
Decodability, Usage, and the Limits of Targeted Adaptation

Ji Dai Rongfeng Guo Haoyan Rong

Abstract

Decoding task-relevant information from a representation does not establish that a model uses it effectively, or that a specified training procedure can improve its use. We examine these distinctions through formation, usage, and value (FUV) diagnostics in in-context regression, physical prediction, language-model checkpoints, and offline meta-reinforcement learning. The development studies do not establish an intervention-selection advantage over performance-and-reference baselines. They do, however, identify limits of diagnostic interpretation. In physical prediction, two source arms with similarly low linear decodability respond differently to formation-oriented training. This contrast persists under a matched-update native-training control, while the specified nonlinear probes do not explain it. Increasing task-information value does not yield reliable treatment-choice headroom, and static value-weighted consumer training increases error under the tested rule. On Ant-Dir, decoded task descriptions improve separately trained reference policies in four source variants, yet none meets the pre-specified consumer-repair criterion across 42 paired treatment runs. Some treatments still yield substantial task gains, including adaptation of a random encoder; a further whitening comparison provides no reliable benefit. A simple example shows why equal decodability and current predictions cannot determine finite-budget adaptation. Together, these results distinguish diagnostic state, reference advantage, and adaptation outcomes, and support intervention-specific testing rather than a general rule for choosing which component to repair.

1 Introduction

When a model performs poorly despite encoding task-relevant information, a natural prescription is to improve the component that uses that information. Conversely, when the information is difficult to decode, improving its representation seems the appropriate intervention. This reasoning is useful for organizing hypotheses, but it conflates three questions: can an external readout extract a factor, does the current model respond appropriately to it, and can a specified training procedure turn it into better task performance? The third question concerns an optimization process, not just the model’s current function. Earlier probing and explicit-latent ICL studies already separate accessibility from useful behavior (Belinkov, 2022; Mittal et al., 2025); we test the additional step from that distinction to finite-budget intervention outcomes.

We investigate this distinction through formation, usage, and value (FUV) diagnostics. Formation denotes recoverability by a specified probe at a specified representation location. Usage denotes changes in a consumer’s output when context is exchanged, including correctness-sensitive comparisons where a target is available. Value denotes task-score differences relative to explicit context controls and task-informed reference consumers. These are operational measurements. In particular, a task-informed reference is a learned comparator, not an information-theoretic ceiling; output sensitivity is not necessarily beneficial use; and a refitted probe is not the consumer already deployed in the model.

Our initial objective was intervention selection: determine whether these diagnostics add information beyond current performance and training progress when choosing which components to adapt. The completed development studies do not establish that claim. In linear in-context learning (ICL), formation is nearly saturated and a fixed treatment leaves little room for a selector. In sinusoidal ICL, the matched reference and training-progress baselines already capture the observed treatment choices. In a physical prediction setting, different source arms exhibit different states, but the aggregate selection margin is small and sensitive to

evaluation uncertainty. A CoPhy-based setting exposes a reference-validity failure. We therefore report the completed work as an empirical study of the limits of translating diagnostics into targeted adaptation, retaining the unsuccessful selection results.

The intervention evidence nevertheless contains informative differences. In the physical prediction setting, two source arms with similarly low linear formation scores have very different *incremental* gains from staged encoder-and-reader adaptation over consumer-only training. This is not a contrast between a repairable model and an unrepairable model: both can improve, and the difference concerns the extra benefit of the staged encoder-side treatment. A later matched-update control preserves a 65.37% error reduction over native continuation in the pseudo-system-pairing arm, with little corresponding benefit in the other low-decoding arm. The specified nonlinear probes do not establish the proposed explanation that one arm merely hides more information nonlinearly. Value-weighted consumer training also worsens performance under the tested weighting rule, with global mean-one weights and a bounded, one-start full-trajectory implementation check.

Offline meta-RL provides a further test of the distinction. On Ant-Dir, decoded task information drives separately trained reference policies to higher returns in all four follow-up source variants. However, 42 paired physical treatment runs, representing 48 logical cells, establish no method-level consumer-only repair under the frozen criterion. Some treatments improve absolute return, and allowing critic gradients into a random encoder yields a substantial gain. Neither observation is equivalent to showing that a diagnosed usage deficit can reliably be repaired by freezing the representation. The planned extensions to new source seeds and an independent environment were not run because the pre-specified repair criterion was not met.

A bounded supplementary study addresses two possible explanations for these limitations. Reducing the physical predictor’s query from four frames to one makes the known-factor reference highly valuable, yet does not reveal reliable treatment-choice headroom. On Ant, an invertible change of representation coordinates, compensated to preserve the initial actor and critics, does not reliably improve fixed-budget consumer adaptation. These supplementary development comparisons retain the earlier stopping decisions.

This paper contributes (i) an operational separation of diagnostic state, reference advantage, and finite-budget repair, including a simple illustration of their logical difference; (ii) paired intervention evidence that similar linear decodability can accompany different incremental benefits from staged encoder-and-reader adaptation, including a matched-update native-training control; and (iii) negative results under pre-specified criteria separating useful reference information from reliable treatment selection and consumer repair. We preserve the chronological distinction between the initial hypothesis, development-driven follow-ups, and decisions frozen before their respective outcomes. The claims are conditional on these tasks, implementations, and small numbers of training seeds; they are not a validated general intervention-selection rule.

2 Operational diagnostics and their limits

Let h be a context, x the current query or observation, and θ an observed task factor. An encoder produces $z = e_\phi(h)$ and a consumer predicts or acts using $c_\psi(x, z)$. This factorization is explicit in our small models and policy networks. For the language-model check, the designated representation is one token at one layer; later computation can access other tokens, so it is not a complete information cut.

Formation and usage. A probe $p(z)$ is fit independently of the consumer, with the setting’s fixed training and validation split. Its held-out performance measures decodability within that probe family. Continuous factors use R^2 and task identity uses classification accuracy or balanced accuracy as specified for that experiment. Early experiments also report a reference-normalized $\tilde{F}$; nonlinear comparisons use raw R^2 to avoid mixing reader-specific denominators. These quantities are not comparable as absolute scales across settings. U_{dep} changes the context while holding the query fixed and measures output change. U_{corr} compares the change to a valid target, when one is available. Both depend on the intervention and metric. A large dependence score can reflect a wrong response, while an alignment score relative to a learned policy inherits that policy’s limitations. Appendix definitions retain the implemented setting-specific measures rather than treating them as one interchangeable scalar.

Task values and references. Write S_{ctx} for the context model’s task score, S_{noctx} for its designated context control, S_{ref} for a task-informed reference, and S_{dec} for that same reference supplied with a probe’s decoded factor. Larger scores are better; regression errors enter with a negative sign. The task-informed reference receives the current input and the factor, not the original support history. Define

$$V_{\text{realized}} = S_{\text{ctx}} - S_{\text{noctx}}, \quad V_{\text{potential}} = S_{\text{ref}} - S_{\text{noctx}}, \quad (1)$$

$$\rho = \frac{S_{\text{ctx}} - S_{\text{noctx}}}{V_{\text{potential}}}, \quad \delta = \frac{S_{\text{dec}} - S_{\text{noctx}}}{V_{\text{potential}}}. \quad (2)$$

Ratios are available only when the denominator exceeds the setting’s frozen positive threshold. Otherwise, the case remains in the report with an invalid-denominator flag. The identity $(1 - \delta) + (\delta - \rho) + \rho = 1$ is algebraic: its terms need not be nonnegative and are not a proven causal decomposition. We therefore use only the two independent ratios as features and retain the component task scores. Ant’s original ρ uses a training-task mean embedding as its context control, and its later repair study uses the unnormalized preintervention gap $J_{\text{decode}} - J_{\text{ctx}}$ directly.

Even accurate task labels do not guarantee a stronger reference. A learned reference may be underoptimized, receive an incomplete descriptor, or lack useful historical information. A negative or reversed comparison diagnoses this comparator problem before it licenses any interpretation of a normalized deficit. In particular, we do not repair unfavorable reference comparisons by clipping them into the range $[0, 1]$.

Intervention outcomes and selection headroom. The main response is the improvement in the actual context task score,

$$g_i(t) = S_{\text{ctx},i}^{\text{after}}(t) - S_{\text{ctx},i}^{\text{before}}. \quad (3)$$

An increase in context advantage over a control is a different quantity and can increase when both task scores deteriorate. When its denominator is valid, the early selection studies also use $G_i(t) = g_i(t) / (S_{\text{ref},i}^{\text{before}} - S_{\text{ctx},i}^{\text{before}})$. We keep raw errors alongside G . For a selector π , regret on the observed treatment set is $r_i(\pi) = \max_t g_i(t) - g_i(\pi(i))$, or the corresponding quantity in G units. The regret of a fixed treatment is the observed headroom available to improve upon that treatment. A small value limits the attainable average advantage of any selector on those outcomes; uncertainty in the treatment differences can further make the empirical best-treatment labels unstable. Baselines that see current performance and the same reference scores are required to separate an advantage from additional information from an advantage from the FUV summaries themselves. Selection and inference use source-trajectory groups rather than treating related checkpoints as independent sources.

Why diagnostic equality need not imply equal repair. A simple scalar example illustrates a missing ingredient without explaining our measured optimizers. Let $\mathbb{E}[x] = 0$, $\mathbb{E}[x^2] = 1$, $y = x$, $z = ax$ with $a \neq 0$, and prediction $bz = qx$ where $q = ab$ is fixed. An unregularized linear probe recovers x perfectly for every a , and all models have identical predictions and squared error. Holding the encoder fixed, one population gradient-descent step of fixed size $\eta > 0$ on $L = \frac{1}{2}\mathbb{E}[(bz - y)^2]$ gives

$$b^+ = b - \eta a(q - 1), \quad q^+ = q - \eta a^2(q - 1). \quad (4)$$

For a positive integer K , after K steps, $L_K = \frac{1}{2}(q - 1)^2(1 - \eta a^2)^{2K}$. Thus for $q \neq 1$ and a fixed positive learning rate, the same decodability and current predictions can accompany rapid improvement, slow improvement, or overshoot. For a shared upstream substitution $x \mapsto x'$, the output remains qx' at each matched initial state; output-only diagnostics built from that substitution are likewise invariant. The conclusion requires unregularized probing and a scale-independent diagnostic intervention; it does not assert that every ridge probe, noise intervention, or optimizer is invariant. Optimization geometry is a separate object from current prediction quality, consistent with longstanding work on parameterization-dependent optimization (ichi Amari, 1998). This example provides a logical counterexample to a universal prescription. We do not claim that it identifies the cause of our Adam-trained models’ results.

3 Study design and evidence chronology

We follow a sequence of development studies, rather than pool different tasks into one confirmatory test. Table 1 lists the evaluated settings and their roles. Appendix A.2 and Appendix B.1 give architecture, data, probe, sampling, and optimization details. All numerical summaries retain the completed analyses and their original thresholds.

Setting	Unit and scope	Question and outcome boundary
I1: linear ICL	15 source trajectories; 30 late and 15 early starts	Intervention prediction and subsequent continuation contrasts; no established incremental selection advantage.
I2: sinusoidal ICL	Six source trajectories; 12 starts	Prediction boundary where progress and reference scores already explain treatment choice.
I3: language tasks	Ten tasks, four checkpoints of one Pythia-1B trajectory	Task-identity readout versus answer accuracy; no model treatment or qualified value reference.
W1: controlled dynamics	Six source arms, three seeds each	54 initial responses; nonlinear probes, weighted consumer updates, repair curves and continuation repeats.
W1 supplement	Six existing starts for native controls; nine natural starts in a one-frame query condition	12 missing native-control repeats; 15 matched reader qualifications and 27 new responses.
W2: CoPhy adaptation	Three source arms, three seeds each	State variation exists, but the task reference fails qualification; no continuation.
Hopper-Param	Three informed and three blind reference policies	Initial value gate fails; bounded implementation and generalization audit.
Ant-Dir	Five source variants, three seeds each	Original state gate fails; a separately frozen four-variant usage follow-up leads to 42 physical repair runs and a failed repair gate.
Ant coordinate supplement	Three end-to-end source seeds	Six paired 40,000-update continuations; no reliable benefit from invertible whitening.

Table 1: Executed scope. Counts describe different units and must not be added as independent replications. The prospective preference-simulator setting from the broad initial plan contributes no results here.

Matched comparisons, distinct treatments. T0 means continuation of the source’s native procedure and T1 means consumer-only training with a frozen representation. However, the concrete training operations depend on the setting. In ICL, T2 adds same-function representation alignment to full-model continuation. In W1, T2 first performs source updates with a same-factor history-alignment objective and then continues the common task reader; its extra gain over T1 includes extra computation. W1 T4 instead matches T1 and changes only fixed sample weights. In Ant, T3 enables critic gradients into the encoder, and T1’ fits a fresh consumer on a frozen encoder for 200,000 updates, five times the 40,000-update continuation budget. Identically numbered labels across settings do not imply identical losses, parameter masks, or FLOPs. Paired comparisons share starts and the specified data streams; they do not force equal gradients.

ICL and W1 prediction. The original executed prediction studies use leave-source-trajectory-out evaluation, with inner source-group validation for ridge regularization. Checkpoints from one trajectory stay together. The FUV model is compared with fixed treatment, current-score, progress, matched score/reference, and other representation-feature policies. The matched baseline has the same current scores and reference information, so FUV cannot win merely by seeing an extra reference. Rank calibration is outcome-free but transductive across the fixed preintervention collection; training-fold imputation and predictor fitting remain separate. These are within-setting development evaluations, not the cross-domain prediction study proposed in the broader initial protocol.

Mechanism follow-ups and qualification. After those outcomes, a follow-up study fixed separate questions about relative versus absolute deterioration, nonlinear readout, value weighting, repair curves, and

continuation seeds. Its resampling retains source and evaluation-unit grouping. Freezing those analysis rules does not erase knowledge of earlier aggregate effects. W1 reuses historically selected sources, and the I1 early extension follows historical campaign test exposure; the reported analysis is development-only and is not called an untouched final validation. W2 and Hopper are retained as failed reference qualifications rather than repaired after the fact to enter a favorable comparison.

Ant usage and repair. The original Ant state gate required a specified high-formation/low-usage pattern and stopped. A separately frozen follow-up tests whether decoded factors improve a fixed task-informed reference and whether movement-direction diagnostics agree. All four follow-up source variants pass this narrower gate. It then evaluates T0, T1, T1', and T3 from twelve origins. Two deterministic treatment aliases reduce 48 logical cells to 42 physical jobs. A consumer alternative must have a nonnegative 95% lower bound for recovering half of the original decoded-reference return gap, a positive direction-change lower bound, and a positive direction-dominance lower bound. At least three methods must qualify to release new-source-seed and independent-environment extensions. Costs and T3 gains cannot change this decision. Full definitions and interval procedures appear in Section B.

Bounded supplementary questions. After the original study was complete, a separately specified supplement added missing W1 native-training repeats, tested one higher-information-demand W1 condition, and compared two Ant representation coordinate systems. Its 60 new scientific training jobs comprise 12 native controls, 15 reader qualifications, 27 W1 responses, and six Ant continuations; they reuse existing sources rather than add 60 independent source trajectories. Rules were fixed before each new comparison's outcomes. The W1 information-value gate released its 27 responses, but the subsequent choice-space gate did not release new-source confirmation. The Ant coordinate comparison does not reopen the original repair or second-environment gates. Appendix D gives the complete supplementary protocols and results.

Uncertainty and data boundaries. Evaluation functions, systems, tasks, and repeated reset seeds are nested within source trajectories, not new trained-model replications. Ant's follow-up resamples its three source and three reference axes separately; the original gates instead use seed- t intervals. We preserve these differing procedures and their limitations. All applicable training runs were completed and archived. Conditional extensions were not run when their pre-specified criteria were not met. We do not use sealed final-test outcomes or retrospectively substitute a new reference or decision rule.

4 In-context learning and physical prediction

4.1 Intervention-selection results

The original prediction hypothesis is not supported. On the 30 late I1 development starts, FUV's mean G -regret was 0.122856, versus 0.066558 for the matched score/reference baseline and 0.032378 for the training-fold fixed baseline. Its regret difference from the matched baseline was 0.056298 (95% interval $[-0.102538, 0.271433]$). The corresponding post-treatment MSEs were 0.037062, 0.036455, and 0.035735. On I2's 12 starts, FUV regret was 0.069817 versus 0.007067 for the matched baseline; the difference was 0.062750 ($[-0.012517, 0.200768]$). The progress-only policy selected the same treatments as the matched baseline at all 12 I2 starts. FUV improved on I2's weaker fixed baseline, but this does not establish incremental value over matched reference information or training progress. These intervals do not justify saying that FUV is universally worse either.

W1 also did not establish a predictive advantage: mean regret was 0.028086 for FUV versus 0.033176 for score/reference, with difference -0.005090 ($[-0.034669, 0.019401]$). RankMe had a smaller mean regret, 0.026909. Moreover, FUV's selected raw coordinate MSE, 2.77243×10^{-4} , was slightly above score/reference's 2.76016×10^{-4} ; lower mean normalized regret is not necessarily lower mean task error. The raw-error headroom between fixed T2 and the observed per-start oracle was only 1.0818% of fixed-T2 error. This is an empirical, noisy choice-space description, not an upper bound on all future methods. Source-level variation and the small number of independent trajectories limit interpretation.

Treatment differences limit stronger W1 conclusions. A later, separately specified noise gate required uniquely separated best and second-best treatments at at least ten of eighteen starts. Only four passed the within-start simultaneous-interval check, all favoring T2 (B3 seed 1 and the three random_B2 starts). The proposed whole-source-arm holdout was therefore not run. This prevents a claim that the learned rule transfers to an unseen source-training method. It does not imply that every treatment contrast is zero. The mechanism and T4 analyses below were specified separately in a subsequent study; their chronology and the decoded-input audit are detailed in Appendix A.3.

An early-state extension changes the choice landscape, not the conclusion. The subsequent I1 extension reconstructed step-700 states by deterministic replay and verified exact agreement with the original step-2,800 checkpoint. It added 15 starts and 45 responses on the same 15 source trajectories, not new independent seeds. A genuinely forward-frozen policy, trained on other trajectories’ late checkpoints before seeing early response labels, chose T1 at all 15 early starts with regret 0.327069 and no 0.01- G hits; the matched baseline’s regret was 0.294133. A separate, pre-specified expanded OOF analysis could use other trajectories’ early labels after they became available. Across all 45 starts, FUV regret fell to 0.040058 versus 0.041356 for score/reference, but the difference -0.001297 had interval $[-0.004001, 0.000123]$. These are distinct analyses and must not be merged into a single prospective success claim.

At the early starts, all three treatments improved the task at all 15 starts. The saved mean native gains were 0.101346 (T0), 0.049995 (T1), and 0.101571 (T2); corresponding mean changes in $\tilde{F}$ were 0.008263, zero, and 0.009529. Yet initial $\tilde{F}$ was already approximately 0.9824–0.9913. This extension therefore does not provide a severe low-formation regime or a causal mediation test. The conservative switching policies made no switches, missing the 15 beneficial early switches identified in expanded OOF; this is a limitation of that fixed rule, not grounds for tuning it after outcomes.

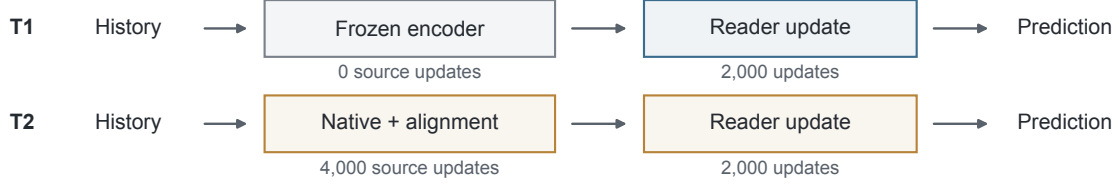
4.2 Continuation and intervention response

Relative underperformance is not absolute deterioration (N-A1). The later analysis explicitly separated $d = L_{T0} - L_{T1}$ from $h = L_{T0} - L_{\text{before}}$. For I1’s 15 final-development checkpoints, $d = 0.004907$ ($[0.002431, 0.007459]$), with 12/15 positive starts, passed the fixed joint requirement of a positive interval and at least two-thirds positive starts. Absolute deterioration did not: $h = 0.001816$ ($[-0.000590, 0.004111]$), with only 8/15 positive starts. In W1’s fixed 12-start subset (B0, B0_split, B2, B3), both intervals crossed zero: $d = 6.651 \times 10^{-6}$ ($[-2.356, 14.647] \times 10^{-6}$) and $h = 4.142 \times 10^{-6}$ ($[-5.902, 12.300] \times 10^{-6}$), despite 11/12 positive starts for each. The W1 joint criteria therefore failed. These results do not support a blanket claim that continued source training harms already formed representations.

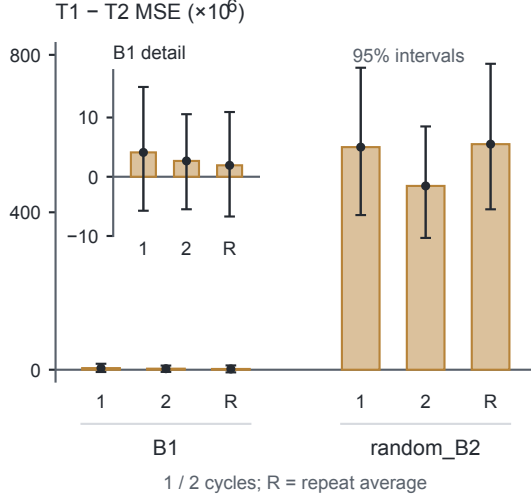
Nonlinear readouts did not explain the repairability contrast (N-A2). Table 12 shows all eighteen preintervention probe results. B1’s three linear R^2 values are 0.2214, 0.2061, and 0.2161; random_B2’s are 0.2056, 0.1911, and 0.2153. This is descriptive similarity, not a statistical equivalence claim. The fixed nonlinear-readout comparison tested whether random_B2 contained an unusually large reservoir missed by a linear probe relative to B1. MLP and k NN improved random_B2’s raw R^2 over the original ridge probe by 0.115181 and 0.062261, respectively. However, the between-arm differences in these increments were -0.001447 (97.5% interval $[-0.047858, 0.042831]$) and 0.039782 ($[-0.016117, 0.094069]$). Neither nonlinear family passed its joint explanation rule. Nonlinear gains within random_B2 do not establish the comparative explanation; failure of this rule also does not prove absent information or equivalence of the arms. All 30 state records and 90 readouts, including the pre-specified descriptive post-treatment records, were retained.

The fixed value-weighted update was a negative result (G1). Across six equally weighted arms and 18 starts, T4’s raw error exceeded T1’s by 159.561×10^{-6} (95% interval $[131.001, 188.827] \times 10^{-6}$). All six arm point estimates and ordinary 95% intervals favored T1. Under the six-arm Bonferroni intervals, five remained above zero; B0_split crossed zero. Thus this particular fixed training-gap weighting did not improve the consumer. Uniform evaluation, all 18 starts, and the original weight formula were retained. This result neither supports the weighting proposal nor establishes that task-value information is useless in every intervention.

a Two continuation schedules



b Extra benefit of staged alignment



c Fixed weighting does not help

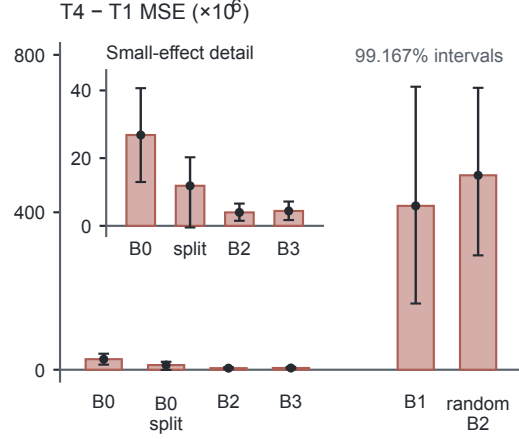


Figure 1: W1 intervention responses depend on the source arm. (a) T1 continues the reader with the history encoder frozen; T2 first performs 4,000 native source updates with added same-factor alignment, then applies the same 2,000 reader updates. The extra source computation prevents an equal-FLOP interpretation and does not isolate the alignment loss. (b) T1-minus-T2 raw coordinate MSE contrasts for B1 and random_B2 after one cycle, two cycles, and averaging three paired continuation repeats within each source. Whiskers reproduce the original 95% hierarchical bootstrap intervals on all 75 effect systems, with three independent source seeds per arm. The inset enlarges all three near-zero B1 effects. (c) T4-minus-T1 raw coordinate MSE contrasts for all six arms, with the original 99.167% six-arm Bonferroni intervals; the inset enlarges all four small-effect arms. T4 matches the reader budget and changes only its fixed training weights. Positive values in (b) favor T2; positive values in (c) favor T1. All MSE contrasts are multiplied by 10^6 ; insets use the same units as their main axes.

Absolute improvement and extra repair benefit differ (G2/G3). For B1, both T1 and T2 reduced absolute error through two standard cycles (Figure 1; full trajectories in Figures 4–5). For example, source seed 0’s full-75-system MSE decreased from 1320.19×10^{-6} to 985.939×10^{-6} and then 721.682×10^{-6} under T1, and to 982.585×10^{-6} and 720.638×10^{-6} under T2. The additional benefit of T2 over T1 was small and uncertain: 4.111×10^{-6} at one cycle ($[-5.737, 15.157] \times 10^{-6}$) and 2.671×10^{-6} at two ($[-5.489, 10.560] \times 10^{-6}$). B1 should not be described as incapable of improvement. In random_B2, the corresponding additional benefits were 565.584×10^{-6} ($[392.943, 767.269] \times 10^{-6}$) and 466.970×10^{-6} ($[334.884, 618.117] \times 10^{-6}$). The frozen standard-budget consistency checks passed for all 12 curves. Different within-arm interval decisions are not themselves a formal test of a between-arm interaction.

Two additional continuation seeds per source preserved this qualitative pattern (G3). Averaging the three paired repeats within each source gave T1-minus-T2 contrasts of 1.933×10^{-6} ($[-6.712, 10.927] \times 10^{-6}$) for B1 and 573.128×10^{-6} ($[407.573, 777.294] \times 10^{-6}$) for random_B2. For the separate T0-minus-T1 contrast, B0’s interval crossed zero (8.600×10^{-6} ; $[-6.604, 22.463] \times 10^{-6}$), whereas B2’s was positive (5.149×10^{-6} ; $[1.351, 9.785] \times 10^{-6}$). There remain only three independent source trajectories per arm: continuation repeats

do not create nine independent sources. Random_B2 is a trained negative control with a fixed pseudo-system pairing; real-factor alignment can repair that deliberately broken relation. This limits extrapolation to naturally occurring representation failures.

A matched-update native control preserves the response contrast. A later supplement completed the twelve missing native T0 repeats for B1 and random_B2, pairing them with the same source and continuation seeds used by T2. Both treatments used 4,000 source updates followed by 2,000 common-reader updates. Averaging the three continuation repeats within each independent source, random_B2’s T0 and T2 errors were respectively 897.926×10^{-6} and 310.948×10^{-6} : a 65.37% reduction. The T0-minus-T2 contrast was 586.978×10^{-6} (97.5% interval $[403.462, 827.249] \times 10^{-6}$). B1’s contrast was 1.057×10^{-6} ($[-7.540, 8.352] \times 10^{-6}$). The directly tested between-arm contrast was 585.921×10^{-6} (95% interval $[423.075, 787.009] \times 10^{-6}$). Thus the earlier separation survives a control for the number of source and reader updates. It remains a result about these objectives and source arms, including the constructed pseudo-system-pairing arm; it does not match FLOPs or identify a unique optimization mechanism (Appendix D.1).

High information value does not ensure treatment-choice headroom. We also fixed one new W1 condition with a single query frame and the same 16-step prediction horizon. All no-history, known-factor, and context readers were trained in that condition. On the diagnostic bank, the known-factor reader reduced error relative to the separately trained no-history reader by 99.923%, passing the pre-specified information-value gate. We then ran T0/T1/T2 from all nine natural-arm starts. A two-fold, system-held-out comparison selected either one fixed treatment or a separate treatment for each start on the opposite fold. Fixed T2 obtained MSE 0.00189003, versus 0.00189508 for per-start selection: relative headroom -0.267% (descriptive 95% interval $[-1.314\%, 0.908\%]$). The pre-specified choice-space gate failed. Three of the 27 corrected pairwise comparisons excluded zero, but no start had a treatment reliably better than both alternatives. No new-source extension followed. This result separates information value from a usable selection margin; it is not a new FUV predictor comparison (Appendix D.2).

4.3 Additional diagnostic settings

Task identity can precede correct task execution (I3/G4). Figure 2 summarizes the four-checkpoint trajectory; Table 15 retains task-specific intervals. Four Pythia-1B checkpoints each completed the same 5,120 diagnostic episodes across ten tasks. At step 1,000, mean task-identity recall was 98.2422%, while ten-shot exact-match accuracy was 0.1172%. At steps 48,000, 96,000, and 143,000, accuracy was 36.6797%, 44.9219%, and 47.2266%, with identity recall 99.9414%, 100%, and 99.9609%. These are the saved macro point estimates; no macro confidence intervals are claimed. Task and layer selection used final-checkpoint calibration, and the checkpoints belong to one pretraining trajectory. Task identity is a narrow label, not complete answer knowledge, and this study does not establish normalized formation, repair benefit, or treatment prediction. The lowest final task accuracy was synonym (0.3906%); English-to-German was 6.6406%, not the minimum.

W2 remained outside response analysis. W2’s nine pre-states passed the state-diversity gate, but all nine fixed reference gaps for G were negative (approximately -0.056877 to -0.003262). Therefore no G -based prediction claim or T0/T1/T2 continuation result is reported. The learned reference was not an input-nested oracle and was not replaced after this failure. This reference-qualification failure leaves W2 outside the intervention analysis under the pre-specified stopping rule.

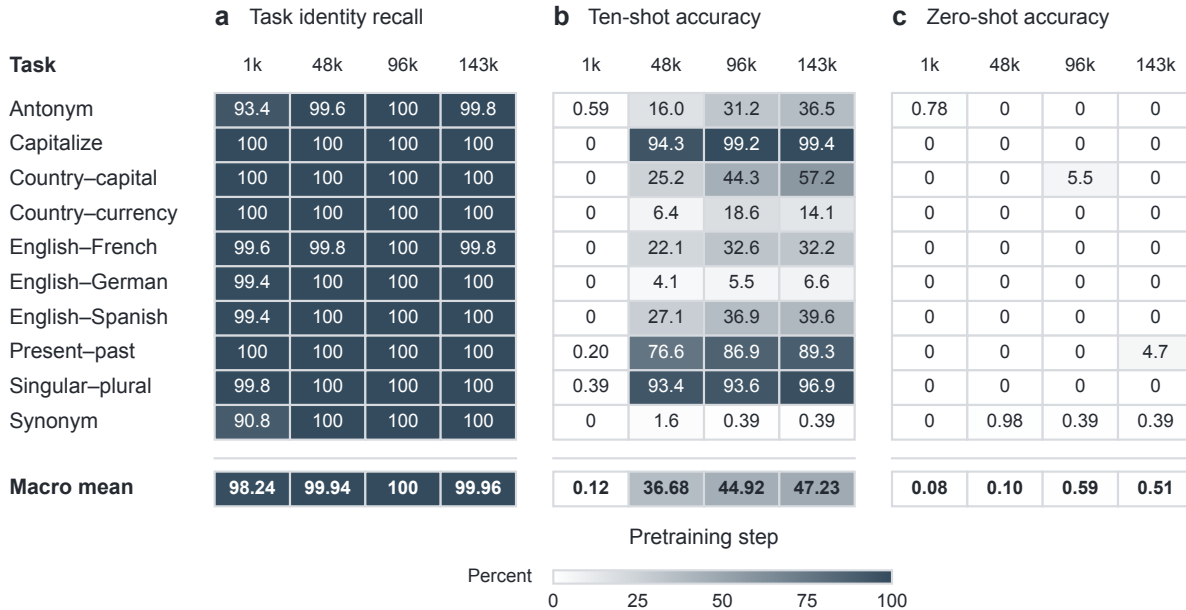


Figure 2: Task identity and task execution develop differently in one Pythia-1B trajectory. Each matrix shows all ten tasks at all four checkpoints; cells report percentages, and the separated bottom row contains the original ten-task macro means. All panels share a 0–100% color scale. Each task/checkpoint has 512 episodes. Tasks and representation location were selected by final-checkpoint calibration. The macro means have no supplied confidence intervals; the original query-cluster intervals are in Table 15. High task-identity recall does not establish executable answer knowledge.

5 Meta-RL results: usable decoded information does not ensure repair

5.1 Reference value and decoded information

The initial gates give different answers in Hopper and Ant. Table 2 reports the original reference-value results. Hopper’s task-informed policy underperforms the blind control, so no Hopper source-stage or original intervention-stage claim follows. Ant passes the value gate and enters its own frozen source assessment. Because Ant was invoked as the fallback development setting after Hopper’s failure, these two outcomes are not independent confirmation of one successful mechanism.

Environment	Task-informed return	Task-blind return	Value gate
Hopper-Param	149.76 [112.29,187.22]	213.28 [196.62,229.93]	Fail
Ant-Dir	81.33 [60.07,102.60]	−3.13 [−9.34,3.08]	Pass

Table 2: Original reference-value gate. Brackets are separate marginal 95% Student-*t* intervals over three training seeds, conditional on ten development tasks and twenty resets per task. The recorded informed-minus-blind differences are −63.52 and 84.46, respectively; the point relative differences are −42.41% and 103.85%. Raw-return ratios depend on the reward zero point.

The original Ant state criterion is not met. All five assessed state features differentiate methods under the frozen dispersion rule, but none of the four nonbaseline variants satisfies the required high-formation/low-usage pattern (Table 3). UNICORN-style exceeds the formation threshold but its action-alignment usage is not below half the end-to-end baseline. FOCAL-style, CORRO-style, and random control do not reach the formation threshold. The original state assessment therefore stops; the stronger original criterion also fails. The later analyses retain this result rather than redefining the original mismatch after seeing it.

Controlled variant	F	U_{dep}	U_{corr}	J_{ctx}	J_{decode}
FOCAL-style	0.660	0.320	0.609	57.57	84.31
CORRO-style	0.424	0.121	0.216	23.85	61.38
UNICORN-style	0.726	0.331	0.614	62.32	90.26
End-to-end	0.915	0.363	0.696	66.50	89.49
Random	0.281	0.162	0.189	11.64	63.03

Table 3: Original Ant state assessment, three source seeds per method. Entries are recorded means, not new computations; the exact source-seed values and Student- t intervals accompany the evidence table. CORRO’s intervals are particularly wide (e.g., F : $[-0.719, 1.568]$ and J_{ctx} : $[-90.13, 137.83]$); its extra pretraining budget prevents an equal-total-budget interpretation.

A1 and A2 establish a narrower decoded-reference advantage. All four methods in the separately frozen follow-up have positive decoded-reference return contrasts, positive direction changes, and positive directional contributions (Table 4). The joint A1/A2 count is four, exceeding the required three; direction coverage is complete. A1 remains positive in the recorded 98.75% sensitivity intervals. This demonstrates an advantage of the decoded input when consumed by a separately trained reference. It does not identify the source consumer as the sole bottleneck: the reference was trained with clean task inputs and differs from the source consumer. Source-to-decoded speed ratios also fail the descriptive $[0.9, 1.1]$ equivalence condition in all four methods, so the return advantage cannot be summarized as pure directional improvement at equivalent speed.

Variant	A1: $J_{\text{decode}} - J_{\text{ctx}}$	A2: Δc	A3: $J_{\text{task-mean-z}} - J_{\text{ctx}}$
FOCAL-style	26.74 [12.62, 44.44]	0.131 [0.100, 0.169]	4.90 [1.90, 7.84]
UNICORN-style	27.94 [16.90, 44.08]	0.110 [0.072, 0.153]	-0.89 [-4.04, 3.28]
End-to-end	22.99 [8.01, 37.16]	0.099 [0.067, 0.135]	6.56 [3.34, 9.90]
Random	51.39 [42.14, 62.05]	0.245 [0.192, 0.290]	14.44 [9.65, 20.33]

Table 4: Already-scored follow-up contrasts with recorded 95% percentile intervals. A1/A2 use three source and three reference seeds with 729 crossed draws; A3 uses three source seeds with 27 draws. A2’s directional-contribution intervals are also positive in every method. A3 changes the available context and is descriptive. The nine source/reference cells are not nine independent source models.

More same-task context helps some consumers but is not a repair diagnosis. The task-specific A3 average improves FOCAL-style, end-to-end, and random returns, while UNICORN-style has an interval spanning zero. Their recorded point gap-recovery fractions are 0.183, -0.032, 0.285, and 0.281, respectively. These ratios have no reported confidence intervals and do not enter the gate. Because A3 averages both more and different episodes, its gains are consistent with sensitivity to context estimation but do not uniquely attribute the deficit to noise. Neither A3 improvement nor its absence proves that consumer retraining will succeed.

5.2 Consumer adaptation and repair criteria

Return gains do not meet the consumer-repair criterion. All 42 treatment runs completed, and all twelve source-specific checks confirmed matching sample prefixes over 40,000 updates. T1 improves return relative to the original policy for FOCAL-style, UNICORN-style, and end-to-end; T1’ improves FOCAL-style and end-to-end (Table 5). These improvements are not evidence that T1 outperforms T0. More fundamentally, every one of the eight consumer alternatives has a negative lower bound for the half-gap contrast. Figure 3 separates these half-gap and absolute-return contrasts. None satisfies the joint half-gap, direction-change, and direction-dominance requirements. The repaired-method count is therefore zero, not three; both the main decision and the sensitivity flags fail. This result rejects the specified repair criterion under these budgets and tasks, not every possible form of consumer adaptation.

Variant	Treatment	Return gain	Half-gap contrast
FOCAL-style	T1	2.86 [0.018,5.20]	-10.51 [-19.76,-3.44]
	T1'	12.11 [5.73,20.02]	-1.27 [-8.42,6.73]
UNICORN-style	T1	3.61 [0.42,7.32]	-10.36 [-18.49,-3.00]
	T1'	5.63 [-4.03,13.40]	-8.34 [-22.60,4.13]
End-to-end	T1	12.41 [1.27,21.19]	0.91 [-12.21,13.45]
	T1'	21.71 [10.14,31.62]	10.21 [-3.08,24.18]
Random	T1 (=T0)	2.04 [-0.038,4.19]	-23.65 [-28.97,-17.73]
	T1'	-1.05 [-2.65,0.18]	-26.74 [-31.98,-21.48]

Table 5: Consumer-only repair alternatives. Return gain is relative to the original context policy; half-gap contrast subtracts half its fixed preintervention decoded-reference gap. Brackets are the already-recorded 95% intervals, using source-only and crossed resampling respectively. T1 has 40,000 additional updates; T1' has 200,000 fresh-consumer updates. No row passes the joint repair criterion.

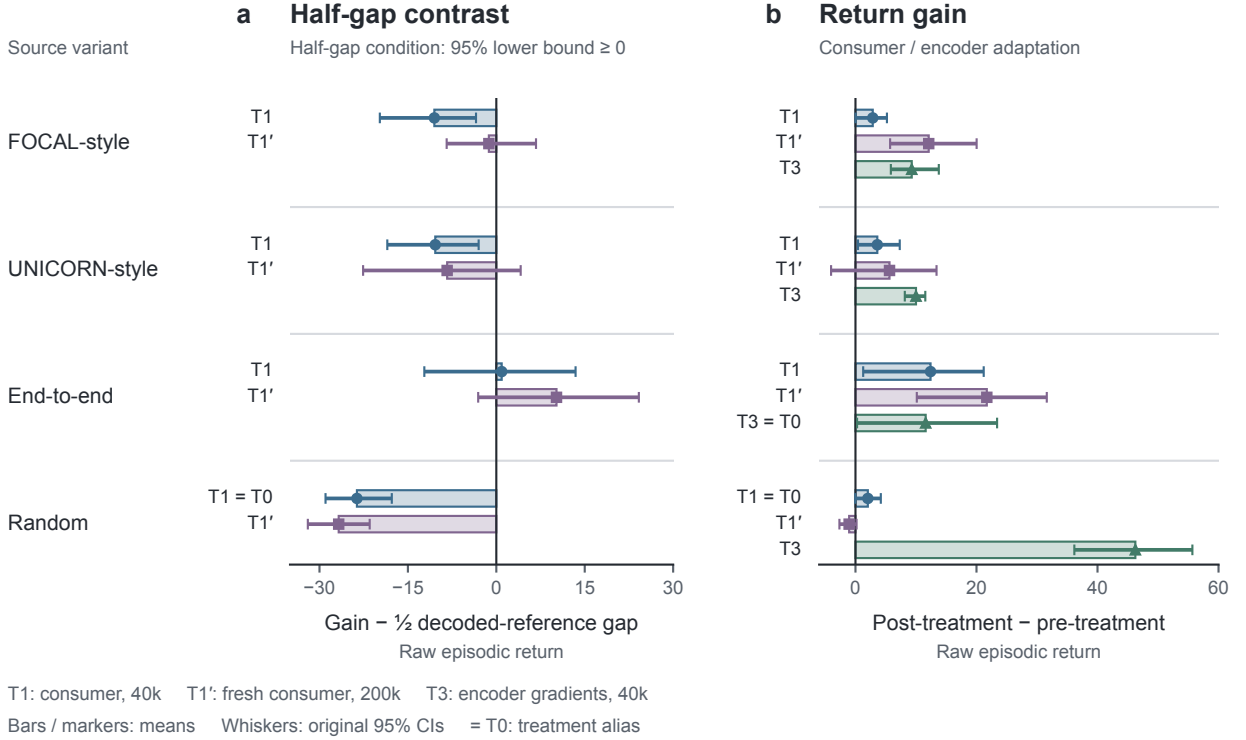


Figure 3: Ant return gains do not establish the specified consumer repair. Bars and symbols show recorded means; whiskers are the original 95% intervals. (a) Consumer-only return gain minus half the fixed preintervention decoded-reference gap. Every T1/T1' interval has a negative lower bound, so none meets the half-gap requirement. These intervals use 729 ordered crossed resamples of three source and three reference seeds. (b) Pre-post return gains, using 27 ordered source-seed resamples. Several gains are positive, including random T3, which changes the encoder. T1 and T3 have 40,000 updates; fresh-consumer T1' has 200,000. Labels identify the random T1 and end-to-end T3 aliases of T0. Fixed development tasks and resets are not resampled; neither aliases nor source-reference cells are extra seeds. The full repair criterion additionally requires direction change and direction dominance.

Random’s large T3 gain has a different interpretation. Allowing critic gradients into the random encoder produces a return gain of 46.25 [36.17,55.68] and a descriptive T3-minus-T1 contrast of 44.21 [34.83,52.02]. Its direction change is 0.394 [0.361,0.427], and its algebraic directional contribution is 41.34 [35.43,47.89]. Thus the record supports a substantial, direction-associated improvement under encoder adaptation. It does not show that the unchanged random representation has been successfully consumed or that the decomposition identifies a causal mediator. The original frozen gate excludes T3 from repair eligibility. FOCAL-style and UNICORN-style also have positive descriptive T3-minus-T1 contrasts, 6.44 [2.22,10.53] and 6.39 [3.94,8.48]; end-to-end’s T3 is its T0 alias and gives no independent intervention evidence.

No cost-advantage claim is established. A review of training retries found no cases requiring a method to be excluded from the compute comparison. The recorded T1-minus-T0 allocated GPU-time differences are -861.65 [$-894.17, -827.20$] seconds for FOCAL-style, -1131.21 [$-1185.27, -1086.17$] for UNICORN-style, and -80.36 [$-86.71, -74.43$] for end-to-end. Nevertheless, the corresponding return-difference intervals span zero: $[-1.526, 2.262]$, $[-2.714, 5.688]$, and $[-6.459, 8.756]$. All three therefore fail the frozen zero-margin noninferiority condition despite lower recorded time and fewer trainable parameters. We do not claim lossless acceleration, a favorable full-study compute tradeoff, or equal-budget efficiency for T1’. Random is excluded because T1 aliases T0. Costs do not alter the scientific repair decision.

The repair criterion does not support further extensions. The pre-specified consumer-repair criterion was not met. The two-method extension to source seeds 3 and 4 and the independent second-environment follow-up were conditional on at least three repaired methods. Because this criterion was not met, neither extension was run, and no cross-environment generalization claim follows. These stopping decisions leave the consumer-repair hypothesis unsupported under the tested conditions.

5.3 A conditioning test for consumer adaptation

A separate supplementary comparison tested whether the end-to-end consumer was hindered by the representation’s coordinate scaling. A train-only, regularized invertible whitening transform was compensated in the first consumer layers to match the initial actor and all critics. The encoder remained frozen, and both the raw-coordinate control and whitened arm reset Adam and used paired samples and random streams for 40,000 updates. Across three source seeds and ten development tasks, mean returns were 81.60 for the raw control and 72.35 for whitening. The paired difference was -9.24 (95% source-task bootstrap interval $[-26.41, 7.78]$), with one positive and two negative source-level differences. This intervention did not establish a reliable benefit; the interval also does not establish deterioration or equivalence. It tests one coordinate change under a fixed optimizer and budget, leaving other conditioning mechanisms unresolved. The original repair decision is unchanged (Appendix D.3).

6 Related work

Probing asks what information a chosen readout can recover from a representation, whereas behavioral interventions ask whether and how that information contributes to the model’s decisions. This distinction is established in the probing literature, rather than introduced here. Control tasks expose the contribution of probe memorization (Hewitt & Liang, 2019), and minimum-description-length probes account for extraction effort beyond final accuracy (Voita & Titov, 2020). Both the methodological review of Belinkov (2022) and the experiments of Ravichander et al. (2021) caution against equating decodability with task relevance. Amnesic probing instead removes information and measures the resulting behavioral change (Elazar et al., 2021). Our formation measure is consequently an operational readout score at a specified representation location, not a claim to measure all encoded information. Nonlinear probes test a particular explanation of low linear scores; failure of these probes does not prove that the information is absent. Our continuation experiments also differ from concept erasure: changing a training objective or trainable module can alter several properties simultaneously.

Representation assessment and adaptation provide a second point of comparison. LogME estimates transferability using label evidence (You et al., 2021), while RankMe uses effective rank to assess self-supervised

representations (Garrido et al., 2023). These methods motivate comparing state diagnostics against other inexpensive representation statistics, but their evidence concerns model or hyperparameter assessment, not the choice among our interventions. Fine-tuning can also distort pretrained features and underperform linear probing under distribution shift (Kumar et al., 2022). This is a relevant precedent for separating feature quality from adaptation outcomes, not an established explanation of the optimizer behavior in our experiments. More generally, optimization depends on parameter-space geometry (ichi Amari, 1998). A simple rescaling example with population gradient descent can therefore illustrate why equivalent current predictions and linear readout scores need not imply equal finite-budget improvement. Such an example establishes a limitation of a specified diagnostic summary under specified updates; it is neither a new general theory of representation learning nor evidence that the same mechanism explains Adam-trained models.

Context-based meta-reinforcement learning makes the distinction between inferred task information and its consumption explicit. PEARL combines probabilistic context inference with off-policy control (Rakelly et al., 2019); its off-policy setting should not be conflated with fully offline learning. FOCAL uses distance-metric context learning and behavior regularization (Li et al., 2021), CORRO learns task representations intended to be robust to changes in the behavior policy generating context (Yuan & Lu, 2022), and UNICORN develops an information-theoretic framework for context-based offline meta-RL (Li et al., 2024). These works motivate the source objectives studied here. We describe the corresponding experimental arms as inspired by these methods unless an exact reproduction is independently established. SAC supplies maximum-entropy actor-critic foundations (Haarnoja et al., 2018), and CQL addresses value overestimation under offline distribution shift (Kumar et al., 2020). MACAW studies optimization-based offline meta-adaptation (Mitchell et al., 2021), whereas IQL learns values and extracts a policy without querying out-of-dataset actions during training (Kostrikov et al., 2022); these are related approaches, not additional evaluated baselines. None of these citations makes a separately trained, parameter-conditioned policy an information upper bound. A learned reference remains an empirical model whose optimization and generalization must be checked.

Our closest ICL antecedent is Mittal et al. (2025), who study explicit latent bottlenecks and find that recovering task variables does not guarantee robust downstream prediction. Our ICL implementation builds on their released code. We do not present the accessibility–use distinction as new; our additional question is whether diagnostic measurements predict matched continuation outcomes, followed by explicitly bounded intervention and repair tests in other controlled settings.

In language models, implicit task inference is closely related to accounts of in-context learning. Xie et al. (2022) analyze a Bayesian interpretation in a latent-concept setting; this does not imply that arbitrary pretrained transformers perform optimal Bayesian inference. Task vectors (Hendel et al., 2023) and function vectors (Todd et al., 2024) provide stronger behavioral evidence than task-identity classification alone, through interventions that transfer or induce task behavior. Our identity probes address the narrower question of whether a specified readout distinguishes tasks at a checkpoint. They do not establish that the model can execute each task correctly. Pythia provides controlled training trajectories for studying how such measurements change over training (Biderman et al., 2023); checkpoints sharing a trajectory remain statistically dependent.

The physical-prediction settings connect to counterfactual dynamics and predictive representation learning. CoPhy studies how observations support prediction under changed physical conditions (Baradel et al., 2020). This motivates testing access to latent physical factors, but does not guarantee that the parameters exposed by a particular adaptation exhaust the information available in observation history. I-JEPA and V-JEPA learn by predicting representations of unobserved image or video content (Assran et al., 2023; Bardes et al., 2024). Variance/covariance regularization (Bardes et al., 2022) and SIGReg (Balestriero & LeCun, 2025) supply objective components used by the W1 source family. Their representation-learning results do not, by themselves, establish identification of physical parameters or correct use of those parameters by a separate consumer. These distinctions motivate checking the task, available inputs, and empirical validity of each reference before interpreting reference-normalized quantities.

Our contribution is an empirical diagnostic and intervention-tracking study within this background. We pair operational state measurements with matched continuation treatments and, where available, repair trajectories. This supports bounded comparisons between information accessibility, present behavior, and

the extra benefit of changing formation-related training rather than only the consumer. It does not establish that the diagnostic state is sufficient for selecting the best treatment, and our experiments do not establish an automatic-selection advantage over the stated baselines. Negative prediction and value-weighting results remain part of the evidence. Follow-up mechanism analyses should therefore refine the interpretation of observed changes without replacing the outcomes of the original decision rules.

7 Discussion and limitations

The evidence supports a distinction between information accessible to a diagnostic readout and information that a particular adaptation procedure can exploit. It does not support a general algorithm that selects the best intervention from FUV state. The distinction matters even where there are real gains. A positive before–after contrast establishes improvement under that treatment; it does not establish superiority to continued training, recovery of a specified reference gap, or an efficient way of obtaining the same return.

What the world-model evidence establishes. The pseudo-system-pairing and within-rollout-alignment arms have similar low linear decodability but different extra benefits from the staged encoder-alignment treatment. This result survives the reported longer-training and continuation-seed checks, and the supplementary matched-update native-training comparison gives a positive between-arm contrast. It is evidence about a response contrast, not about one arm being intrinsically impossible to repair. Nor does it establish the proposed hidden-nonlinear-information account: the joint nonlinear-probe criterion fails. The two-course training curves describe how the responses evolve, but they do not identify a unique gradient or optimizer mechanism. We did not perform the proposed gradient-angle experiments. The analytic example in Section 2 establishes that a universal implication would be unwarranted; it is not a fitted explanation of those trajectories.

The single-query-frame supplement addresses a different concern: perhaps the original task made contextual information too dispensable to reveal a useful intervention contrast. In the new condition, the known-factor comparator sharply outperforms the matched no-history reader, but the held-out treatment-choice margin remains indistinguishable from zero. Thus increasing information value was insufficient to produce reliable selection headroom in this fixed family of natural starts and treatments. Whether other tasks provide such headroom remains unresolved.

What the Ant follow-up establishes. The decoded-reference route demonstrates that a separately trained task-informed policy can benefit from the decoded task description, with improvements in both return and direction-related summaries. The subsequent adaptation study directly tests the further inference that a frozen representation can be successfully consumed after a specified retraining procedure. That inference is not established: no method meets the joint repair criterion. The end-to-end fresh-consumer point result is compatible with substantial recovery, but its half-gap interval still crosses zero. The supplementary whitening comparison likewise gives no reliable gain despite matching initial functions and resetting both optimizers. It constrains that particular conditioning intervention, not every possible optimization remedy. The random encoder’s large T3 response instead changes the representation. These distinctions prevent a favorable result from one comparison from silently standing in for a different claim.

Interpretation of unsuccessful weighting. T4 tests a particular static value-weighting rule, not the general proposition that all samples have equal value. The true-factor reference and the no-context reader define the weights; their difference need not measure the marginal usefulness of a sample for the current consumer. Even with a unit global mean and correct sample indexing, nonuniform weights change the training objective and the gradient distribution. Concentration, reference mismatch, and correlations with difficult or atypical samples are possible explanations, not established mechanisms. The negative result is therefore informative about this rule and budget without ruling out all value-aware objectives.

Scope and uncertainty. Most arm-level comparisons use three independently initialized source models. Additional continuation seeds, evaluation systems, tasks, reset seeds, source–reference combinations, and deterministic aliases do not multiply that number. The reported bootstrap and t intervals retain their

original procedures and are conditional on the specified development design. They should not be interpreted as broad population guarantees. We do not pool their numerical scales across ICL, physical prediction, and control. The diagnostic interventions differ across settings; a single-position language probe can overlook distributed information, and task identity is weaker than executable task knowledge.

Our chronology also limits the claim. ICL outcomes motivated changes in the development sequence; the Ant usage study followed an unsuccessful original state gate. Freezing each later question before its own outcomes protects that comparison from subsequent retuning, but does not make the overall research program a single prospective confirmation. The original stop decisions are retained. The independent second-environment study was not run, and neither sealed prospective-domain results nor final-test arrays were used as evidence of cross-domain validation. We make no claim of a verified prior qualitative prediction for the Ant follow-up.

Finally, the offline-RL variants are controlled implementations inspired by named methods and share a particular SAC/CQL kernel. They are not exact benchmark reproductions. Source pretraining costs differ, the reference policies are imperfect, and the evaluated Ant tasks and resets are fixed. These limitations motivate replication under independently trained sources and other environments if warranted by a new study; they are not additional experiments selected here to overturn the frozen stopping rule.

8 Conclusion

FUV diagnostics can make an intervention hypothesis precise, but they do not by themselves determine which finite-budget training change will work. The completed studies show three separations: high or similar decodability need not imply similar task performance or incremental repair benefit; decoded information can improve a reference policy without satisfying consumer-only repair criteria; and explicit value weighting can worsen uniformly evaluated task error. Some targeted interventions yield substantial gains, while the general intervention-selection and consumer-repair claims remain unsupported. Reporting both outcomes, their data roles, and the preserved stopping decisions is the central result of this study.

References

- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.
- Randall Balestriero and Yann LeCun. LeJEPa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL <https://arxiv.org/abs/2511.08544>.
- Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual Learning of Physical Dynamics. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_SkeyppEFvS.html.
- Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL <https://arxiv.org/abs/2105.04906>.
- Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video. *Transactions on Machine Learning Research*, 2024. URL <https://arxiv.org/abs/2404.08471>.
- Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL <https://aclanthology.org/2022.cl-1.7/>.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL <https://aclanthology.org/2021.tacl-1.10/>.
- Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974, 2023. URL <https://proceedings.mlr.press/v202/garrido23a.html>.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, 2018. URL <https://proceedings.mlr.press/v80/haarnoja18b.html>.
- Roe Hendel, Mor Geva, and Amir Globerson. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL <https://aclanthology.org/2023.findings-emnlp.624/>.
- John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL <https://aclanthology.org/D19-1275/>.
- Shun ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746. URL <https://direct.mit.edu/neco/article/10/2/251/6143/Natural-Gradient-Works-Efficiently-in-Learning>.
- Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=68n2s9ZWf8>.

-
- Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations*, 2022. URL <https://arxiv.org/abs/2202.10054>.
- Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html>.
- Lanqing Li, Rui Yang, and Dijun Luo. FOCAL: Efficient Fully-Offline Meta-Reinforcement Learning via Distance Metric Learning and Behavior Regularization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=8cpHIfgY4Dj>.
- Lanqing Li, Hai Zhang, Xinyu Zhang, Shatong Zhu, Yang Yu, Junqiao Zhao, and Pheng-Ann Heng. Towards an Information Theoretic Framework of Context-Based Offline Meta-Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2408. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8a30aba6514b56d02976f49797f6338a-Abstract-Conference.html.
- Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline Meta-Reinforcement Learning with Advantage Weighting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7780–7791, 2021. URL <https://proceedings.mlr.press/v139/mitchell121a.html>.
- Sarthak Mittal, Eric Elmoznino, Leo Gagnon, Sangnie Bhardwaj, Guillaume Lajoie, and Dhanya Sridhar. Does learning the right latent variables necessarily improve in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 44504–44530. PMLR, 2025. URL <https://proceedings.mlr.press/v267/mittal25a.html>.
- Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340, 2019. URL <https://proceedings.mlr.press/v97/rakelly19a.html>.
- Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL <https://aclanthology.org/2021.eacl-main.295/>.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.
- Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL <https://aclanthology.org/2020.emnlp-main.14/>.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022. URL <https://arxiv.org/abs/2111.02080>.
- Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, 2021. URL <https://proceedings.mlr.press/v139/you21b.html>.
- Haoqi Yuan and Zongqing Lu. Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25747–25759, 2022. URL <https://proceedings.mlr.press/v162/yuan22a.html>.

A Detailed experimental methods

A.1 Development settings and measurement contract

Our non-meta-learning experiments separate a representation’s factor readout, its influence on the task consumer, and its task value. They comprise two synthetic in-context regression settings (I1 and I2), a controlled dynamics setting (W1), a CoPhy-derived setting (W2), and a checkpoint study of Pythia-1B (I3). These experiments first evaluated a treatment-selection hypothesis; later, separately specified development analyses examined particular mechanisms. The latter do not retroactively validate the original prediction hypothesis. We report all applicable completed analyses, including negative findings and W2’s failed reference qualification. No prospective R1/R2/A1 or final-test results are used in this section.

Formation, use, and value are different measurements. Let z denote the measured context or history code, x the unchanged current query, and θ the known task or physical factors. In I1/I2 and W1, formation is operationalized by a held-out linear factor probe:

$$\tilde{F} = R^2(\theta, \hat{\theta}(z)) / R^2(\theta, \hat{\theta}(\theta)). \quad (5)$$

The denominator uses the same ridge-probe pipeline with the true factors as its representation; it is not assumed to equal one. Neither R^2 nor the ratio is clipped. This measures the chosen probe’s recoverability of specified factors, not all information or knowledge in z . Probes are fitted on a training bank, regularized using a separate calibration bank, and scored on D_{diag} . The consumer and the measured representation are unchanged by this diagnostic fit.

For regression, U_{dep} is the mean squared displacement between the consumer’s original and context-swapped predictions at the same query. I1/I2 use native output units; W1 uses the fixed training-state normalization. It measures sensitivity, which can be large without useful dependence. Where legal counterfactual targets exist, U_{corr} is the pooled projection of the prediction displacement onto the target displacement, $\sum \langle \Delta \hat{y}, \Delta y \rangle / \sum \|\Delta y\|^2$; it is not clipped to $[0, 1]$. In W1 the counterfactual changes only drag in an exact simulator replay from the unchanged query endpoint and future controls. I3 instead uses task-identity classification and candidate-distribution KL; these are deliberately not renamed $\tilde{F}$ and regression U_{corr} .

Task scores S have larger-is-better sign: negative MSE for regression and greedy exact-match accuracy for I3. A separately trained, fixed known-factor consumer defines S_{ref} ; it is an empirical reference, not an analytic oracle or a guaranteed upper bound. S_{dec} uses that same consumer with probe-decoded factors, without replacing or refitting it. Define

$$V_{\text{realized}} = S_{\text{ctx}} - S_{\text{noctx}}, \quad V_{\text{potential}} = S_{\text{ref}} - S_{\text{noctx}}, \quad (6)$$

$$\rho = V_{\text{realized}} / V_{\text{potential}}, \quad \delta = (S_{\text{dec}} - S_{\text{noctx}}) / V_{\text{potential}}. \quad (7)$$

These dimensionless ratios can lie outside $[0, 1]$; invalid denominators are recorded as missing. The denominator must strictly exceed both zero and a fixed calibration threshold: half the width of the learned-reference score’s 95% percentile interval from 10,000 calibration-unit bootstrap draws, after averaging the three reference seeds. The response target uses a different gap,

$$G_T = \frac{S_{\text{ctx}, T}^{\text{after}} - S_{\text{ctx}}^{\text{before}}}{S_{\text{ref}}^{\text{before}} - S_{\text{ctx}}^{\text{before}}}. \quad (8)$$

Its reference and validity threshold remain anchored before treatment. Diagnostic features use D_{diag} ; paired treatment outcomes use a separate D_{effect} . A valid state ratio does not ensure a valid G_T .

In-context regression. I1 learns linear functions of eight input features and an intercept (nine factors), with label-noise standard deviation 0.1. I2 learns a one-dimensional sinusoidal family with three coefficients and three fixed frequencies, without label noise. The frequency generator is the pinned upstream rule 5rand(1,3) at seed 1, with no additional 2π factor. Both explicit models have separate four-layer context and prediction Transformers: width 256, four heads, feed-forward width 512, a 256-dimensional context code,

and zero dropout. Sources use Adam, learning rate 10^{-4} , zero weight decay, batch 128, and 7,000 updates in FP32. Context length is sampled from 16–128 and query count equals context count. Factors and examples are generated online; these are not 7,000 fixed datasets.

I1 has five arms (EXP, EXP_AUX, EXP_ALIGN, EXP_VIB0, EXP_VIB), each with source seeds 0/1/2. I2 has EXP_SAME and EXP_NEAR with the same three seeds; the latter uses $x_q = x_c + 0.1\epsilon$ during source training. Checkpoints at 2,800 and 7,000 updates yield 30 I1 starts and 12 I2 starts, but only 15 and six independent source trajectories. At each start, T0 continues the native objective; T1 freezes the entire context encoder and optimizes only the consumer; T2 continues the full model with an added same-function, two-context representation-alignment term. Each runs 1,400 updates with paired data streams and the source Adam moments (filtered to the consumer for T1). T2 is not the same intervention as W1’s staged treatment.

Controlled dynamics. W1 reuses six source arms—B0, B0_split, B1, B2, B3, and random_B2—at 20,000 updates, with three source seeds each (18 starts). The measured code contains only a donor history: 24 states and 23 observed actions. B0 exposes its native 128-dimensional monolithic code; split models expose their 64-dimensional persistent code. A common, newly trained reader receives a separate current query window and known future controls. The main condition uses four query frames and predicts the horizon-16 endpoint. Formation probes use the native code before padding; the reader standardizes the code using training histories and pads split codes to width 128 at its input.

The dataset has 1,000 training systems and 200 development systems, split by fixed system identity into 50 calibration, 75 diagnostic, and 75 effect systems. Each measurement bank has 32 windows per system. The base reader uses 10,000 AdamW updates, effective batch 512, learning rate 3×10^{-4} , weight decay 0.05, a 500-update warmup followed by cosine decay, and norm clipping at 1. T1 adds 2,000 reader updates with the source fixed. T0 first performs 4,000 native source updates, then the same reader continuation; T2 adds calibrated same-factor history alignment during those source updates before continuing the reader. Native batches contain 48 pairs and alignment batches 96 examples. Thus T2-versus-T1 includes extra source computation and is not an equal-FLOP comparison. Main prediction analysis uses the 54 T0/T1/T2 responses; optional T3 is excluded from this comparison.

Treatment prediction and uncertainty. The executed development analyses leave out a whole source trajectory: both late checkpoints for I1/I2 and one arm-by-seed start for W1. Regularization is selected by inner leave-one-source-group validation from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, with an unpenalized intercept. One ridge model predicts each treatment’s G . The matched score/reference baseline and FUV share S_{ctx} , S_{noctx} , S_{ref} , ρ , its missing indicator, progress, context fraction, and distribution-shift indicator. FUV additionally receives $\tilde{F}$, U_{dep} , δ , and their missing indicators; U_{corr} is a reported diagnostic, not an input to this predictor. Rank calibration uses the fixed collection of pre-intervention features, including held-out features, without outcomes: this is a transductive feature protocol, not fully inductive preprocessing. Imputation and model fitting remain inside the training folds.

We report regret relative to the observed best treatment, a hit within 0.01 G , and task error in its native scale. The fixed-treatment baseline chooses its treatment using training-fold mean G , not the held-out outcome. Existing 95% intervals for predictor contrasts resample whole source trajectories 10,000 times, preserving checkpoint and method pairing. They are descriptive, do not refit the predictor on each draw, and are not multiplicity-adjusted. The observed per-start best treatment is a noisy hindsight comparator, not a guaranteed attainable prediction improvement. Later night analyses use the separately specified hierarchical units and corrections detailed in Appendix A.2.

A.2 Non-meta-learning implementation and evidence details

This appendix records the executed development design. Preparation files retain some historical “not yet generated” status strings; execution and completion receipts, not those stale preparation labels, establish completion. The accompanying evidence map pins the original reports, numerical aggregates, protocols, and implementation definitions. It does not claim that historical uncommitted source-training revisions were retroactively certified.

ICL generators and source objectives. The ICL implementation derives from the code of Mittal et al. (2025), `explicit_implicit_icl`, commit `caf74532cd2010cab73372873e0263bc4077f037`. Linear factors, including the intercept, are independent standard-normal draws; inputs are standard normal. In I2, $y(x) = \sum_{j=1}^3 a_j \sin(\omega_j x)$, with $a_j \sim \mathcal{U}[-1, 1]$ and the fixed frequencies described in the main text. Source data and model randomness have separate checkpointed streams. EXP minimizes query MSE. EXP_AUX adds unit-weight factor MSE through a separate linear head, with gradients into the encoder. EXP_ALIGN adds same-function consistency. EXP_VIB0 and EXP_VIB have identical stochastic 256-dimensional diagonal-Gaussian bottlenecks, with KL coefficients zero and 10^{-3} . Their mean head starts as identity, log-variance head as zero weights and bias -2 , and log variance is clipped to $[-10, 5]$. Both sample during training, including VIB0; diagnostics probe the posterior mean, while consumer scores average eight fixed posterior-noise draws paired across comparisons. The detached online source probe is not the formal formation measurement.

ICL alignment compares L2-normalized codes (normalization epsilon 10^{-8}) from the first $\lfloor n/2 \rfloor$ and remaining context points of the same function, keeping the full query batch. Both sides backpropagate; posterior means are used for VIB alignment. Twenty fixed training batches set $\lambda = 0.1 \bar{L}_{\text{native}} / \bar{L}_{\text{align}}$, with weight one if mean alignment is below 10^{-12} . Calibration restores randomness and does not use outcomes. The native loss includes applicable source auxiliary, KL, and pre-existing alignment terms; T2 adds its newly calibrated term. In EXP_ALIGN, T0 already retains source alignment and T2 adds to it. This is not presented as a wholly different objective in that arm.

Each setting has 8,192 probe-training functions and separate banks of 4,096 calibration, 4,096 diagnostic, and 4,096 effect functions. The main evaluation has 72 context and 72 query points. Pre-fixed sensitivity conditions vary context size (16, 32, 72, 128), query distribution (same, wide, far), and context noise (I1: 0.1/0.3/1.0; I2: 0/0.1/0.3). Conditions share factors and paired random draws; they are not extra independent training starts. The “wide” condition scales the base query by three; “far” uses $0.1x$ plus a length-three independent direction. Three KnownLatent references train for 7,000 updates with the same prediction architecture and Adam recipe, then are averaged equally at their fixed last step. The true function is used for simulation and counterfactual labels, never substituted for the learned reference task score. No-context evaluation passes a genuinely empty context through the unchanged native encoder, retaining its learned token/prior. The measurement pipeline consistently disables the attention fast path after an empty-context numerical incompatibility; it does not replace missing context by a favorable fabricated representation.

Ridge probes select regularization by five-fold calibration-only CV grouped by function identity (seed 660023). The grid is the seven values in the main text. Features and targets are standardized using only fit-fold statistics, with scale floor 10^{-8} and an unpenalized intercept; final fitting uses only the separate probe-training bank. R^2 averages factor dimensions and is neither clipped nor computed against a pooled all-bank target mean. RankMe uses the first 2,048 raw diagnostic codes and uncentered singular-value entropy (epsilon 10^{-7}); LogME uses the frozen Gaussian-prior evidence implementation and factor labels. Neither substitutes for task utility. All ICL responses use seed 1000 + source seed, shared data across T0/T1/T2 and the two original checkpoints, with audited batch hashes.

W1 source reuse and consumers. W1’s four-dimensional state consists of the simulator state used by the bound source, with two-dimensional controls. A 24-frame source history is encoded through an observation width of 128 and a history Transformer of width 192, depth four, eight heads, and dropout 0.1. Split variants use 64-dimensional transient and persistent components; B0 has one 128-dimensional context. Source metadata records 20,000 training updates, learning rate 3×10^{-4} , weight decay 0.05, and warmup 500. The native objective combines future latent prediction at horizons 1/4/16 and SIGReg (Balestriero & LeCun, 2025) (weight 0.02). B0_split changes the representation split without the persistent-pair loss. B1/B2/B3 add the persistent VICReg-style term (Bardes et al., 2022), with B1 using non-overlapping windows from the same rollout and B2/B3 same-system pairing. B3 additionally uses a cross-context prediction term. The bound original coefficients and optimizer state are retained during native continuation. Random_B2 uses B2 architecture with a hash-bound, fixed pseudo-system relation; it is trained, not an untrained random network.

The common reader uses a GRU with input width seven (state four, preceding control two, and a known-control flag) and hidden width 128, a condition-to-64 GELU projection, and a 289–256–256–4 GELU MLP. Known future controls are padded to 32 steps with validity masks, and horizon is an additional scalar. State and control normalization use only the original training corpus. Source-specific code normalization uses 8,192 fixed training histories; the no-context slot is zeroed after this normalization. The known-factor reader uses the same query/control architecture but receives only scalar drag in its condition slot, never donor history.

W1’s 32 measurement rows per system are eight query rollouts times anchors 23, 25, 28, and 31. A separate donor rollout and a fixed hash determine the history. Training-system histories used to fit probes are not unseen source systems. Ridge CV groups all 32 rows of a calibration system, while R^2 scores all 2,400 diagnostic rows. W1 reference training uses three seeds and the same 10,000-update reader recipe; decoded drag is injected without clipping. Diagnostic scores and U_{dep} use squared residuals divided by fixed training-state scales, whereas the reported task-error contrasts use *raw coordinate MSE*: mean over coordinates, then windows within each system, then equally over systems. A displayed factor of 10^6 is only a unit scaling. The two error scales are not interchangeable.

W1 T2 aligns raw history codes with MSE, unlike ICL’s L2-normalized alignment. The paired histories have the same drag and unchanged complete query/control inputs. Twenty fixed batches calibrate the added term to 10% of native loss. T0/T2 restore source optimizer moments where available; absence means a shared, explicitly recorded reset from the bound source weights, not an exact resume of the earlier trajectory. Reader continuation restores the common completed base reader’s moments under a new, shared 2,000-update schedule. Microbatch 128 accumulates an effective batch of 512. A supplementary joint T3 existed but did not enter the main treatment-selection analysis. Historical source selection had already exposed development data; eight legacy source records lacked an original test-read flag and carry explicit reuse evidence. These limitations are retained rather than rewritten as pristine held-out training.

Development prediction chronology and early extension. The original broad campaign protocol proposed cross-setting/domain prediction. The I1/I2/W1 analyses reported here are the executed within-setting, leave-source-trajectory development evaluations, and cannot be labeled a successful cross-domain validation. I1 has 15 outer folds with 14 training sources; I2 has six with five training sources; W1 has 18 with 17 training sources. Source groups never split across inner and outer folds. Prediction ties use T0, then T1, then T2. A seeded uniform treatment permutation supplies the random policy. Missing continuous features use training-fold medians (or 0.5 when a column is wholly missing), with explicit missing indicators.

The step-700 I1 extension was specified after the earlier development results and after the campaign’s historical final-test exposure; it is not a fresh confirmatory holdout. This manuscript section uses only its development artifacts, not the historical final-test results. Exact replay to step 2,800 checks the reconstructed trajectory but does not turn step 700 into an originally saved checkpoint. The forward-frozen and expanded-OOF views have separate artifacts and must remain separate. The latter has 45 starts, 135 responses, and the same 15 source groups; rank calibration includes the fixed 45 pre-treatment feature rows. The conservative rule’s large leave-group-out radii and missed switches are reported as a fixed-rule limitation, without post-outcome retuning.

N-A1: relative versus absolute changes. This development reanalysis was specified after historical aggregate effects were already known. It selected all five I1 arms at their final development checkpoint (15 starts) and the fixed W1 arms B0/B0_split/B2/B3 (12 starts). For each setting, the rule required both a positive 95% interval and positive direction in at least two-thirds of starts, separately for d and h . Ten thousand bootstrap draws (seed 920101) resampled source trajectories within each fixed arm and used paired, shared evaluation-unit draws. W1 units are 75 systems; I1 units are 4,096 functions. A positive d means T0 finishes worse than T1; only h addresses worsening from the starting state. Their scientific meanings cannot be collapsed.

N-A2: a fixed nonlinear explanation test. The three readouts were the original ridge, a 128-unit ReLU MLP with scalar output, and uniform Euclidean 15-nearest-neighbor regression. The MLP used AdamW (learning rate 10^{-3} , weight decay 10^{-4}), batch 512, maximum 200 epochs, patience 20, minimum improve-

ment 10^{-5} , and seed 660031. Training-only feature/target normalization was fixed; calibration selected the best epoch using equal-system standardized MSE. k NN used training-only feature normalization and fixed bank-order tie breaks. Readouts were trained on probe_train, selected on calibration, and scored on the same diagnostic systems. The primary comparison is raw R^2 ; common original-ridge reference normalization and same-family reference normalization are retained as distinct secondary summaries. An unchanged source under T1 reuses its pre-treatment readout after exact source/bank/probe identity checks.

There were 18 pre-states plus B1/random_B2 post-T1/post-T2 states for all three seeds: 30 state records and 90 readouts. The explanation required that at least one nonlinear family have positive lower bounds for both its random_B2 nonlinear increment and that increment’s difference from B1. The 97.5% intervals account for testing two readout families. Bootstrap seed 920102 used 10,000 draws, source resampling within arms and shared diagnostic-system draws, recomputing the R^2 denominator inside each original analysis draw. No additional bootstrap is performed for this paper. Post-treatment readouts remain descriptive and do not rescue the failed pre-treatment explanation rule.

G1: exact value-weighted treatment and bounded implementation control. T4 freezes the source and changes only reader training weights. Before continuation, its fixed bank covers $1000 \times 8 \times 25 = 200,000$ training queries, keyed by system storage row, rollout, and anchor 23–47. With the starting reader’s normalized no-context error e_i^0 and the equal mean of three per-example reference errors e_i^{ref} , define

$$g_i = e_i^0 - e_i^{\text{ref}}, \quad u_i = \text{clip}(g_i/\bar{g}, 0, 5), \quad w_i = u_i/\bar{u}. \quad (9)$$

The means are over the entire training bank. There is no batch-wise renormalization, no online weight update, and no second clipping; final weights may exceed five. A nonpositive global denominator fails the job. The objective is the example-and-coordinate mean of weighted normalized squared error. T4 otherwise uses the same 2,000 reader updates and paired sampling as T1, and evaluation remains uniformly weighted. The overall contrast equally weights the six arms; 10,000 bootstrap draws (seed 920103) resample source trajectories within arm and shared effect systems. Arm-wise intervals additionally use six-arm Bonferroni correction.

A later, separately accepted technical control used the real B0_s0 start and continuation seed 1000 on single-threaded CPU. T1 and all-one T4 each completed all 2,000 updates: 8,000 paired microbatch losses were exactly equal, with equal final runtime model/optimizer/scheduler/RNG states. The runtime also verified 1,024,000 actual query-to-bank lookups; independent local review checked 24 saved real lookup samples and the receipts. This is a bounded implementation check on one start, not 18 additional GPU replications, not equality to the historical GPU endpoint, and not a new scientific benefit test. It leaves the negative G1 conclusion unchanged.

G2/G3: budget trajectories and continuation repeats. G2 compares B1 and random_B2 at three source seeds each, with T1 and T2 (12 complete curves). A second standard cycle continues each phase’s own optimizer moments and random stream, restarts its original phase schedule, and reuses the initially calibrated alignment weight. T1 has 2,000 reader updates per cycle; T2 has 4,000 source then 2,000 reader updates. The plotted budget fraction divides by these respective budgets, not by common wall time or FLOPs. Fixed deciles from zero to two give 21 T1 points; the T2 source/reader boundaries add two points, for 23 per T2 curve and 264 measurement points in total. All curve points use the same hash-selected 25 effect systems. Full-75-system endpoint values at zero/one/two cycles are reported separately and are not spliced into those curves. The pre-specified endpoint comparisons use 10,000 paired draws (seed 920104). Other times and formation trajectories are descriptive; pointwise intervals are not simultaneous curve bands. No best time point is selected after seeing outcomes.

G3 adds continuation repeat IDs 1 and 2 to the original ID 0 for four arms: B0/B2 compare T0–T1 and B1/random_B2 compare T1–T2. Seeds are $1000 + \text{source seed} + 10000 \times \text{repeat ID}$. Each arm completed 12 new jobs, for 48 total. The estimator first averages paired repeats within source, then sources. Its 10,000 draws (seed 920105) resample source trajectories, paired repeats within sampled source occurrences, and a shared effect-system draw. Replayed standard G2 trajectories are not counted as new seeds. With only three independent sources per arm, these intervals have limited source-level support despite the extra continuation repeats.

I3/G4: checkpoint identity study rather than intervention evidence. The public Pythia-1B architecture is GPT-NeoX: 16 layers, hidden width 2,048, eight attention heads, feed-forward width 8,192, and maximum sequence length 2,048. The ten tasks derive from `function_vectors` commit `fb9eac7b6dc707ea1475a717379916007fe448d5`. A globally shared NFC-normalized input-string hash allocates roles, keeping related translation queries together across tasks. Demonstrations come only from the training input pool, excluding the current query. Each probe-training/calibration/ diagnostic bank has 1,024/256/512 episodes per task, respectively. Their 10,240/2,560/5,120 episodes contain 847/873/783 distinct global query IDs; episodes with different demonstrations are not independent queries.

The fixed measurement site is the 2,048-dimensional output of complete block 9 at the last query-prefix token, before answer tokens. The ten-class linear probe uses standardized features and L2 logistic regression (LBFGS, maximum 2,000 iterations, tolerance 10^{-4} , $C = 1/\alpha$ with the seven-value grid). Five-fold calibration CV groups repeated query IDs; the final fit uses only `probe_train`. Both task selection and layer selection use the final checkpoint’s calibration data. Nine tasks passed the 0.05 ten-minus-zero-shot selection threshold; synonym was explicitly added for required antonym/synonym pair coverage despite its below-threshold calibration gain. It is retained in all reports. The four executed checkpoints are steps 1,000, 48,000, 96,000, and 143,000, not the earlier candidate-checkpoint list in preliminary metadata. Earlier checkpoint probes are fitted by the fixed recipe; the existing final-checkpoint result is reused.

Correctness is unconstrained greedy exact match after NFC and stripping surrounding whitespace, preserving case, stopping at newline/EOS with at most 32 generated tokens. Ten-shot and zero-shot scores share queries. The initial final-checkpoint I3 study also reports U_{dep} as KL over 16 fixed candidate answers, based on summed conditional token log probabilities with no length normalization. Its U_{corr} is donor-task answer accuracy (and change from the original-context donor score), only for same-input task pairs with disjoint valid answers. Capitalize, present-past, and singular-plural have no such legal counterpart here; missing U_{corr} is not zero. G4 does not add new KL, U_{corr} , reference training, or model interventions. Its saved within-task 95% intervals use 10,000 query cluster draws (seed 920106) shared across checkpoints; repeated demonstrations stay grouped. The ten-task macro values have no supplied macro interval. No matched reference was qualified for these I3 summaries, so no normalized $\tilde{F}$, ρ , δ , or treatment-selection evidence is claimed.

W2: stop on reference qualification, not a replacement experiment. W2 reuses FullCross, Structure, and Random sources with three seeds each, a 64-dimensional persistent code, three support histories, a three-frame query, and a horizon-12 trajectory target. Formation concerns recorded mass/friction/restitution, not every physical property. The existing consumer uses the source’s 128-dimensional combined code; this is distinct from the 64-dimensional probed persistent code. There are 14,000 training systems and 4,000 development systems split into 1,000 calibration, 1,500 diagnostic, and 1,500 effect systems. Both active objects remain in the same probe fold. The known-factor reference uses the original `Head(3,4,3,12)` with query pose, public detection/type, and the three raw factors, and no history. Three references train for 11,000 updates (100 epochs), AdamW learning rate 3×10^{-4} , weight decay 10^{-4} , effective batch 128 (tail 48), and no scheduler, with fixed last-update selection.

The state-diversity rule required the maximum arm-mean range to exceed twice the pooled within-arm standard deviation. Formation, dependence, and δ passed; ρ did not. Nevertheless, all nine response gaps $S_{\text{ref}} - S_{\text{ctx}}$ were negative, so their G targets were unavailable. The state-ratio and response denominators must not be conflated; negative δ and $\rho > 1$ are also retained where their own denominator is valid. W2 has no defined U_{corr} in this measurement contract. The trained reference does not receive the learned consumer’s complete latent input and is not an input-nested upper bound. Architectural/optimization or scale mismatch are possible explanations, not identified causes. There was no reference replacement, threshold change, or downstream continuation to turn this stopped branch into a positive result.

A.3 W1 task, uncertainty gate, and input audit

W1 dynamics. W1 is a deterministic planar damped-force system with state (p_x, p_y, v_x, v_y) , unit mass, and dynamics $\dot{p} = v$, $\dot{v} = u - \gamma v$. Its scalar drag is uniform on $[0.5, 4]$ and fixed within a system. The simulator uses the exact constant-force solution at time step 0.05, producing 64 states per rollout and eight

rollouts per system. Initial positions are uniform in $[-1, 1]^2$; speed is uniform on $[0, 2]$ with uniform planar direction. Piecewise-constant force segments last 3–8 steps. Each segment is zero with probability 0.3; otherwise magnitude is uniform on $[0, 4]$ with uniform direction. The dataset seed is 20260818. These definitions describe the bound simulator; the historical source-training provenance limitations stated above still apply.

The W1 treatment-noise gate precedes the later mechanism study. After aggregate W1 outcomes were known, an additional cost gate was fixed before reading per-system error details. It required at least ten of eighteen starts to separate their observed best and second-best treatments. The existing analysis used 10,000 paired draws of the 75 effect systems (seed 920031), keeping 32 windows per system together and sharing indices across treatments and starts. It applied 98.333% intervals to each of the three fixed treatment pairs within a start, giving an approximate 95% Bonferroni family interval within that start, not across all eighteen starts. Only four starts passed (six with ordinary 95% intervals), and all four favored T2: B3 seed 1 and all three random_B2 seeds. The gate failed, so the proposed whole-source-arm holdout was not run. This is insufficient separation of unique best treatments, not absence of every treatment effect. It also does not establish reliable state-dependent differences in which treatment is optimal. T4 was not run under that earlier decision; the later, separately specified mechanism study authorized the reported T4 experiment.

Matched task conditions and bounded decoded-input audit. The diagnostic and effect banks use the same four-frame query and horizon-16 endpoint task on different, disjoint systems. Normalized and raw errors derive from the same predictions; separate data roles need not be removed to reconcile the units. A later saved-model GPU replay examined all $3 \times 2400 = 7200$ random_B2 decoded-drag predictions, reproduced 51 archived checks exactly, and found no prediction outside $[0.5, 4]$. Clipping to that support changed neither the decoded-reference error nor δ . This excludes the proposed out-of-range injection explanation for these records, without establishing a unique mechanism for negative δ . The initial CPU replay mismatch was retained as a numerical execution limitation; the original GPU measurement and frozen threshold were not changed.

B Offline meta-RL: separating available information from successful use

We use offline meta-RL to distinguish information encoded in a context representation from information used by its policy. A context consists of transitions from one task, and a policy receives an observation and an aggregated context embedding. The reference intervention instead supplies a task description to a separately trained policy. This comparison tests what a specified consumer can achieve with a supplied input; it does not make the reference an optimal policy or establish that the original consumer can be repaired.

Experimental sequence and data roles. The initial sequence contains a reference-value gate followed, only if that gate passes, by a representation-state assessment. Hopper-Param was the first development setting. Its failed value gate permitted Ant-Dir as a fallback development setting, not as independent confirmation. The original Ant state assessment also stopped. The later Ant usage study was a new question motivated by these observed development results and was frozen separately; it does not retrospectively replace either original decision. All training uses designated training tasks. Evaluation and diagnostics use designated development tasks, and no sealed or final-test arrays enter these analyses. The source and reference training seeds are 0, 1, 2. Development evaluation uses ten tasks and twenty fixed reset seeds per task, giving 200 rollouts per model. The horizon is 200 steps. Tasks and resets are averaged before training-seed uncertainty is reported.

Reference-value gate. For each environment, we train three task-informed and three task-blind policies, each for 200,000 updates with batch size 256. Both arms have the same input dimensions; the blind arm replaces the task features with zeros. In Hopper-Param, the recorded physical descriptor has 41 coordinates: body masses, inertias, damping, and friction. Ant-Dir uses the two coordinates $(\cos \theta, \sin \theta)$ of the author’s recorded direction. Observation and task-feature normalization are fit on training data only. The gate requires a positive return difference, a point relative difference of at least 20%, and the lower marginal 95% interval of the informed policy to exceed the upper marginal interval of the blind policy. The relative

difference divides by the absolute informed-policy return; a near-zero denominator ($\leq 10^{-6}$) is ineligible. These are point and marginal-interval criteria, not a confidence guarantee that the relative improvement is at least 20%. The marginal intervals are Student- t intervals over three training seeds, with two degrees of freedom and critical value 4.302652729911275.

Controlled implementations. We implement a common offline SAC plus two-proposal CQL(H) kernel inspired by Haarnoja et al. (2018); Kumar et al. (2020), rather than claiming an exact reproduction of any author’s complete pipeline. The actor is a tanh-Gaussian policy; two critics and their targets use two 256-unit ReLU hidden layers. Learning rates are 3×10^{-4} , discount 0.99, target-update coefficient 0.005, fixed entropy coefficient 0.2, and conservative-loss coefficient 2. At temperature one, the conservative partition estimate uses ten uniform actions and ten current-policy actions with their proposal-density corrections and equal proposal weighting. Its log-sum-exp is normalized by the twenty candidates. Time-limit truncations are not treated as terminal events. Policy evaluation uses the deterministic action. Ant has 27 observation and eight action coordinates; Hopper has twelve and three. We do not interpret results from this controlled kernel as benchmark reproductions of the original algorithms.

The original Ant state assessment trains five controlled source variants for three seeds each. An episode contains 200 transitions, each represented as the concatenation of normalized observation, action, reward, and normalized next observation. For Ant, the transition encoder is $63 \rightarrow 128 \rightarrow 128 \rightarrow 41$. Its deterministic embedding is aggregated into a 41-dimensional latent, giving actor input dimension 68 and critic input dimension 76. These 41 latent coordinates are not physical task coordinates. Each update draws sixteen training-task slots, one complete context episode per slot, and sixteen RL transitions per slot, for 256 RL transitions. The methods share the paired sampling design and the RL-network initialization within each seed.

The labels FOCAL-style, UNICORN-style, and CORRO-style indicate controlled constructions informed by Li et al. (2021; 2024); Yuan & Lu (2022). They are not assertions of exact author-code reproduction. FOCAL-style uses mean pooling and a task-identity metric loss on context embeddings. With mean squared latent distance d_{ij} and $\epsilon = 10^{-3}$, same-task pairs contribute $\sqrt{d_{ij} + \epsilon}$ and different-task pairs contribute $1/(d_{ij} + 100\epsilon)$; the two sums are separately normalized by their pair counts plus ϵ . Its metric weight is one. UNICORN-style combines the same metric at weight 1.5 with a learned Gaussian dynamics-and-reward reconstruction loss at weight one. Its decoder has two 128-unit hidden layers and learned log-variance bounds. For these two variants, auxiliary losses update the encoder, while critic gradients stop at the representation. End-to-end control uses mean pooling and critic gradients into the encoder, without the auxiliary losses. Random control freezes a randomly initialized encoder and uses mean pooling. Actor and target branches do not send representation gradients in any variant.

CORRO-style first trains a conditional VAE, then a deterministic, transition-level L2-normalized encoder using InfoNCE with generated negatives, and finally freezes that encoder while training an attention aggregator and the RL consumer. The negatives retain the query state/action and generate reward/next-observation values from independent Gaussian prior latents. Positive pairs are distinct transitions from the same task. This distinct-row rule and the shared controlled RL backend are explicit implementation choices. The local primitive audit references CORRO commit 210ad407a0a7807b7789ab426d8b5be6c72268a7; it is not an end-to-end equivalence claim. Each CORRO seed uses 1,406,300 VAE updates at batch size 128, 200,000 contrastive encoder updates at batch size 64, and 200,000 downstream RL updates. The VAE count corresponds to the recorded nominal 100-epoch budget for 1,800,000 training transitions, with task-uniform IID row sampling rather than epoch shuffling. The contrastive temperature is 0.1 with sixteen negatives per query. Consequently, equal downstream update counts do not imply equal total training cost across the five source variants.

Original formation, usage, and value measurements. We fit a task-grouped ridge probe from embeddings to the recorded physical task descriptor using only training tasks. The fixed design contains twenty training tasks, 32 contexts per training task, ten development tasks, and twenty contexts per development task. Five folds each hold out four training tasks; feature/target scaling is fit within each training fold. The ridge coefficient is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ by task-balanced out-of-fold raw macro-

R^2 , with ties resolved toward stronger regularization. Formation F is the development macro- R^2 of the refit probe using raw, unclipped physical predictions. Negative scores are retained.

Usage measurements compare actions at identical observations while exchanging task contexts. The design includes all 45 unordered development-task pairs and fifty observation queries per pair. U_{dep} is the mean RMS action difference normalized by action half-range; U_{corr} is its cosine alignment with the action difference of the task-informed reference, whose differences are averaged across the three reference seeds before taking the cosine. Results average states within pairs and then pairs equally. Near-zero reference differences are handled with explicit coverage, not replaced by positive evidence. A separate return diagnostic is

$$\rho = \frac{J_{\text{ctx}} - J_{\text{mean-z}}}{J_{\text{ref}} - J_{\text{mean-z}}}, \quad (10)$$

when the denominator exceeds 10^{-6} . Here $J_{\text{mean-z}}$ uses an equal-task mean embedding from the training tasks, whereas J_{ref} averages the three task-informed references. This control differs from the later task-specific A3 average.

For J_{decode} , the probe prediction supplies the task input of each frozen task-informed reference policy. The true environment task is unchanged. Ant predictions are clipped to the training-coordinate box and projected to a unit direction within that box, with a fixed near-zero fallback. These bounded injection inputs do not replace the raw predictions used for F . We retain three source and three reference axes rather than treating the nine combinations as nine independently trained source models.

The original state gate requires detectable state differentiation (at least one feature’s between-method range exceeds twice the pooled within-method seed standard deviation) and at least one nonbaseline method with $F \geq 0.7$ and $U_{\text{corr}} < 0.5U_{\text{corr}}^{\text{2e}}$. A stronger point criterion additionally requires a reference-gap fraction of at least 0.3. Reference validity is checked separately. These thresholds and the accompanying seed- t intervals are distinct objects; intervals are not clipped to the feasible range of a statistic.

Separately frozen usage follow-up: A1–A3. The new follow-up covers FOCAL-style, UNICORN-style, end-to-end, and random control, using the existing twelve source checkpoints and six reference checkpoints. CORRO remains in the original five-method result table and is outside this four-method follow-up. A1 asks whether $J_{\text{decode}} - J_{\text{ctx}}$ has a strictly positive 95% interval. A2 checks whether the same comparison improves movement direction and has a positive directional contribution to the return difference. It replays the old conditions and requires exact agreement with the archived returns and trace hashes before adding motion diagnostics. With planar velocity v_t , target unit direction θ , mean speed $s = H^{-1} \sum_t \|v_t\|$, and direction cosine $c = \sum_t v_t^\top \theta / \sum_t \|v_t\|$, the Ant return is Hsc , where $H = 200$. We use the symmetric identity

$$\Delta J = D + S, \quad (11)$$

$$D = \frac{H}{2}(s_b + s_a)(c_b - c_a), \quad S = \frac{H}{2}(c_b + c_a)(s_b - s_a). \quad (12)$$

This is an algebraic decomposition, not causal mediation. Stationary paths retain their rows but produce unavailable direction diagnostics; complete direction coverage is required for the gate. At least three methods must jointly pass A1, positive direction-change lower bound, and positive D lower bound. Speed equivalence is descriptive and cannot change this gate.

A3 supplies each original consumer with the mean of 128 separately encoded context episodes from the same development task. These episodes exclude the original twenty contexts and all episodes containing the original action-query states. The fixed embedding is reused over the twenty reset seeds. This changes both context volume and context composition; it is neither an oracle representation nor a uniquely identifying test of representation noise. A3 is descriptive and does not change the gate. The execution comprises twelve source jobs and six shared-reference jobs, with 10,800 replay rollouts plus 2,400 task-mean rollouts (13,200 total); no training occurs in this phase.

Consumer-repair treatments. After the joint A1/A2 gate passes, the second follow-up phase starts from all twelve complete 200,000-update source states. T0 continues the original training procedure for

40,000 updates. T1 freezes all representation modules and continues the existing RL consumer for 40,000 updates. T3 permits critic gradients into the encoder and any trainable aggregator for 40,000 updates while retaining the treatment’s existing auxiliary schedule. T1’ freezes the source representation, independently initializes the actor and both critics with matching fresh targets and empty consumer Adam states, and trains this fresh consumer for 200,000 updates. It preserves the original sampling and private update-noise starting states. Continuations restore the complete original optimizer, RNG, sampler, and normalization state. The recorded collector verifies identical full 40,000-update sampling prefixes within each of twelve origins, including the first 40,000 updates of T1’. This is data pairing, not a claim that different treatment gradients are identical.

There are 48 logical cells but 42 physical models: random T1 is T0, and end-to-end T3 is T0, at all three seeds. Each physical model has 200 development context rollouts, for 8,400 new rollouts. Aliases add neither independent seeds nor independent training costs. T1’ has five times the continuation budget and is excluded from the equal-40,000-update cost comparison. Original before-treatment formation and action-usage diagnostics are retained; we do not present this follow-up as a new post-treatment F/U study.

For treatment $t \in \{T1, T1'\}$, define the return gain $g_t = J_t - J_{\text{ctx}}$ and the fixed preintervention gap $d = J_{\text{decode}} - J_{\text{ctx}}$. A consumer alternative passes only if the 95% lower bounds satisfy

$$\text{LB}(g_t - \tfrac{1}{2}d) \geq 0, \quad \text{LB}(\Delta c_t) > 0, \quad \text{LB}(D_t - \tfrac{1}{2}g_t) > 0, \quad (13)$$

with a positive fixed gap and complete direction coverage. A method may pass through either alternative, but the choice is made across all three source seeds, never separately per seed. At least three methods must pass to release either conditional extension. T3, cost, and sensitivity analyses cannot alter this decision.

Uncertainty and cost. A1/A2 and phase-two half-gap contrasts use all $3^3 \times 3^3 = 729$ ordered crossed bootstrap draws, resampling the three source and three reference axes separately with indices shared across methods and treatments. Source-only motion, A3, and cost contrasts use all $3^3 = 27$ ordered draws. Quantiles use linear interpolation. Tasks and resets are fixed, not resampled, and the repeated use of one reference does not create additional independent units. A1 additionally reports 98.75% sensitivity intervals; the eight phase-two consumer alternatives additionally report 99.375% intervals. These sensitivity intervals do not replace the frozen main gate.

The cost comparison is T1 versus T0 for the three nonrandom methods. It requires complete allocated GPU time, strictly fewer trainable parameters in every source, a strictly negative upper bound for the time difference, and a nonnegative lower bound for the return difference (zero-margin noninferiority). The recorded training-time scope includes startup/loading, training, logging/checkpoints, saving, and final verification, and excludes later evaluation and backup. Missing full-attempt accounting is NA. The separate retry audit determines whether such an accounting caveat applies; absence of a relevant retry does not itself establish a cost advantage.

B.1 Meta-RL implementation and reproducibility details

Table 6: Executed Meta-RL settings. These reproduce the controlled study’s configuration, not an assertion of equivalence with an original author’s complete pipeline. Exact settings and file hashes are retained in the evidence handoff.

Component	Setting
Reference/source training seeds	0, 1, 2; three trained seeds per condition
Development evaluation	Ten tasks; reset seeds 50000–50019; horizon 200; 200 roll-outs/model
Ant task partition	Twenty training tasks; ten development tasks; sealed/final-test excluded
Ant training transitions	1,800,000 (recorded source configuration)
Hopper dimensions	Observation 12, action 3, physical descriptor 41
Ant dimensions	Observation 27, action 8, physical direction 2; source latent 41
Reference training	200,000 updates, batch size 256, both informed/blind arms
Source training	200,000 downstream updates per method/seed
Reference input	Normalized physical descriptor; blind receives zero task slots
Actor and twin critics	Two ReLU hidden layers of width 256; target critics initialized from critics
Actor distribution	Tanh Gaussian; log standard deviation clipped to $[-20, 2]$; deterministic evaluation
Actor/critic learning rates	3×10^{-4} , Adam
Discount / target coefficient	$\gamma = 0.99$, $\tau = 0.005$
Entropy / conservative coefficients	Fixed 0.2 / 2.0
CQL proposals	Ten uniform and ten current-policy actions; proposal corrections; temperature one
Ant transition encoder	63–128–128–41; two hidden layers; deterministic
Encoder output / aggregator	Tanh and mean pooling except CORRO-style: per-transition L2 and attention
Source sampling	Sixteen task slots; one 200-step episode and sixteen RL rows/slot
Representation/decoder optimizers	Adam; learning rates 3×10^{-4} ; betas (0.9, 0.999); epsilon 10^{-8} ; weight decay zero
FOCAL-style metric	Weight 1; pair epsilon 10^{-3}
UNICORN-style objectives	Metric weight 1.5; reconstruction weight 1; decoder 128–128 hidden layers
UNICORN-style log-variance bounds	Initial minimum -10 , maximum 0.5; learned bounds; regularizer 0.01
Representation seeds	Encoder $10000 + s$, aggregator $20000 + s$, decoder $30000 + s$, noise $40000 + s$; consumer and sampler seed s
CORRO-style VAE	Latent 64; hidden width 128; KL weight 0.2; observation/reward MSE weights 1
CORRO-style VAE budget	1,406,300 updates, batch 128; nominal 100 epochs; task-uniform IID rows
CORRO-style contrastive budget	200,000 updates, batch 64; sixteen negatives/query; temperature 0.1
CORRO-style positive pairs	Distinct rows from the same task; frozen VAE during contrastive training
CORRO-style downstream	Frozen pretrained encoder; attention updated through critic; 200,000 RL updates

Component	Setting
Fixed diagnostic contexts	32/train task; 20/development task; complete 200-transition episodes
Action comparison design	45 unordered task pairs; 25 query states from each task per pair
Ridge cross-validation	Five task-grouped folds of four training tasks; equal task then context weighting
Ridge candidates	$10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100$; unpenalized intercept; tie selects larger coefficient
Ridge scaling / score	Training-fold weighted population scales; pooled task-equal out-of-fold raw macro- R^2
Probe numerical tolerances	Scale epsilon 10^{-8} ; SST epsilon 10^{-16} ; CV tie tolerance 10^{-12}
Design seeds	Contexts271828; action queries314159; probe folds161803
Usage convention	Action half-ranges one; near-zero epsilon 10^{-6} ; equal pair then state
Ant decoded injection	Train-box clipping plus feasible unit-circle projection; raw R^2 remains unclipped
A3 task-specific mean	128 disjoint episodes/task; selection seed2026100403; no original contexts/query episodes
A2 motion	$dt = 0.05$; horizon200; stationary path threshold 10^{-9} m
T0/T1/T3 continuation	40,000 updates from the same complete 200,000-update origin
T1' fresh consumer	200,000 updates; new actor/critics/targets and empty consumer Adam; frozen original representation
Continuation checkpoints / logs	Checkpoints every10,000; logs every1,000 updates
Original-stage uncertainty	Marginal or source-seed Student- t intervals; $n = 3$, df2
Follow-up crossed uncertainty	729 ordered draws of separate three-source/three-reference axes
Follow-up source-only uncertainty	27 ordered three-source draws; shared indices; linear percentile interpolation
Follow-up interval levels	95% main; A1 98.75% sensitivity; repair alternatives99.375% sensitivity
Task/reset resampling	None; all intervals conditional on the fixed development design
Ant rollout runtime	Recorded Gym0.18.0, mujoco-py2.1.2.14, MuJoCo native version210; explicit seed and one raw reset

Source variant	Auxiliary E update	Critic E update	Critic A update	Extra pretraining
FOCAL-style	Metric	No	No	None
UNICORN-style	Metric+reconstruction	No	No	None
CORRO-style	No	No	Yes	VAE+InfoNCE
End-to-end	No	Yes	No	None
Random	No	No	No	None

Table 7: Executed source gradient ownership. E is the transition encoder and A the context aggregator. All but CORRO-style use parameter-free mean pooling. Actor and target branches do not update either representation module. UNICORN-style additionally updates its decoder through reconstruction. These are source masks; T1 and T3 explicitly change them.

B.2 Hopper reference audit

Task parameters are the archived recorded values. Ant directions and Hopper physical parameters were recorded with limited decimal precision; current consistency checks do not recover unrecorded acquisition precision. A post-result Hopper diagnostic verified the recorded parameter/environment/policy-input mapping for thirty tasks and 1,230 coordinates, with zero maximum discrepancy, without changing the original failed gate. A separate descriptive diagnostic found informed-minus-blind training-task return differences of 25.92, 29.72, and 24.95 across the three seeds, while the original development differences remained negative. The train and development task sets differ, and behavior episodes do not share rollout resets with model evaluation. These observations neither provide a new significance test nor prove that additional training would fix the failure. The original Hopper implementation runs a fixed 200-step horizon without health termination, so low return alone does not establish falling.

The archived 29,000 training behavior episodes have mean return 201.26. The six learned policies have training-task means from 231.96 to 273.06, exceeding this descriptive behavior average. Their behavior and evaluation resets are unpaired; this is not a paired policy-improvement test. Together with the preserved negative development comparison, the record is compatible with a task-generalization problem, rather than evidence that both policies simply failed to learn anything. No new training or revised gate follows from this diagnostic.

C Evidence accounting and reproducibility

Completed units. The night study completed N-A1 for 27 starts, N-A2 for 30 state records (90 readouts), G1 for 18 weighted-reader updates, G2 for twelve two-course trajectories, G3 for 48 additional continuation jobs, and G4 for four checkpoints. These counts have different units and are not an effective sample size. The original Meta-RL development sequence completed six Hopper references, six Ant references, fifteen Ant sources, and the three CORRO-style pretraining sequences. The separately specified Ant usage phase completed eighteen physical evaluation jobs. Its repair phase completed 42 physical training jobs and 8,400 new development rollouts, representing 48 logical treatment cells.

Frozen outcomes and operational verification. The Ant repair collection requires all 42 per-job back-ups, complete evaluations, and all twelve same-origin 40,000-update sampler-prefix checks before its single complete scoring pass. That pass records zero methods meeting consumer repair. The new-source-seed and independent second-environment branches are consequently not applicable. The original Ant state-gate stop remains part of the record. We distinguish replay and packaging checks from new scientific evidence: copying byte-identical historical traces into a compact delivery repairs the artifact layout without changing results or increasing the sample count.

The evidence bundle retains exact numerical tables, analysis definitions, source and reference identities, training and evaluation receipts, logical alias maps, seed-grouping rules, and content hashes. Large checkpoint and trajectory files remain in persistent storage with explicit locations and SHA-256 inventories; they are not represented as small local table files. Manuscript figure generation reads already-scored summaries and does not refit predictors, recompute confidence intervals, or reevaluate policies. Reproduction of a plotted value and reproduction of model training are separate tasks.

Availability and external validity. The private author handoff contains the source maps, full numerical exports, and reproducibility scripts used for this manuscript. No public data or code release is claimed. The paper describes the completed configurations and their methodological boundaries so the controlled experiments can be assessed independently of the private operational system. Access to original datasets and large model artifacts may be needed for full training reproduction. A public anonymized release and journal submission are separate author decisions, not actions performed as part of this manuscript preparation.

Source arm	T0	T1	T2	T0–T2 [97.5% CI]
B1	929.195	930.072	928.138	1.057 [−7.540, 8.352]
random_B2	897.926	884.076	310.948	586.978 [403.462, 827.249]

Table 8: Matched-source-update W1 control study. All entries are raw coordinate MSE or differences multiplied by 10^6 ; lower MSE is better and positive T0–T2 favors alignment. Each arm has three source trajectories and three paired continuation repeats per source. T1 is a descriptive comparator.

D Bounded supplementary tests

These studies were specified after the original manuscript and experiments were complete. They address alternative explanations and task-design limitations raised by the earlier outcomes. Their individual rules were frozen before their own outcomes, and all positive and negative results are retained. They are post-hoc additions to the research program, not an untouched prospective validation. No sealed final-test data were read, no prior stopping rule was changed, and no additional setting was selected after these results.

D.1 W1: matching source updates isolates an additional alignment benefit

The earlier W1 results leave a specific alternative explanation: T2 updates the source representation as well as the consumer, whereas T1 updates only the consumer. Its advantage could therefore reflect additional source training. After those results were known, we specified a separate supplementary study comparing T2 with native continuation T0 at the same source and consumer update counts. This study retains the original results and stopping decision (Appendix A.3); it is a post-hoc control study, not an independent prospective test of intervention selection.

For B1 and random_B2, each with three source seeds, we added T0 continuation repeats 1 and 2: twelve new runs. Together with existing repeat 0 and the existing T2 repeats, this gives eighteen paired T0–T2 comparisons. Existing T1 repeats are included descriptively. The three repeats are averaged within each source; there are three independent source trajectories per arm, not nine.

Both T0 and T2 perform 4,000 source updates followed by 2,000 consumer updates on the original four-frame, horizon-16 task. T0 retains the native objective; T2 adds the calibrated alignment objective. Pairing preserves the initial checkpoint, native training batches, consumer data order, optimizer initialization, and learning-rate schedule. Continuation seed is $1000 + s + 10000r$ for source seed s and repeat r . Consumer continuation restores the AdamW moments from the 10,000-update base reader; source optimization follows the same inherited restore/reset policy in both treatments. We retain learning rate 3×10^{-4} , weight decay 0.05, effective batch size 512, gradient clipping at 1, and the original warmup and cosine schedules. The native batches contain 48 pairs; T2 additionally uses alignment batches of 96 examples. Thus *update counts are matched, but FLOPs are not*: T2 still performs additional alignment computation. The original random_B2 pairing map is unchanged.

Raw coordinate MSE is evaluated on the same 75 development effect systems, keeping the 32 windows of each system together. The frozen analysis uses 10,000 bootstrap draws (seed 202610071), resampling source trajectories within each arm and sharing paired system draws across treatments. The two arm-specific intervals are 97.5% intervals, with Bonferroni correction for two comparisons; the between-arm contrast has a 95% interval.

T2 reduces random_B2 error by 65.37% relative to T0, versus a descriptive 0.114% reduction for B1 (Table 8). The between-arm difference in T0–T2 is 585.921×10^{-6} , with 95% interval $[423.075, 787.009] \times 10^{-6}$. The large random_B2 advantage therefore survives a control for the number of source updates. B1 is not unrepairable: the earlier consumer continuation already improves it; the present comparison finds little additional benefit from alignment. Nor does the control establish why the responses differ. In particular, it does not overturn the failed comparative nonlinear-readout criterion in Section 4.1. Similar initial linear readout scores coexist with different additional alignment benefits, but the proposed hidden-nonlinear-information explanation remains unsupported. Finally, random_B2 is a trained, deliberately mismatched pairing con-

Source arm	Before continuation	T0	T1	T2
B0_split	168.045	170.019	164.389	168.478
B1	5554.401	5539.910	5530.497	5462.822
B2	39.052	43.876	37.750	38.790

Table 9: One-frame W1 responses: descriptive raw coordinate MSE multiplied by 10^6 , averaged over three source seeds per arm. Each treatment has one continuation repeat per source. Apparent arm-wise differences in the best mean do not by themselves establish reliable treatment selection.

trol, not an untrained random network or a naturally arising source family. The result is a bounded repair contrast, not evidence of general FUV predictive superiority.

D.2 W1: increasing information demand does not guarantee treatment-selection space

A second, separately specified supplement asks whether the original task made context too easy to replace. We reduce the query from four state frames to one, retaining the horizon-16 target, future controls, and the independent same-system support history. The history contains 24 states and 23 actions. We use three existing source arms, B0_split, B1, and B2, with three seeds each; no source is deliberately damaged for this experiment. This is a new task condition with new consumer fits, so its absolute errors are not pooled with the four-frame study.

Qualifying information value before continuation. We first train fifteen readers: three no-history Null readers, three known-drag Oracle readers, and nine context readers with frozen sources. Null and Oracle have the same architecture, initialization seed, data order, and 10,000-update budget. Null receives the training-mean factor, hence zero after normalization; Oracle receives the true drag, normalized once using training statistics. Neither receives support history. Both are learned references: “Oracle” denotes its known-factor input, not a certified information upper bound. Reader training uses AdamW with learning rate 3×10^{-4} , weight decay 0.05, effective batch size 512, clipping at 1, 500 warmup updates and cosine decay. We use the final 10,000-update checkpoint, without checkpoint selection.

Qualification uses only the 75 diagnostic systems, D_{diag} . The specified gate requires at least a 10% Null-to-Oracle relative MSE reduction and a positive lower 95% bound on the paired raw difference. Its 10,000-draw bootstrap (seed 202610072) resamples reader seeds and shared systems. Null MSE is 0.006906969 and Oracle MSE is 0.000005293, a 99.923% reduction. The raw difference is 0.006901676, with 95% interval [0.005326983, 0.008709697]. This passes the information-value gate without consulting D_{effect} to select the condition. It establishes that the task benefits strongly from knowing drag; it does not establish that different starts need different treatments.

Responses and the predeclared selection-space gate. We then complete all $9 \times 3 = 27$ T0/T1/T2 responses, one continuation per cell. Each starts from its new one-frame consumer, including its optimizer moments. T0 and T2 again use 4,000 source updates followed by 2,000 consumer updates; T1 freezes the source and uses 2,000 consumer updates. Training data, native samples, and consumer samples are paired across applicable treatments. Scores use the disjoint 75 development effect systems, with 32 windows per system; no final-test data are used. The analysis is performed once after all 27 outputs, pairing checks, and backups are complete.

The gate, fixed before response outcomes, measures selection space without using FUV features. A deterministic hash of system IDs partitions the 75 systems into folds of 38 and 37. On one fold, we choose both a separate best treatment for each start and a single fixed treatment minimizing mean error across starts; we evaluate both choices on the other fold, then swap. Ties follow T0, T1, T2 order. Held-out errors weight systems equally, accounting for the unequal fold sizes. Define $\Delta = E_{\text{fixed}} - E_{\text{per-start}}$. Qualification requires $\Delta/E_{\text{fixed}} \geq 5\%$ and a positive lower 95% bound for Δ . The 10,000-draw bootstrap (seed 202610074) resamples source trajectories within each arm and paired systems within each fixed fold, reselecting both policies in every draw.

The gate additionally requires at least two distinct treatments, each a reliable winner at two or more independent starts. A reliable winner must be chosen in both folds and beat both alternatives. We use 50,000 paired system-bootstrap draws (seed 202610075) and a predeclared family of 27 unordered pairwise contrasts: three per start. Bonferroni intervals have individual coverage $1 - 0.05/27$ (approximately 99.815%). These are approximate intervals conditional on the trained models; they do not resolve continuation-seed uncertainty.

The fixed choice is T2 in both folds. Its cross-fitted MSE is 0.001890030, compared with 0.001895075 for per-start selection: a relative improvement of -0.267% , with descriptive 95% interval $[-1.314\%, 0.908\%]$. The raw difference is -5.045×10^{-6} , with 95% interval $[-25.315, 17.046] \times 10^{-6}$. Three individual family-corrected contrasts exclude zero: T2 beats T0 for B1 seed 1 and B2 seed 1, and T1 beats T0 for B2 seed 2. Yet no start has a reliable winner over *both* alternatives. Consequently, the selection-space gate fails and the new-source extension is not released. This is not an absence of every treatment effect. It shows that substantial information value can coexist with insufficient demonstrated selection space for these interventions and budgets. With three sources per arm and one continuation per cell, this remains a development study; no FUV predictor is evaluated and the original W1 stopping decision is unchanged.

D.3 A function-preserving conditioning test on Ant

The unsuccessful consumer-repair study leaves open whether the latent coordinates make adaptation unnecessarily difficult. We tested one concrete version of this explanation: an invertible, train-only change of coordinates at the representation-consumer interface. This is a separately frozen mechanism follow-up on the three existing end-to-end Ant sources, not a new evaluation of the original repair gate. Its primary comparison is regularized ZCA conditioning against a raw-coordinate control, both with reset consumer Adam states and the same continuation budget.

Coordinate change and initial function preservation. For each source, we used its 640 archived training-context embeddings: 32 contexts from each of twenty training tasks. The encoder and context aggregator remain frozen. We computed a single global mean μ and covariance C in the original 41-dimensional latent space, weighting tasks equally and then contexts equally within task. No physical task labels were supplied to the consumer, and no development outcomes selected the transform. With $C = Q \text{diag}(c_i) Q^\top$, we fixed

$$\lambda = 0.01 \text{tr}(C)/41, \quad S = Q \text{diag}(\sqrt{c_i + \lambda}) Q^\top, \quad u = S^{-1}(z - \mu). \quad (14)$$

The positive additive ridge makes S invertible. There is no eigenvalue clipping, projection, dimension deletion, or clipping of transformed inputs. This is regularized ZCA: the transformed covariance need not equal the identity. Centering is global rather than task-specific.

For each first affine layer, write the contribution of the latent coordinates as $W_z z + b$. In the conditioned arm, we set $W_u = W_z S$ and $b' = b + W_z \mu$, leaving observation and action weights unchanged. Thus $W_u u + b' = W_z z + b$ before continuation. We apply this compensation to the actor, both critics, and both target critics, retaining the source’s target-network weights. Every formal job passed a train-only initial check of deterministic actor actions, actor means and log standard deviations, and all four critic outputs, using absolute tolerance 10^{-5} and relative tolerance 10^{-4} . These checks neither advance the training sampler nor consume its private reinforcement-learning random streams.

Paired continuation and evaluation. Both arms start from the same complete 200,000-update source checkpoint. They reset the actor and critic Adam states, retain the original source sampler cursor and private random states, and run exactly 40,000 updates with the original SAC/CQL kernel and hyperparameters. In particular, the raw arm is an Adam-reset control; it is not a relabeling of the historical T1 continuation. The encoder, aggregator, auxiliary modules, and their optimizer states remain unchanged. Training receipts verify identical sample-index chains and private-random-state hashes within each source pair, together with exactly 40,000 steps for each consumer optimizer. All six cells are included, regardless of effect direction; there is no selected checkpoint, adaptive budget extension, or hyperparameter retuning.

At inference the adapter is applied exactly once, immediately before the consumer. The diagnostic cut remains the original raw z , so this intervention does not alter the representation used by the existing formation probe. We do not infer that a newly refitted, regularized probe in transformed coordinates would have an identical finite-sample score.

Evaluation uses the original ten development tasks, twenty fixed context/reset pairs per task (reset seeds 50000–50019), and 200-step episodes with deterministic actions. There are 200 rollouts per cell and 1,200 in total, but only three independently trained source models. The prespecified primary contrast is conditioned minus raw-control return at 40,000 updates. We first average the twenty resets within each source–task cell, then average the paired differences. The 95% percentile interval uses 10,000 crossed bootstrap draws of three source indices and ten task indices, with the same draws for both arms (PCG64 seed 2026100705); reset indices are held fixed. This procedure differs from the original repair gate’s source/reference resampling and does not replace that gate.

Source seed	Raw + Adam reset	ZCA + Adam reset	ZCA – raw
0	71.59	79.84	+8.25
1	94.74	77.86	–16.88
2	78.46	59.36	–19.10
Mean	81.60	72.35	–9.24

Table 10: Fixed-endpoint Ant conditioning comparison. Each entry averages ten development tasks and twenty resets. The paired mean contrast is -9.24 , with a 95% crossed source–task bootstrap interval of $[-26.41, 7.78]$. Source seeds, not individual rollouts or the six treatment cells, are the model replications.

Result and interpretation. The tested conditioning intervention did not demonstrate a reliable return gain over the paired raw-coordinate control (Table 10). Its point estimate is lower, with a positive difference for one source and negative differences for two. Because the interval crosses zero, this is neither an established performance decrease nor an equivalence result. It shows that this specific, function-preserving coordinate change was insufficient to establish improvement under the fixed optimizer and budget. It does not rule out other regularization strengths, optimizers, learning rates, budgets, or conditioning procedures, and it does not identify a unique cause of the earlier repair failure. No original STOP decision is changed, and this development follow-up supplies no prospective intervention-selection or cross-environment claim.

Implementation corrections and evidence preservation. Before any affected scientific rollout, two operational faults were corrected through separately recorded evaluation entry points: missing checkpoint identity in cached-source metadata, and selection of an interpreter without the accepted MuJoCo runtime. The corrected evaluator retains the original numerical evaluation body and uses the previously verified Ant interpreter and native-library fingerprint. The two already-completed seed-0 training jobs were adopted by their unchanged checkpoint and receipt hashes, rather than retrained. Failed zero-rollout attempts remain archived. All six jobs used the same corrected evaluator; their complete trajectories, terminal checkpoints, and receipts were persistently backed up and hash-verified before the single scientific scoring pass. These corrections did not change the transform, training, evaluation grid, scoring rule, or sealed-data boundary.

E Additional numerical results

All entries below format already-scored summaries; no statistical analysis is rerun for these tables. Complete precision, additional per-start intervals, and source hashes are retained in the accompanying numerical exports.

Policy	I1 regret	I1 MSE	I2 regret	I2 MSE	W1 regret	W1 MSE
FUV	0.12286	37061.973	0.06982	989.518	0.02809	277.243
Score/reference	0.06656	36454.778	0.00707	908.057	0.03318	276.016
Current score	0.03238	35734.905	0.00803	909.665	0.04494	277.358
Progress	0.03238	35734.905	0.00707	908.057	0.03785	277.180
Fixed treatment	0.03238	35734.905	0.29563	1236.957	0.03785	277.180
Probe only	0.03238	35734.905	0.07399	997.705	0.03160	277.004
RankMe	0.03238	35734.905	0.16869	1035.881	0.02691	276.774
LogME	0.03238	35734.905	0.01729	927.465	0.03065	276.870
FUV without δ	0.08286	36735.623	0.08434	993.648	0.03900	277.618
Random	0.23790	39338.240	0.26065	1079.696	0.05660	311.064

Table 11: Original within-setting development prediction. Regrets are in G units; all MSE columns display raw MSE multiplied by 10^6 (native output units for ICL, raw coordinates for W1). There are 30/12/18 starts but 15/6/18 source groups. These point estimates do not establish differences without their paired uncertainty. RankMe and LogME are feature baselines, not independently selected treatment budgets.

Source arm	Seed	Linear R^2	MLP R^2	k NN R^2
B0	0	0.9760	0.9826	0.9493
B0	1	0.9746	0.9819	0.9519
B0	2	0.9737	0.9826	0.9532
B0_split	0	0.9803	0.9874	0.9766
B0_split	1	0.9793	0.9877	0.9788
B0_split	2	0.9791	0.9869	0.9772
B1	0	0.2214	0.3280	0.2346
B1	1	0.2061	0.3387	0.2465
B1	2	0.2161	0.3268	0.2300
B2	0	0.9962	0.9967	0.9969
B2	1	0.9961	0.9967	0.9971
B2	2	0.9962	0.9967	0.9971
B3	0	0.9960	0.9964	0.9966
B3	1	0.9960	0.9967	0.9971
B3	2	0.9963	0.9969	0.9970
random_B2	0	0.2056	0.3075	0.2450
random_B2	1	0.1911	0.3309	0.2830
random_B2	2	0.2153	0.3192	0.2709

Table 12: All eighteen preintervention W1 probe comparisons. Scores are raw held-out R^2 , not reader-specific normalized ratios. Each seed is a distinct source model; the three readers share its data split. In particular, the six B1/random_B2 linear scores are similarly low descriptively, not a statistical equivalence result.

Arm	Before	T0	T1	T2	Best by seed 0/1/2
B0	176.34	182.51	172.88	177.59	T1,T1,T1
B0_split	162.95	163.39	158.66	161.97	T2,T1,T0
B1	1282.19	930.40	932.99	928.87	T0,T2,T2
B2	37.00	41.85	35.92	36.86	T2,T2,T1
B3	37.85	42.97	36.65	37.57	T2,T2,T1
random_B2	1267.90	900.62	885.80	320.22	T2,T2,T2

Table 13: W1 raw task MSE $\times 10^6$, averaged over three source seeds and the full 75-system effect bank. Best labels are unthresholded point winners, not claims that differences exceed noise. B1 improves absolutely under all treatments.

Arm	Seed	Median	95th pct.	Max.	Zero (%)	Cap (%)	ESS/ N
B0	0	0.29	5.70	6.29	0.04	4.42	0.281
B0	1	0.28	5.87	6.36	0.10	4.52	0.275
B0	2	0.26	6.10	6.47	0.12	4.65	0.267
B0_split	0	0.30	5.53	6.20	0.18	4.30	0.288
B0_split	1	0.28	5.81	6.34	0.20	4.46	0.277
B0_split	2	0.30	5.57	6.23	0.25	4.35	0.286
B1	0	0.49	4.04	5.31	0.01	2.65	0.384
B1	1	0.48	4.09	5.35	0.01	2.79	0.379
B1	2	0.47	4.17	5.35	0.01	2.86	0.374
B2	0	0.36	4.91	5.80	0.01	3.73	0.320
B2	1	0.35	4.98	5.86	0.02	3.78	0.315
B2	2	0.36	4.91	5.82	0.01	3.74	0.320
B3	0	0.37	4.83	5.73	0.01	3.65	0.327
B3	1	0.35	5.01	5.88	0.01	3.82	0.314
B3	2	0.36	4.90	5.79	0.01	3.72	0.321
random_B2	0	0.27	5.90	7.25	0.04	3.92	0.250
random_B2	1	0.24	6.10	7.34	0.07	4.04	0.241
random_B2	2	0.25	6.05	7.29	0.04	4.01	0.245

Table 14: T4 fixed training-bank weights for every start ($N = 200,000$ each). Final mean is one up to floating-point precision. Cap is the fraction clipped at the intermediate upper cap five; final global renormalization permits weights above five. The recorded concentration diagnostic is $ESS = (\sum_i w_i)^2 / \sum_i w_i^2$, not a count of independent systems. It describes concentration without proving the cause of worse task error.

Table 15: I3 per-task percentages and original 95% query-cluster intervals. Each cell has 512 episodes, with repeated demonstrations grouped by query identity. Checkpoints share one trajectory. Final-calibration task/layer selection is retained; these are not independent final-test confirmations. Macro curves use saved point summaries, not averages of confidence endpoints.

Task	Step	Identity recall	Ten-shot accuracy	Zero-shot accuracy
antonym	1k	93.36 [90.75,95.82]	0.59 [0.00,1.32]	0.78 [0.00,2.43]
antonym	48k	99.61 [99.01,100.00]	16.02 [11.02,21.53]	0.00 [0.00,0.00]
antonym	96k	100.00 [100.00,100.00]	31.25 [24.85,38.12]	0.00 [0.00,0.00]
antonym	143k	99.80 [99.38,100.00]	36.52 [29.90,43.48]	0.00 [0.00,0.00]
capitalize	1k	100.00 [100.00,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
capitalize	48k	100.00 [100.00,100.00]	94.34 [89.51,98.02]	0.00 [0.00,0.00]
capitalize	96k	100.00 [100.00,100.00]	99.22 [97.99,100.00]	0.00 [0.00,0.00]
capitalize	143k	100.00 [100.00,100.00]	99.41 [98.42,100.00]	0.00 [0.00,0.00]
country-capital	1k	100.00 [100.00,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
country-capital	48k	100.00 [100.00,100.00]	25.20 [12.90,38.06]	0.00 [0.00,0.00]
country-capital	96k	100.00 [100.00,100.00]	44.34 [26.19,62.52]	5.47 [0.00,16.77]
country-capital	143k	100.00 [100.00,100.00]	57.23 [38.04,75.65]	0.00 [0.00,0.00]
country-currency	1k	100.00 [100.00,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
country-currency	48k	100.00 [100.00,100.00]	6.45 [1.31,14.15]	0.00 [0.00,0.00]
country-currency	96k	100.00 [100.00,100.00]	18.55 [7.17,32.39]	0.00 [0.00,0.00]
country-currency	143k	100.00 [100.00,100.00]	14.06 [5.30,24.95]	0.00 [0.00,0.00]
english-french	1k	99.61 [99.01,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
english-french	48k	99.80 [99.39,100.00]	22.07 [17.66,26.78]	0.00 [0.00,0.00]
english-french	96k	100.00 [100.00,100.00]	32.62 [27.27,38.12]	0.00 [0.00,0.00]
english-french	143k	99.80 [99.40,100.00]	32.23 [27.04,37.53]	0.00 [0.00,0.00]
english-german	1k	99.41 [98.65,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
english-german	48k	100.00 [100.00,100.00]	4.10 [2.28,6.16]	0.00 [0.00,0.00]
english-german	96k	100.00 [100.00,100.00]	5.47 [3.25,7.95]	0.00 [0.00,0.00]
english-german	143k	100.00 [100.00,100.00]	6.64 [4.12,9.38]	0.00 [0.00,0.00]
english-spanish	1k	99.41 [98.65,100.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
english-spanish	48k	100.00 [100.00,100.00]	27.15 [22.44,32.06]	0.00 [0.00,0.00]
english-spanish	96k	100.00 [100.00,100.00]	36.91 [31.74,42.22]	0.00 [0.00,0.00]
english-spanish	143k	100.00 [100.00,100.00]	39.65 [34.07,45.33]	0.00 [0.00,0.00]
present-past	1k	100.00 [100.00,100.00]	0.20 [0.00,0.60]	0.00 [0.00,0.00]
present-past	48k	100.00 [100.00,100.00]	76.56 [62.35,89.35]	0.00 [0.00,0.00]
present-past	96k	100.00 [100.00,100.00]	86.91 [73.41,97.72]	0.00 [0.00,0.00]
present-past	143k	100.00 [100.00,100.00]	89.26 [76.64,99.37]	4.69 [0.00,14.37]
singular-plural	1k	99.80 [99.39,100.00]	0.39 [0.00,1.23]	0.00 [0.00,0.00]
singular-plural	48k	100.00 [100.00,100.00]	93.36 [84.17,100.00]	0.00 [0.00,0.00]
singular-plural	96k	100.00 [100.00,100.00]	93.55 [83.74,100.00]	0.00 [0.00,0.00]
singular-plural	143k	100.00 [100.00,100.00]	96.88 [90.45,100.00]	0.00 [0.00,0.00]
synonym	1k	90.82 [87.98,93.42]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
synonym	48k	100.00 [100.00,100.00]	1.56 [0.39,3.03]	0.98 [0.00,2.61]
synonym	96k	100.00 [100.00,100.00]	0.39 [0.00,0.99]	0.39 [0.00,1.23]
synonym	143k	100.00 [100.00,100.00]	0.39 [0.00,1.00]	0.39 [0.00,1.23]

Table 16: Original Ant assessment intervals, three source seeds and Student- t critical value for two degrees of freedom. Return metrics use raw episodic return; the others are dimensionless. Wide or physically out-of-range interval endpoints are retained, not clipped. The later follow-up uses a different, explicitly specified empirical bootstrap.

Variant	Metric	Mean	95% interval
focal	F	0.660	[0.618,0.702]
focal	U_dep	0.320	[0.312,0.328]
focal	U_corr	0.609	[0.556,0.662]
focal	rho	0.711	[0.458,0.965]
focal	J_ctx	57.572	[36.524,78.620]
focal	J_decode	84.313	[74.882,93.745]
focal	J_meanz	-0.835	[-2.575,0.905]
corro	F	0.424	[-0.719,1.568]
corro	U_dep	0.121	[-0.368,0.611]
corro	U_corr	0.216	[-0.690,1.122]
corro	rho	0.311	[-1.051,1.674]
corro	J_ctx	23.854	[-90.126,137.833]
corro	J_decode	61.385	[-14.506,137.276]
corro	J_meanz	-1.307	[-5.090,2.477]
unicorn	F	0.726	[0.699,0.754]
unicorn	U_dep	0.331	[0.318,0.344]
unicorn	U_corr	0.614	[0.590,0.637]
unicorn	rho	0.768	[0.591,0.945]
unicorn	J_ctx	62.321	[47.759,76.884]
unicorn	J_decode	90.260	[80.112,100.408]
unicorn	J_meanz	-0.582	[-1.059,-0.106]
e2e	F	0.915	[0.884,0.947]
e2e	U_dep	0.363	[0.341,0.386]
e2e	U_corr	0.696	[0.662,0.731]
e2e	rho	0.820	[0.475,1.165]
e2e	J_ctx	66.496	[38.083,94.909]
e2e	J_decode	89.487	[83.606,95.367]
e2e	J_meanz	-1.079	[-2.034,-0.123]
random	F	0.281	[0.265,0.297]
random	U_dep	0.162	[0.130,0.195]
random	U_corr	0.189	[0.039,0.340]
random	rho	0.221	[0.043,0.399]
random	J_ctx	11.640	[-3.738,27.018]
random	J_decode	63.027	[52.713,73.341]
random	J_meanz	-8.176	[-14.402,-1.950]

F Complete W1 repair trajectories

The following plots retain all 264 saved measurements from the twelve two-course trajectories. They use the fixed 25-system subset; the full-75-system endpoint comparisons in the main text are separate. Curves are descriptive, without a newly computed simultaneous confidence band.

B1: reader and staged adaptation

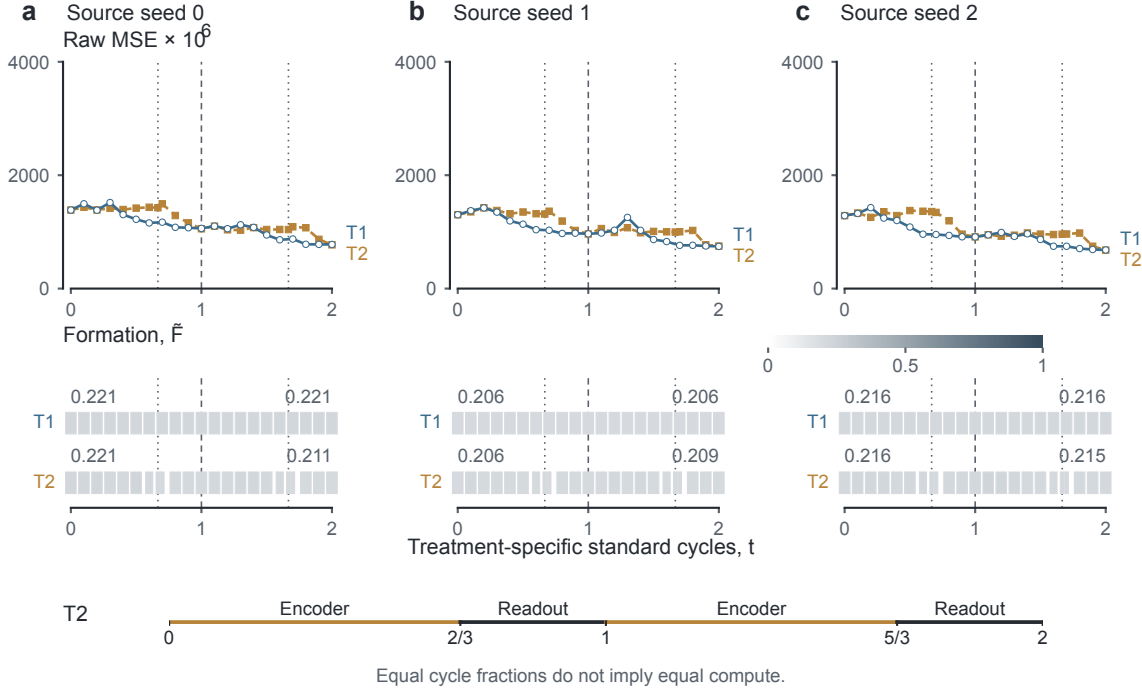


Figure 4: **B1: complete individual-source repair trajectories.** Columns show source seeds 0–2. Upper panels retain all recorded fixed-25-system raw-coordinate MSE values ($\times 10^6$) for T1 (consumer-only updates; blue circles and solid lines) and T2 (staged native-plus-alignment source updates followed by reader updates; ochre squares and dashed lines). Lower panels encode each recorded linear formation score $\tilde{F}$ as a discrete tile centered at its original measurement position; endpoint labels are rounded to three decimals. Tile widths provide separation between adjacent observations and do not represent time intervals, interpolation, or additional measurements. All 132 records are displayed in both metrics: 21 T1 and 23 T2 observations per source. The MSE range (0–4000) and formation color scale (0–1) are shared with the random-source trajectories in Figure 5. Dotted guides at $t = 2/3$ and $5/3$ mark the recorded T2 source-to-reader boundaries; the dashed guide at $t = 1$ marks the second cycle. The shared stage ruler describes T2: one cycle comprises 4,000 source updates and 2,000 reader updates, whereas T1 uses 2,000 reader updates. The horizontal coordinate is a treatment-specific update-budget fraction, not matched time or compute. The lines connect observed MSE values without smoothing. These individual-source trajectories are descriptive: no new intervals or simultaneous confidence bands are computed, and the 75-system endpoint estimates are not substituted for curve measurements. Formation values are reused only when the source hash is unchanged, including T1 and the reader phases of T2; tiles show recorded values, not independent probe evaluations.

Random-pair source: staged adaptation

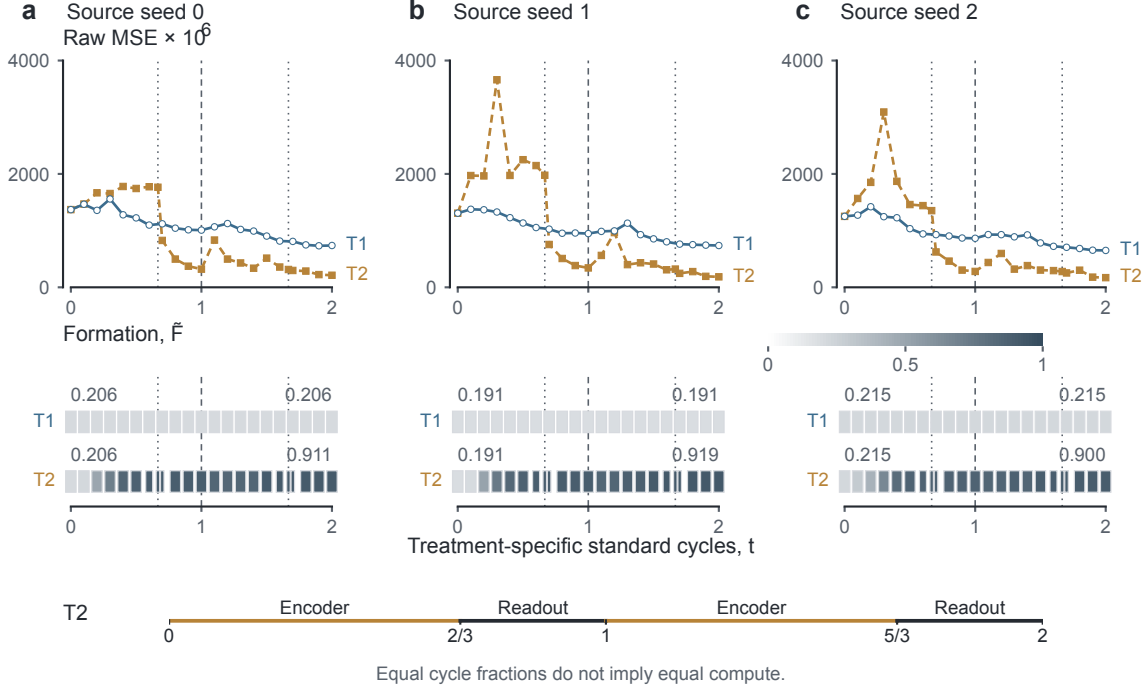


Figure 5: **Formation and prediction error during random_B2 adaptation.** Each column follows one of the three trained random_B2 sources, whose initial training used a fixed pseudo-system pairing. Top: mean raw-coordinate MSE on the same 25 effect systems at every measurement. Bottom: linear formation $\tilde{F}$, with one discrete color tile centered at each measured budget fraction. Tile width only separates measurements; it does not represent an interval or interpolation. All 21 T1 and 23 T2 measurements per source are shown, including T2 source-to-reader boundaries at $2/3$ and $5/3$ cycles. The dashed vertical line at one cycle marks the start of the second cycle. T1 keeps the source fixed and updates the reader; T2 alternates source (native plus alignment) and reader updates. A standard cycle comprises 2,000 reader updates for T1, versus 4,000 source plus 2,000 reader updates for T2, so the horizontal axis is treatment-specific budget fraction, not matched time or compute. Formation gains and prediction-error reductions occur at different stages in these descriptive trajectories. No new uncertainty bands are computed, and the separately evaluated 75-system endpoints are not inserted into the curves. Formation values are reused only when the source hash is unchanged, including T1 and the reader phases of T2; tiles show recorded values, not independent probe evaluations. The MSE range (0–4000) and formation color scale (0–1) are shared with Figure 4.