accurate. A depth-4 head incurs excess MSE of 0.514749 at depth 16 for the fixed core and 0.108383 for the balanced core. In the reverse direction, excess MSE falls from 1.261267 to 0.083139. Equation (7) explains the asymmetry through the target depth's feature metric; a single Euclidean distance between heads cannot describe both directions.

Supporting state recovery. Affine maps fitted only to paired early and late states recover prediction accuracy on a separate 4,096-episode test; shuffled pairings fail (Appendix C, Table 4). At the selected penalty, the map composed with the early head equals ridge regression onto that head's predictions. This intervention recovers predictions from a supervised teacher without new answer labels and complements the answer-supervised capacity analysis.

A.2 POPULATION PROJECTION IDENTITY AND RISK ACCOUNTING

The transfer costs and sharing penalties both follow from the same least-squares identity. We establish it for singular as well as full-rank feature distributions.

Assume $\mathbb{E}\,\|X_t\|^2 < \infty$ and $\mathbb{E}Y^2 < \infty$. Define $\Sigma_t = \mathbb{E}X_tX_t^\top$, $b_t = \mathbb{E}X_tY$, and $c = \mathbb{E}Y^2$. For any $a \in \ker \Sigma_t$, we have $a^\top X_t = 0$ almost surely, which implies $a^\top b_t = 0$. Hence $b_t \in \operatorname{range}(\Sigma_t)$, so $\Sigma_t v_t = b_t$ has a solution even when the features are singular. We expand the risk to obtain

$$R_t(v) = v^\top \Sigma_t v - 2v^\top b_t + c, \tag{16}$$

$$R_t(v) - R_t(v_t) = (v - v_t)^\top \Sigma_t (v - v_t). \tag{17}$$

A common minimizer solves $\overline{\Sigma}\overline{v} = \overline{b}$, where $\overline{\Sigma} = \sum_t \pi_t \Sigma_t$ and $\overline{b} = \sum_t \pi_t b_t$; the same range argument proves existence. A common head is feasible at each depth, so $L_{\text{shared}} \geq L_{\text{sep}}$. The native head is also a feasible common head, so $L_{\text{native}} \geq L_{\text{shared}}$. The sum of these risk differences gives Equation (2). All three terms can change when training changes the features. The identity decomposes risk under one evaluation distribution; it does not identify causal mediation effects.

A.3 PROOF OF PROPOSITION 1

The training objective applies the preceding risk decomposition to the supervised depths. The key question is when their individually optimal predictors admit a common choice.

Fix θ and restrict every sum to the positive-weight support of q. The preceding range argument applies to each depth and to their weighted mixture, so every independent-head minimum and the shared-head minimum are attained. Subtract and add the shared optimum in J_q to obtain Equation (4). The remaining gap E_q is nonnegative and vanishes at a common minimizer.

If a head belongs to every $\mathcal{A}_t(\theta)$, its weighted risk equals B_q, so $\Gamma_q = 0$. Conversely, choose a common minimizer $\overline{v}$ when $\Gamma_q = 0$. Each difference $R_t^\theta(\overline{v}) - \min_a R_t^\theta(a)$ is nonnegative. Their positive-weight sum is zero, so every difference is zero and $\overline{v}$ lies in every optimal set. This reasoning uses optimal sets rather than unique coefficients and therefore includes singular features. For a singleton support, the intersection is a nonempty independent optimal set. Training and evaluation can use different depth distributions; the identity holds separately under each specified distribution.

A.4 Proof of Proposition 3

The circle example makes the capacity curve explicit because the optimal heads occupy a single plane. We identify those heads and compute the two nonzero eigenvalues that determine their approximation error.

Orthogonality gives $\mathbb{E}Z_tZ_t^\top = I_m$ and $Y = (A^te_1)^\top Z_t$. Since $\mathbb{E}Z_t = \mathbb{E}Y = 0$, the augmented feature $(Z_t^\top, 1)^\top$ has second moment I_{m+1}, and the unique optimal affine head is $v_t = ((A^te_1)^\top, 0)^\top$. Each independent risk is zero. Put $\alpha_t = 2\pi t/K$. The two nonzero coordinates of v_t are $\cos\alpha_t$ and $\sin\alpha_t$.

For $K \geq 3$, both $\sum_{t=1}^K e^{i\alpha_t} = 0$ and $\sum_{t=1}^K e^{2i\alpha_t} = 0$. These identities imply $\overline{v} = 0$, $K^{-1}\sum_t \cos^2\alpha_t = K^{-1}\sum_t \sin^2\alpha_t = 1/2$, and $\sum_t \cos\alpha_t \sin\alpha_t = 0$. Hence $M_\pi M_\pi^\top$ has exactly two nonzero eigenvalues, both $1/2$. Equation (14), proved independently in Appendix B, gives costs 1, $1/2$, and 0 for zero, one, and two directions. The spectral lower bound includes arbitrary affine adaptation, including intercept changes. Thus an intercept direction cannot improve these optima.

The circle example uses a common feature metric. For two arbitrary depths, the following expression instead keeps both depth-specific metrics and measures the cost of their disagreement.

Proposition 4 (Two-depth sharing cost)**.** *For depths s, t with weights $p, 1-p$, $0 < p < 1$, put $Q = p\Sigma_s$, $S_0 = (1-p)\Sigma_t$, and $\delta = v_s - v_t$. Then*

$$\Gamma_p = \delta^\top (Q : S_0)\delta, \qquad Q : S_0 = Q - Q(Q+S_0)^\dagger Q. \tag{18}$$

If both second-moment matrices are positive definite, $Q : S_0 = (Q^{-1} + S_0^{-1})^{-1}$. In the semidefinite case, the displayed pseudoinverse expression applies and the result is independent of the chosen optimal-head representatives.

The matrix parallel sum is classical (Anderson & Duffin, 1969).

A.5 Proof of Proposition 4

For two depths, the common-head problem reduces to one quadratic minimization. We keep the semidefinite case throughout so that invisible feature directions contribute zero cost.

Let $x = v - v_t$ and $\delta = v_s - v_t$. The sharing cost equals

$$\min_x\{(x-\delta)^\top Q(x-\delta) + x^\top S_0 x\} = \min_x\{x^\top(Q+S_0)x - 2x^\top Q\delta + \delta^\top Q\delta\}. \tag{19}$$

For positive semidefinite Q, S_0, the identity $\ker(Q+S_0) = \ker Q \cap \ker S_0$ implies that $Q\delta$ lies in the range of $Q+S_0$. We substitute the minimizer $x = (Q+S_0)^\dagger Q\delta$ to obtain

$$\Gamma_p = \delta^\top [Q - Q(Q+S_0)^\dagger Q]\delta. \tag{20}$$

For positive-definite matrices, the bracket equals $(Q^{-1}+S_0^{-1})^{-1}$. A change to v_s within $\ker\Sigma_s$, or to v_t within $\ker\Sigma_t$, leaves the original minimization unchanged. The result therefore does not depend on the chosen optimal-head representatives, as required by the classical parallel-sum properties (Anderson & Duffin, 1969).

A singular counterexample. Let Y be uniform on $\{-\sqrt{2}, \sqrt{2}\}$, $X_s = (Y, 0, 1)^\top$, $X_t = (0, Y, 1)^\top$, and $p = 1/2$. Choose optimal heads $v_s = (1,0,0)^\top$ and $v_t = (0,1,0)^\top$. Then $Q = \mathrm{diag}(1, 0, 1/2)$, $S_0 = \mathrm{diag}(0, 1, 1/2)$, and $\delta = (1, -1, 0)^\top$. The correct parallel sum is $\mathrm{diag}(0, 0, 1/4)$, yielding zero sharing cost: $(1,1,0)^\top$ predicts exactly at both depths. The incorrect replacement $(Q^\dagger + S_0^\dagger)^\dagger = \mathrm{diag}(1, 1, 1/4)$ would instead give cost 2.

Common metric. For $\Sigma_t = \Sigma$ at every depth, $\overline{v} = \sum_t \pi_t v_t$ is a shared minimizer and

$$\Gamma_\pi = \sum_t \pi_t \|v_t - \overline{v}\|_\Sigma^2 = \frac{1}{2}\sum_{s,t} \pi_s\pi_t \|v_s - v_t\|_\Sigma^2, \qquad \|a\|_\Sigma^2 = a^\top\Sigma a. \tag{21}$$

Equation (7) gives $C^*_{s\to t} = \|v_s - v_t\|_\Sigma^2$ and proves Equation (8). The proof also holds for a common semidefinite Σ, with coefficients identified on its range. Under different second moments, unweighted

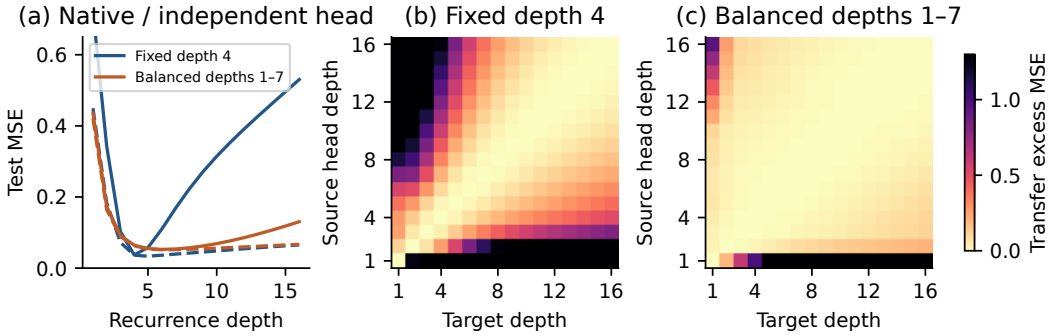

Figure 3: **Depth-specific predictability differs from transfer compatibility.** All panels use the same new episodes. Left: native heads (solid) and separate affine heads (dashed). Center/right: excess target-depth MSE from a source-depth head. The common color scale clips values above 1.3 and displays negative finite-sample values at zero; saved arrays retain every value.

coefficient variance need not equal sharing cost, and transfer can be asymmetric. Singular depth-specific metrics introduce another distinction: a source-null direction can affect target predictions, so $C^*_{s\to t}$ can depend on the source-head representative. Shared cost remains unchanged by the choice of representative, while the experimental ridge rule selects a definite head for transfer. Estimated-head comparisons also include estimation error.

A.6 An exact frequency characterization for repeated computation

The circle construction isolates one rotation frequency. We now allow several frequencies and ask how the distribution of supervised depths weights their contribution to sharing cost.

Consider an explicit information-preserving recurrence

$$Z_t = A^t Z_0, \qquad A^\top A = I, \qquad \mathbb{E}[Z_0 Z_0^\top] = I, \qquad Y = u^\top Z_0. \tag{22}$$

In this example, heads are linear and have no intercept. Each depth admits zero risk through $w_t = A^t u$, so any sharing cost comes entirely from incompatibility.

Theorem 1 (Depth distribution and recurrence frequency)**.** *Under Equation (22), for linear predictions $w^\top Z_t$, the optimal common head and sharing cost are*

$$\overline{w}_\pi = \sum_t \pi_t A^t u, \tag{23}$$

$$\Gamma_\pi = \|u\|^2 - \|\overline{w}_\pi\|^2 = \frac{1}{2}\sum_{s,t} \pi_s \pi_t \left\|A^s u - A^t u\right\|^2. \tag{24}$$

Let $A = U\,\mathrm{diag}(e^{i\theta_j})U^*$ be the complex spectral decomposition, with $c = U^* u$ and $\varphi_\pi(\theta) = \sum_t \pi_t e^{it\theta}$. Then

$$\Gamma_\pi = \sum_j |c_j|^2 \left(1 - |\varphi_\pi(\theta_j)|^2\right). \tag{25}$$

Equation (25) separates task energy across recurrent modes, each mode's rotation per step, and the distribution of supervised depths. The same transition can support a common head for one task direction but require changing heads for another. Even preservation of the entire state therefore does not guarantee a common prediction interface.

Corollary 1 (Consecutive and spaced supervision)**.** *Under Theorem 1, for uniform supervision on $1, \ldots, T$ with integer $T \geq 1$,*

$$\Gamma_T = \sum_j |c_j|^2 \left[1 - \left|\frac{\sin(T\theta_j/2)}{T\sin(\theta_j/2)}\right|^2\right], \tag{26}$$

with continuous extension at multiples of 2π. *For any finite support with at least two depths, let* $g = \gcd\{t - s_0 : t \in S\}$ *and* $s_0 = \min S$. *Then*

$$\Gamma_\pi = 0 \quad \Longleftrightarrow \quad A^g u = u. \tag{27}$$

Consecutive supervised depths require one-step invariance of the task direction for zero sharing cost. Depths spaced four steps apart permit a four-step cycle, so agreement at supervised depths can hide disagreement between them (Figure 4). At finite T, Equation (26) weights non-invariant task energy by frequency, and its value need not increase with T. Appendix A gives examples of both effects.

A.7 Proof of Theorem 1

Orthogonality puts every depth in the same feature geometry. We can therefore express sharing cost as the variance of its optimal heads and evaluate that variance in the recurrent eigenbasis.

Orthogonality implies $\mathbb{E}Z_t Z_t^\top = I$ and

$$R_t(w) = \mathbb{E}[((A^{t\top} w - u)^\top Z_0)^2] = \left\|A^{t\top} w - u\right\|^2 = \left\|w - A^t u\right\|^2. \tag{28}$$

The unique optimal head is therefore $w_t = A^t u$, with zero risk. Let $\overline{w} = \sum_t \pi_t w_t$ and complete the square:

$$\sum_t \pi_t \left\|w - w_t\right\|^2 = \left\|w - \overline{w}\right\|^2 + \sum_t \pi_t \left\|w_t\right\|^2 - \left\|\overline{w}\right\|^2. \tag{29}$$

Every head satisfies $\|w_t\| = \|u\|$, which proves the first risk expression. The expansion of pairwise squared distances gives the second expression in Equation (24). A real orthogonal matrix is normal over $\mathbb{C}$ and therefore has a decomposition $A = U \operatorname{diag}(e^{i\theta_j}) U^*$ with unitary U. In this basis, coordinate j of $\overline{w}$ equals $c_j \varphi_\pi(\theta_j)$. Since unitary transformations preserve norms, the difference between $\sum_j |c_j|^2$ and $\sum_j |c_j|^2 |\varphi_\pi(\theta_j)|^2$ gives Equation (25). The complex coordinates represent real rotation planes compactly.

Affine heads and nonzero means. Theorem 1 optimizes linear heads. If additionally $\mathbb{E}Z_0 = 0$, then $\mathbb{E}Z_t = \mathbb{E}Y = 0$ and, for any common intercept b,

$$\sum_t \pi_t \mathbb{E}[(w^\top Z_t + b - Y)^2] = \sum_t \pi_t \mathbb{E}[(w^\top Z_t - Y)^2] + b^2. \tag{30}$$

The optimal intercept is therefore $b = 0$, and the theorem extends to affine heads under the zero-mean condition. Nonzero means can change the result. Let $A = -1$, $u = 1$, $\mathbb{E}Z_0 = \mu$, and $\mathbb{E}Z_0^2 = 1$, with equal weights on two consecutive depths. The shared affine risk is $w^2 + (b - \mu)^2 + 1 - \mu^2$, whose minimum is $1 - \mu^2$. In contrast, the linear-head sharing cost equals 1, although each depth still admits exact prediction separately. This example also shows why second moments and covariances are not interchangeable here.

A.8 Proof of Corollary 1 and examples

The spectral formula becomes especially simple for consecutive depths. For spaced depths, the same zero-cost condition is determined by the common period of their depth differences.

The geometric series yields

$$T^{-1} \sum_{t=1}^{T} e^{it\theta} = e^{i(T+1)\theta/2} \frac{\sin(T\theta/2)}{T \sin(\theta/2)}. \tag{31}$$

The modulus of this expression gives Equation (26). Positive weights ensure that the pairwise formula vanishes exactly when $A^t u = A^{s_0} u$ for every $t \in S$. Invertibility reduces this condition to $A^{t - s_0} u = u$. If $A^g u = u$, every depth difference is a multiple of g, so the condition follows. Conversely, Bézout's identity expresses g as an integer combination of depth differences. Powers that fix u also fix it under inversion and composition, which gives $A^g u = u$.

For a quarter-turn rotation and a unit task direction, supervision at $1, 5, 9, 13$ gives zero cost, while supervision at $1, 2, 3, 4$ gives cost one. With two consecutive depths, $\Gamma_2 = \sin^2(\theta/2)$ can be arbitrarily

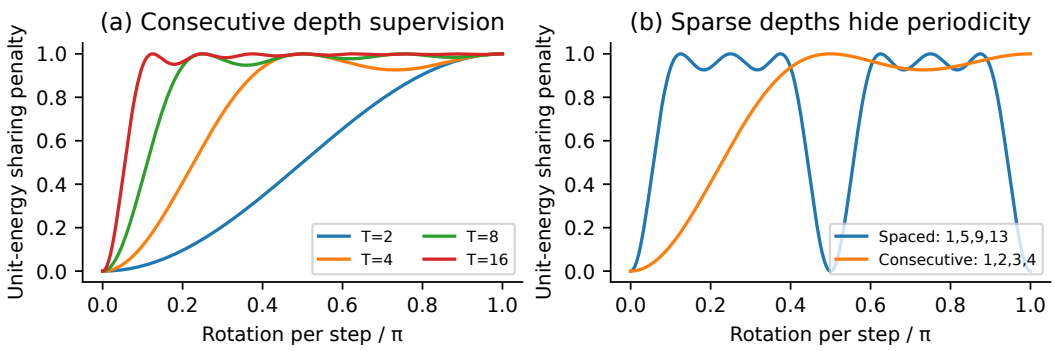

Figure 4: **The depth distribution filters recurrent frequencies.** Exact sharing cost for a unit task direction under a two-dimensional rotation. Left: uniform supervision on consecutive depths. Right: supervision at $1, 5, 9, 13$ misses a quarter-turn cycle ($\theta = \pi/2$). The curves describe this linear system; they do not estimate the trained transformers' spectra.

small even when all task energy lies outside the invariant subspace. For fixed finite-dimensional A, uniform consecutive averaging removes every nonzero frequency as $T \to \infty$. Only in that limit does the penalty equal all non-invariant task energy. Finally, $A = -I$ and unit u give $\Gamma_2 = 1$ and $\Gamma_3 = 8/9$, so cost need not increase with the depth range.

Invertible, non-orthogonal recurrence. For invertible A, a full-rank initial second moment, and $Y = u^\top Z_0$, exact heads take the form $w_t = A^{-t\top} u$. The same greatest-common-divisor argument shows that an exact common head over S exists precisely when $A^{-g\top} u = u$. The frequency formula additionally requires orthogonality because a general transition can change the feature metric. Singular support also prevents a unique coefficient trajectory. For $Z_s = (x, 0)$, $Z_{s+1} = (x, x)$, and $Y = x$, the minimum-norm heads are $(1, 0)$ and $(1/2, 1/2)$. An invertible transport can instead produce the equally valid later head $(1, 0)$. The relevant object is the prediction function, rather than a particular minimum-norm representative.

Local derivatives and fitted readouts. For differentiable functions with $f_t = f_{t+1} \circ F$, the chain rule gives $\nabla f_t(h) = J_F(h)^\top \nabla f_{t+1}(F(h))$. Separately fitted linear heads generally do not satisfy this functional relation. For example, let h be uniform on $\{-2, -1, 1, 2\}$, with $Y = h$ and $F(h) = h^3$. The optimal slopes are $w_0 = 1$ and $w_1 = \mathbb{E}[h^4]/\mathbb{E}[h^6] = 17/65$, yet $J_F(h)w_1 = 51h^2/65$ differs from w_0. The exact nonlinear decoder $f_1(z) = z^{1/3}$ does satisfy the chain rule at the support points. Thus local function gradients and global regression coefficients require different analyses. In a transformer, the full token state evolves recurrently, while the normalized query feature alone need not define a closed transition. A derivative analysis must therefore specify both the state and the decoded function.

B Shared adaptation capacity

The main capacity formula allows feature geometry to change with depth. We prove it by solving for the depth coefficients and common head at a fixed adaptation subspace, then examine the simpler common-metric case.

B.1 Heterogeneous geometry: proof of Proposition 2

We first optimize the depth-specific coefficients while holding the shared directions fixed. The remaining objective is quadratic in the common head, which gives the stated formula even when the feature metrics are singular.

Assume finite second moments and strictly positive depth weights. The projection identity gives Equation (9) without requiring unique optimal heads or a common feature metric. For fixed U and w, define $d_t = v_t - w$, $K_t = U^\top \Sigma_t U$, and $S_t = \Sigma_t^{1/2}$. Let Π_t project orthogonally onto $\mathrm{range}(S_t U)$.

Least squares over the depth coefficient gives

$$\inf_{a_t} \|S_t(Ua_t - d_t)\|^2 = d_t^\top S_t(I - \Pi_t)S_t d_t = d_t^\top P_t d_t, \tag{32}$$

$$P_t = S_t(I - \Pi_t)S_t = \Sigma_t - \Sigma_t U K_t^\dagger U^\top \Sigma_t. \tag{33}$$

The projector expression implies $P_t \succeq 0$. A minimizing coefficient is $a_t = K_t^\dagger U^\top \Sigma_t d_t$. To verify the range condition, take $z \in \ker K_t$: then $S_t U z = 0$ and $z^\top U^\top \Sigma_t d_t = 0$. Hence the right-hand side of the normal equation lies in the range of K_t.

After we eliminate the depth coefficients, the objective becomes

$$w^\top A_U w - 2w^\top b_U + \sum_t \pi_t v_t^\top P_t v_t. \tag{34}$$

If $x \in \ker A_U$, then $0 = x^\top A_U x = \sum_t \pi_t x^\top P_t x$. Each summand is nonnegative and every weight is positive, so $P_t x = 0$ at every depth. Consequently $x^\top b_U = 0$, which proves $b_U \in \mathrm{range}(A_U)$. The choice $w = A_U^\dagger b_U$ then gives Equation (11). If we replace v_t by $v_t + n_t$ with $n_t \in \ker \Sigma_t$, then $S_t n_t = P_t n_t = 0$, leaving the expression unchanged. The infimum over U completes the proof. Both coefficient problems attain their minima for fixed U; the outer search over subspaces can fail to attain its infimum.

Why the outer optimization uses an infimum. Singular depth metrics can prevent attainment even when the infimum is zero. Let Y be uniform on $\{-1, 1\}$ and use equal weights on three depths with scalar features

$$Z_1 = Y - 1, \qquad Z_2 = (Y - 1)/2, \qquad Z_3 = 0, \qquad X_t = (Z_t, 1)^\top. \tag{35}$$

The shared transition $F(z) = -z(z + 1)/2$ produces these features. The first two second-moment matrices are positive definite, with unique optimal heads $v_1 = (1, 1)^\top$ and $v_2 = (2, 1)^\top$. At depth three, any slope with zero intercept is optimal. For every $\varepsilon > 0$, choose

$$w = (1, 1)^\top, \quad U_\varepsilon = (1, \varepsilon)^\top, \quad (a_1, a_2, a_3) = (0, 1, -1/\varepsilon). \tag{36}$$

The heads $w + U_\varepsilon a_t$ predict exactly at depth one, have constant error ε at depth two, and achieve the independent optimum at depth three. Their excess risk is therefore $\varepsilon^2/3$, giving $C_\pi(1) = 0$. A zero-cost interface would need a line through $(1, 1)^\top$ and $(2, 1)^\top$. Every point on that line has intercept one, which conflicts with the zero intercept required at depth three. Thus no one-direction interface attains the infimum, although the fixed-U minima in Proposition 2 remain attained.

Capacity and comparison. The nested interface families make $C_\pi(r)$ nonnegative and nonincreasing in r. Choose $w = v_{s_0}$ and let U span the differences $v_t - v_{s_0}$. This construction represents every optimal head with at most $\min(D, |S| - 1)$ directions. The difference between $B_M + C_M(r)$ and $B_F + C_F(r)$ gives Equation (12). Since $C_M(r) \geq 0$, the condition $C_F(r) < \delta$ suffices for fixed-core superiority. However, $C_F(r) - C_M(r)$ need not be monotone, so the decomposition does not imply a unique crossing. The small capacity at which the observed comparison reverses remains an empirical finding.

Population optima, fitted references, and penalized objectives. For any fitted reference $\widehat{v}_t$, put $e_t = v - \widehat{v}_t$ and $g_t = \mathbb{E}[X_t(Y - X_t^\top \widehat{v}_t)]$. Direct expansion gives

$$R_t(v) - R_t(\widehat{v}_t) = e_t^\top \Sigma_t e_t - 2e_t^\top g_t. \tag{37}$$

The residual term vanishes at a population least-squares optimum evaluated on the same distribution. A fitted or calibration-selected ridge head need not satisfy this condition on test data. For fixed $P \succeq 0$ and $\lambda \geq 0$, define $J_t(v) = \widehat{R}_t(v) + \lambda v^\top P v$. Its minimizer solves $(\widehat{\Sigma}_t + \lambda P)v = \widehat{b}_t$, and the completed-square identity uses $\widehat{\Sigma}_t + \lambda P$. The heterogeneous formulas therefore also describe this penalized fitting objective. Our calibration procedure can select different penalties across capacities, so one fixed-penalty identity does not describe the full test curve. Test-risk differences from a fitted independent head remain achieved-risk comparisons and can be negative.

Coordinates and numerical checks. A common invertible feature transform preserves the unpenalized function family and affine head dimension when coefficients and metrics transform together. Depth-specific transforms change the sharing constraint, while an unchanged penalty under new coordinates changes the estimator. We compare the fixed-U formula with stacked least squares on forty deterministic positive-definite and rank-deficient problems. The maximum absolute discrepancy is 1.96×10^{-14}. Factorizations evaluate the residual projectors without subtractive cancellation. The pseudoinverse threshold is 10^{-10} relative to the largest eigenvalue, with the corresponding 10^{-5} singular-value cutoff in the stacked solve. These numerical checks assess the algebra and implementation, separately from learned-model evidence or optimization over U.

B.2 Common-metric specialization

When every depth has the same metric, one coordinate change turns all risk terms into ordinary squared distances. The capacity problem then reduces to approximating a matrix of centered optimal heads.

For a common positive-definite second moment Σ, define $\overline{v} = \sum_t \pi_t v_t$ and let σ_j be the singular values of the matrix with columns $\sqrt{\pi_t}\Sigma^{1/2}(v_t - \overline{v})$. Classical low-rank approximation (Eckart & Young, 1936) gives

$$C_\pi(r) = \sum_{j>r} \sigma_j^2. \tag{38}$$

Proof. Suppose $\Sigma_t = \Sigma \succ 0$, and define $b_t = \Sigma^{1/2} v_t$ and $\overline{b} = \sum_t \pi_t b_t$. After we multiply the candidate head and basis by $\Sigma^{1/2}$, the excess risk becomes

$$\min_{b,V,\{a_t\}} \sum_t \pi_t \left\| b + V a_t - b_t \right\|^2. \tag{39}$$

Center the coefficients so that $\sum_t \pi_t a_t = 0$; the optimal common component then equals $b = \overline{b}$. Let M contain columns $\sqrt{\pi_t}(b_t - \overline{b})$, and let P_V project onto a candidate column space of dimension at most r. For that space, the best approximation is $P_V M$, with squared error $\mathrm{tr}(MM^\top) - \mathrm{tr}(P_V MM^\top)$. In an eigenbasis of $MM^\top$, the second trace equals $\sum_j \sigma_j^2 p_j$, where $0 \le p_j \le 1$ and $\sum_j p_j \le r$. Its maximum is $\sum_{j \le r} \sigma_j^2$, which the leading eigenspace attains. The projection preserves $M(\sqrt{\pi_t})_t = 0$, so the approximation remains feasible for centered coefficients. The minimum residual is therefore $\sum_{j>r} \sigma_j^2$, and the centered matrix has at most $|S| - 1$ nonzero singular values. For semidefinite Σ, the proof applies on its range; null-space components of optimal heads remain unidentifiable.

For depth-dependent moments, the objective instead contains different quadratic forms,

$$\sum_t \pi_t (v + U a_t - v_t)^\top \Sigma_t (v + U a_t - v_t). \tag{40}$$

A common whitening transform generally leaves these quadratic forms different. An empirical coefficient spectrum alone therefore does not bound task-risk capacity. Our experiments measure achieved task risk and record the fitting and optimization conditions.

C Finite-sample identities and state recovery

Population identities use optimal readouts, whereas the experiments evaluate fitted heads on held-out episodes. We retain the resulting residual cross term and relate the supporting state-recovery intervention to regression onto an early head's predictions.

For any predictions p_i, fitted probe outputs q_i, and labels y_i, the exact finite-test identity is

$$\widehat{R}(p) = \widehat{R}(q) + n^{-1} \sum_i (p_i - q_i)^2 + 2n^{-1} \sum_i (q_i - y_i)(p_i - q_i). \tag{41}$$

The cross term need not vanish for ridge heads evaluated on held-out data. It is small but nonzero in the recorded evaluations, while the complete identity agrees to numerical precision across depths and models. This check verifies the risk calculation without assuming away the residual term.

Teacher-regression equivalence. For centered late states Z and early states H, with $\lambda > 0$, matrix ridge regression gives

$$M = (Z^\top Z/n + \lambda I)^{-1} Z^\top H/n. \tag{42}$$

For an early head slope w, the composition satisfies $Mw = (Z^\top Z/n + \lambda I)^{-1} Z^\top (Hw)/n$. This is exactly ridge regression onto the early head's predictions at the same penalty. The affine version follows after we restore the fitting means. State-reconstruction error and teacher-output error can select different penalties, so the equivalence applies to a fixed selected map.

We fit the map on 2,048 state pairs and select its penalty using state-reconstruction error on 512 calibration pairs. Target labels and original readout parameters enter scoring only after we save the map; the models and native heads remain frozen. The unrestricted affine map is fitted separately for each source depth, and its composition with a scalar head reduces inference to one affine readout. Shuffled controls independently permute fitting and calibration pairings.

Table 3: Original 512-episode depth-16 alignment results, averaged over three seeds per setting. The independent 4,096-episode confirmation appears in Table 4.

Setting	Native	Orthogonal	+ one scale	Affine	Shuffled
$d = 4$, fixed 4	.514715	.364974	.341633	.061745	.947634
$d = 8$, fixed 4	.609649	.510140	.418978	.041446	1.000129
$d = 4$, uniform 1–8	.092311	.085902	.086978	.061723	.951022

These controls show that general affine maps recover accuracy beyond what orthogonal rotation achieves. The maps use no new answer labels during fitting, although the early features and teacher come from supervised training.

Table 4: **Frozen recovery maps on a separate confirmation set.** Each setting includes three models and 4,096 episodes; all columns report depth-16 MSE. Uniform 1–8 has a higher mean training depth than fixed 4, so it provides a supporting recovery check rather than a matched intervention.

Setting	Native	Affine map	Separate head	Shuffled map
$d = 4$, fixed 4	.548848	.066708	.064222	.983484
$d = 8$, fixed 4	.683512	.043228	.042030	.998041
$d = 4$, uniform 1–8	.090751	.062864	.059589	.987987

D FULL MAIN-EXPERIMENT RISK TABLES

The shared-head curves in Section 5.3 summarize how each core responds to different depth preferences. The tables below give every sampled risk pair and the per-initialization results underlying that comparison.

Table 5: All eleven frozen depth weights on the fresh 4,096-episode evaluation. Each cell averages three paired seeds. The final column is balanced minus fixed weighted risk.

	Fixed depth 4		Balanced depths 1–7		Weighted-risk
p	Depth 4	Depth 16	Depth 4	Depth 16	difference
0.0	1.297482	.065144	.145790	.066791	+.001647
0.1	.291666	.076593	.093360	.067412	-.028094
0.2	.219282	.086930	.083697	.068821	-.041604
0.3	.176618	.099053	.078707	.070344	-.049470
0.4	.144701	.114092	.074965	.072244	-.053003
0.5	.118147	.133480	.071915	.074634	-.052539
0.6	.095316	.158852	.069308	.077714	-.048060
0.7	.074313	.195423	.066935	.082563	-.039023
0.8	.057076	.245325	.065273	.087373	-.025032
0.9	.042739	.331936	.063898	.095016	-.004649
1.0	.036215	.579893	.062651	.175175	+.026436

Table 6: Per-pair equal-weight results on the fresh evaluation.

Pair	Fixed shared	Balanced shared	Fixed cost	Balanced cost	Risk difference
1	.113155	.047252	.079530	.005849	-.065904
2	.113550	.073452	.058304	.008843	-.040099
3	.150734	.099119	.087568	.010967	-.051615

For the equal-weight comparison, the conditional paired 95% interval for balanced minus fixed sharing cost is $[-0.070073, -0.063365]$, and the interval for shared average MSE is $[-0.056290, -0.048985]$. At the depth-16-only endpoint, the separate-head difference has interval $[-0.002133, 0.005474]$. The measured native-risk reduction of approximately 0.186164 combines a 0.066581 reduction in sharing cost, a 0.133625 reduction in the native-to-shared gap, and a 0.014042 increase in separate risk. This achieved-risk accounting retains finite-sample residual cross terms.

The paired bootstrap averages each episode's contrast across the three seeds, then resamples the common episode indices. Intervals condition on the observed models; per-seed results separately establish the sign of each trained pair's effect.

E Reproduction details

The controlled comparison depends on matching training exposure and separating readout fitting from evaluation. We specify those choices here, together with the architecture and data splits needed to reproduce the experiments.

Matched depth schedule. Fixed-depth training supervises the native output after four loops. Balanced multi-depth training supervises one terminal depth per update, with depths $1, \ldots, 7$ represented equally up to the unavoidable remainder. Both protocols use 1,500 updates and batches of 64 fresh episodes, for 96,000 training episodes. Each paired run shares its initial weights and input–target stream, while a separate generator selects the supervised depth. We use AdamW with cosine learning-rate decay and gradient clipping, and backpropagate through the selected trajectory. The protocols match cumulative recurrent computation.

Aggregation and intervals. We average errors over episodes within each model and then over the three initialization seeds, using common episodes for every paired comparison. Paired bootstrap intervals resample complete episodes jointly across all seeds and protocols. They measure episode uncertainty conditional on the six trained models; we also report each seed's effect. The independent training sample consists of three pairs, regardless of the number of test episodes. Sampled risk curves describe the tradeoffs achieved by the specified fitting rule, rather than the full population Pareto frontier.

Table 7: Architecture and training configuration for the main matched comparison.

Component	Main matched experiment
Input dimension / context examples	4 / 8
Embedding	Affine $\mathbb{R}^6 \to \mathbb{R}^{32}$
Shared block	Pre-LayerNorm transformer encoder block
Attention	2 heads, full episode attention, no mask
Feedforward	Width 64, GELU
Dropout / positional or depth embeddings	0 / none
Output	Affine scalar head after trained LayerNorm
Optimizer	AdamW with cosine learning-rate decay
Updates / batch size	1,500 / 64
Evaluation recurrences	1 through 16

The implementation uses $r_t = \mathrm{block}(H_t) - H_t$ and $H_{t+1} = H_t + r_t$, with no engineered rotations. The matched runs share this implementation, and exact-equality checks verify their initial parameters and training input–target streams. Three additional fixed-depth models use $d = 8$, 16 context

examples, and width 48. Three original uniform-depth models use the $d = 4$ architecture and depths 1–8, giving a higher mean training depth of 4.5. These settings contain twelve independently trained models; deterministic reruns retain the same training seeds.

Table 8: Data splits, episodes per setting. Every paired run shares the same evaluation episodes.

Purpose	Count	Parameter selection
Probe/map fit	2,048	Fit coefficients
Probe/map calibration	512	Select penalty
Original analysis test	512	None; viewed in research
Transport confirmation	4,096	Maps/probes frozen
Risk-curve confirmation	4,096	Eleven heads frozen
Capacity confirmation	4,096	All rank-specific heads frozen
Calibration confirmation	4,096	All control heads frozen

We save the risk-curve heads and their checksums before generating the test set. Its 4,096 episodes differ from the transport confirmation set. Once a test set has informed a new analysis protocol, we use new episodes to confirm that analysis.

For affine ridge fitting, we center features and targets using fitting means, then solve

$$(Z_c^\top Z_c/n + \lambda I)\widehat{w} = Z_c^\top y_c/n, \qquad \widehat{b} = \overline{y} - \overline{z}^\top \widehat{w}. \tag{43}$$

For two-depth sharing, we compute means and second moments with weights $p, 1 - p$ and total loss weight one. Calibration uses the same depth-weighted objective, and the endpoints reproduce independent-head fitting. A log-spaced ridge grid is selected on calibration data; the intercept is unpenalized, and we do not refit after selection. Test labels enter scoring only after every fitting decision is fixed.

F PUBLIC-MODEL PROTOCOL AUDIT

The regression experiments fix the answer format across all readout interventions. This supporting language-model audit examines how prompt and scoring choices affect measured accuracy before any comparison across recurrent depths.

We evaluate the publicly released Huginn-0125 model (Geiping et al., 2025) using greedy generation that recomputes full prefixes without a key–value cache. The cached path produced an implementation error in our environment, which the logs distinguish from model prediction errors.

The 32 development questions ask for sums of operands drawn from 10 through 499. An answer is correct only if it completes an integer equal to the sum; truncation at eight generated tokens counts as a format failure. We compare a raw numeric prompt with a beginning-of-sequence token, the official chat template, and that template with four solved examples ($17 + 24$, $36 + 27$, $58 + 16$, $29 + 45$).

Table 9: Huginn development audit at 32 recurrences on the same 32 questions.

Numeric interface	Correct	Accuracy	Completed integer
Raw + beginning token	6/32	18.75%	16/32
Official chat template	23/32	71.875%	31/32
Official chat + four examples	28/32	87.50%	32/32

The official template adds 17 correct answers, and four examples add a net five by correcting six answers while introducing one error. All four-example responses finish within the token cap. The remaining errors are $395 + 409 \mapsto 704$ (correct 804), $207 + 83 \mapsto 300$ (290), $74 + 471 \mapsto 529$ (545), and $493 + 176 \mapsto 679$ (669).

Letter-based multiple choice retains a strong position bias. At depth 32, the official template answers 8/32 questions correctly by always selecting D. The four-example template answers 12/32 correctly,

selecting A six times and B twenty-six times, with no C or D. The available 64-recurrence records use letter protocols, so they do not provide a direct-numeric comparison between 32 and 64 recurrences.

Candidate likelihood scoring also mismatches termination tokens: numeric candidates end in a newline, while chat demonstrations end in an end-of-turn token. If we remove the saved newline contribution, four-example candidate accuracy changes from 7/32 to 11/32. This post-hoc sensitivity check remains separate from the primary score. Prompt and scoring changes affect both hidden computation and the task interface. The audit therefore establishes protocol sensitivity, rather than a frozen-state readout mechanism in the public model.

G SUPPLEMENTARY SOURCE MAP

The anonymous archive provides the main regression pipeline as source code. Its staged entry point generates synthetic episodes, trains the six matched cores, extracts frozen features, and runs the two-depth, capacity, and calibration analyses. Checkpoints, datasets, cached features, measured outputs, and paper files are maintained separately and are not included in this code-only archive.

Table 10: Source entry points in the anonymous code supplement. Experimental files are under `experiments/`; the last three entries are at the archive root.

Analysis or procedure	Source file
Fixed-depth core training	`trained_pilot.py`
Exactly balanced depth training	`budget_matched_depth.py`
Frozen features and separate heads	`readout_probe.py`
Two-depth risk curves and transfer	`readout_frontier.py`
Shared adaptation and random controls	`shared_adapter_capacity.py`
Scalar and joint-direction controls	`readout_calibration_control.py`
Paired capacity analysis	`analyze_adapter_capacity.py`
Paired calibration analysis	`analyze_calibration_control.py`
Optimization continuation	`capacity_optimizer_audit.py`
Heterogeneous risk identities	`heterogeneous_risk_checks.py`
Recurrence-frequency identities	`recurrence_spectral_checks.py`
Training objective and circle checks	`story_theory_checks.py`
Staged execution	`run_pipeline.py`
Small integration check	`smoke_test.py`

The execution instructions specify stage order, dependencies, and output paths. Runs retain the information needed to audit fitting choices and per-episode errors. The small integration check exercises training, feature extraction, readout fitting, and evaluation. Numerical checks compare regression formulas with direct solvers and verify the recurrence identities. The six-core pipeline regenerates the main regression analyses; the supplementary package contains no historical numerical records.

H CAPACITY EXPERIMENT PROTOCOL AND FULL RESULTS

What the frozen-core comparison identifies. We use the independent-head reference to separate the available prediction risk from the additional error of each constrained interface. Both cores receive the same set of evaluated depths and the same fitting supervision at every capacity. Their head parameters are fitted separately because their feature coordinates differ, but their permitted function families and selection procedures agree. This gives a counterfactual comparison on each saved trajectory: what predictions become achievable if we change the interface budget while keeping every recurrent state fixed? The independent reference locates the endpoint, the small-capacity curve measures how rapidly the constraint relaxes, and the random-direction and scalar-calibration controls test which structure produces the improvement.

One coordinate system for every depth. The capacity analysis uses the same 2,048 fitting and 512 calibration episodes as the earlier probes. We estimate one mean and pooled covariance from fitting

states across all sixteen depths. A common whitening map removes the pooled numerical null space and rescales the retained coordinates. Every depth uses this transform, so it supplies no depth-specific alignment. We append a constant feature and constrain the augmented head to $v + Ua_t$. Rank zero shares both slope and intercept; depth-specific intercept changes consume adaptation capacity. Ridge penalties apply to slopes in the common whitened coordinates, with the intercept unpenalized. This estimator differs explicitly from the earlier two-depth fits in the original normalized-state coordinates.

Fitting and controls. We evaluate $r = 0, 1, 2, 4, 8, 15$, solving rank zero by the common normal equations. Rank 15 represents all sixteen affine heads, fitted separately under one shared penalty selected from a log-spaced calibration grid. Intermediate ranks use a data-derived start and a fixed random start. An exact fixed-basis regression initializes each start, after which L-BFGS-B optimizes task risk with analytic gradients. The lower fitting objective selects the start without using test scores. The random-basis control fits coefficients and calibrates penalties for three fixed random bases at each rank, then averages their risks. Each basis remains a separate interface; the control does not ensemble predictions.

We freeze every selected interface for all twelve models before generating an independent set of 4,096 confirmation episodes. The main comparison uses six $d = 4$ fixed and balanced models. The $d = 8$ fixed and unmatched $d = 4$ uniform 1–8 models supply supporting checks. All 16 depths enter both fitting and evaluation, so the experiment measures adaptation at those depths rather than generalization to unseen depths.

Interface size and fitting cost. For width-32 models, pooled whitening retains 31 slope coordinates, giving an augmented head dimension of 32. A common head, two directions, and sixteen coefficient pairs use $32 + 32 \times 2 + 16 \times 2 = 128$ scalars, compared with $16 \times 32 = 512$ for independent heads. This storage count excludes the common feature transform and does not measure latency. Adapted heads use answer labels from the fitting and calibration splits. In contrast, the state-recovery maps in Appendix C use paired states without new answer labels. Both interventions leave the trained cores fixed.

Table 11: All-depth capacity MSE on the new 4,096-episode set. Three models per setting. Rank 15 is the common-penalty separate-head reference.

Setting	$r = 0$	$r = 1$	$r = 2$	$r = 4$	$r = 8$	$r = 15$
Fixed 4, $d = 4$	.146120	.088886	.085642	.083908	.083731	.083730
Balanced 1–7, $d = 4$	.093959	.089446	.089084	.088680	.088629	.088629
Fixed 4, $d = 8$	.111871	.060711	.056172	.054671	.054449	.054448
Uniform 1–8, $d = 4$	.088615	.084740	.084157	.083802	.083749	.083749

Define remaining cost as $\widehat{R}_r - \widehat{R}_{15}$ and the recovered fraction as $(\widehat{R}_0 - \widehat{R}_r)/(\widehat{R}_0 - \widehat{R}_{15})$. We retain fractions outside $[0, 1]$ because finite-sample risks can fall on either side of the references. In table order, mean rank-two recovery is $96.94\%, 91.45\%, 97.00\%, 91.61\%$. All six matched models exceed 90%; among the other models, one uniform-depth seed reaches 88.44%, giving eleven of twelve above 90%. Multi-depth training lowers absolute incompatibility without reducing the observed sufficient capacity. The fixed-depth mean already exceeds 90% recovery at rank one, while balanced training reaches that level at rank two.

Table 12: Random shared-basis controls, MSE on all sixteen depths. Each cell averages risks from three fixed random bases within each model, then averages three trained models.

Setting	$r = 1$	$r = 2$	$r = 4$
Fixed 4, $d = 4$	.144411	.142585	.136247
Balanced 1–7, $d = 4$	.093843	.093624	.092992
Fixed 4, $d = 8$	.109748	.106691	.104066
Uniform 1–8, $d = 4$	.088482	.088394	.087998

At rank one, balanced minus fixed error is 0.000560, with conditional paired 95% interval $[-0.001685, 0.002792]$. At rank four the difference is 0.004772, with interval $[0.002538, 0.006994]$. The three paired differences are 0.005387, 0.007204, and 0.001724.

Supporting configurations. The existing eight-dimensional fixed models recover 97.00% of mean sharing cost with two directions. The unmatched uniform-depth setting recovers 91.61% on average, although one seed reaches 88.44%. Thus 11 of 12 models exceed 90% recovery, including all six matched models. These supporting settings extend the capacity analysis to a larger input dimension and an unmatched depth schedule.

Removing the earliest depths. A diagnostic recorded before confirmation evaluates the existing heads only at depths 4–16, without refitting on this subset. Fixed $d = 4$ MSE at ranks $0, 1, 2, 15$ is $0.100136, 0.051898, 0.050621, 0.049041$; balanced MSE is $0.061335, 0.058009, 0.057971, 0.057658$. Rank-two recovery remains 96.91% and 91.50%, respectively, so the low-capacity result extends beyond the first three states.

Optimization continuation audit. Some intermediate-rank fits reached the original optimization limit, so we continued the selected rank-$1, 2, 4$ heads with penalties fixed and using only the fitting objective. All runs terminated normally, and the largest change in an individual model's confirmation MSE was below 1.4×10^{-5}. The rank-four reversal persisted in every pair. This audit assesses optimization sensitivity on an already observed test set; it adds neither independent statistical confirmation nor a global-optimality guarantee.

Metric and estimator checks. Pooled whitening leaves the per-depth second moments different, so the common-metric spectral identity cannot give an exact prediction of neural risk. We verify rank-zero and rank-fifteen solutions against the normal equations and check task-objective gradients by finite differences. The earlier original-coordinate heads and the new common-penalty references agree closely in test MSE (approximately 0.0836 for fixed and 0.0885 for balanced training). This comparison checks whether the coordinate and penalty changes alone account for the observed adaptation benefit.

The results show that a small learned interface family achieves low risk. A failed lower-rank fit would not establish that more directions are mathematically necessary. The supplementary results contain the capacity outputs, optimization audits, and deterministic common-metric checks, together with the recorded fitting choices.

I Calibration controls and independent confirmation

The capacity curves motivate a simpler explanation to test: the useful prediction direction may be shared, with only its scale and offset changing across depths. We fit two interfaces with this structure and compare them with general adaptation on new episodes.

Two restricted interface families. A depth-specific gain and offset transform a frozen prediction into $a_t \widehat{y}_t^{\mathrm{shared}} + b_t$. For shared augmented head w, the resulting head is $a_t w + b_t e_b$, where e_b selects the constant feature. These heads have an affine span of dimension at most two. A jointly fitted common slope gives $s_t v^\top z_t + c_t$, with shared v and depth-specific s_t, c_t; these augmented heads also span at most two dimensions. General heads $w + U a_t$ can change the prediction directions beyond either restricted family.

Frozen-output calibration. For each depth, let p_i be the existing common-head prediction on fitting episode i. We minimize $n^{-1} \sum_i (a p_i + b - y_i)^2 + \alpha (a - 1)^2$ without penalizing b. When $\widehat{\mathrm{Var}}(p) + \alpha > 0$, the fitting means and moments give

$$a = \frac{\widehat{\mathrm{Cov}}(p, y) + \alpha}{\widehat{\mathrm{Var}}(p) + \alpha}, \qquad b = \overline{y} - a\overline{p}. \tag{44}$$

If the denominator is zero, then $\alpha = 0$ and p is constant on the fitting set. Any a is optimal with $b = \overline{y} - a\overline{p}$; we take $a = 1$. Average calibration error across all sixteen depths selects one penalty

$\alpha \in \{0, 10^{-8}, 10^{-7}, \ldots, 10^{1}\}$. We keep the original common head and feature transform fixed, and check the unregularized solution against direct two-column least squares.

Joint single-direction fitting. We use the capacity experiment's pooled feature coordinates. Let S_t be the fitting covariance of centered features and q_t their covariance with the centered target. For fixed nonzero v and $\lambda > 0$, we minimize over gain and intercept with penalty $\lambda T^{-1} \sum_t \|s_t v\|^2$, obtaining

$$s_t = \frac{v^\top q_t}{v^\top (S_t + \lambda I) v}, \qquad c_t = \overline{y} - s_t v^\top \overline{z}_t, \tag{45}$$

and the profiled fitting objective

$$\widehat{\mathrm{Var}}(Y) - \frac{1}{T} \sum_t \frac{(v^\top q_t)^2}{v^\top (S_t + \lambda I) v}. \tag{46}$$

We normalize the direction to unit norm to fix its scale. For each penalty, analytic-gradient L-BFGS-B uses three fitting-data starts: the common slope, the leading uncentered singular direction of independent slopes, and a fixed random direction. The penalized fitting objective selects the start, and calibration selects the penalty. The selected runs terminate normally and reach numerically indistinguishable final objectives for each model. This agreement supports fitting stability while leaving the direction search nonconvex. Finite differences check the profiled gradient, and reconstructed heads reproduce the original penalized objective.

Data chronology. Both controls use the existing 2,048 fitting and 512 calibration episodes. We designed them after inspecting the original capacity test, then froze all control and comparator heads before generating a separate set of 4,096 episodes. We report both controls on this new set and on the original set, without test-based model selection. All cores and capacity heads remain unchanged.

Table 13: Control MSE on the original capacity test, a post-hoc diagnostic. The independent confirmation is Table 1.

Interface	Fixed 4	Balanced 1–7
Common head	.146120	.093959
Frozen head + gain/offset	.130169	.093530
Joint direction + gain/offset	.117791	.092608
Two general directions	.085642	.089084
Four general directions	.083908	.088680
Independent reference	.083730	.088629

Table 14: Individual paired differences, balanced minus fixed MSE. Positive values favor fixed-depth training. Old denotes the original capacity set and new denotes the independent confirmation set.

Set	Pair	Shared	Joint direction	Rank 2	Rank 4	Independent
Old	0	-.067711	-.019450	+.004424	+.005387	+.005428
Old	1	-.033539	-.021186	+.006819	+.007204	+.007297
Old	2	-.055235	-.034914	-.000915	+.001724	+.001972
New	0	-.065216	-.018265	+.002661	+.003650	+.003645
New	1	-.036162	-.024801	+.006630	+.007234	+.007290
New	2	-.051158	-.032649	+.001980	+.004854	+.005064

On the new set, the rank-four mean paired difference is 0.005246, with conditional episode-bootstrap interval $[0.002995, 0.007782]$. The fixed core with two directions also achieves lower mean risk than the balanced core's fitted independent-head reference: difference -0.003414, interval $[-0.005944, -0.001102]$. This comparison uses an attained reference under the recorded estimator. Intervals resample complete episodes jointly across six models and sixteen depths.

Same-penalty optimization bounds. For a selected capacity head, let J_r be its penalized fitting objective and J_r^* the infimum within its rank-constrained family. Let J_{ind}^* be the exact independent-head minimum at the same penalty. Since independent heads relax the constraint,

$$0 \le J_r - J_r^* \le J_r - J_{\mathrm{ind}}^*. \tag{47}$$

We evaluate this analytical bound with normal-equation solves from the fitting moments; no test data enter the calculation. The bound decreases with rank. At rank four it is below 5×10^{-4} for every fixed model and below 2×10^{-4} for every balanced model. These bounds concern penalized fitting objectives; they do not bound test MSE or population risk gaps.