
Beyond Accuracy and Alignment: A Diagnostic Evaluation Protocol for Feedback Alignment

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Modern feedback-alignment evaluation on deep residual networks is still summar-
2 ized by a deceptively simple pair: headline accuracy and headline cosine align-
3 ment Γ to the backpropagation gradient. We show that this pair can silently fail in
4 two distinct ways on standard CIFAR-10 pre-LayerNorm ResMLP and ViT-Mini
5 settings: first, *measurement degeneracy*, where residual-stream growth drives
6 hidden-layer BP gradients to the numerical floor and makes Γ uninterpretable;
7 and second, *low intrinsic credit-direction quality*, where random-feedback credit
8 remains essentially unaligned with BP on the deep blocks even when the reference
9 gradient is still meaningful. The headline result is that the field-standard reporting
10 pair walks back none of the methods we audit, whereas a four-diagnostic proto-
11 col walks back the three degenerate methods and passes the two trustworthy con-
12 trols. Intervention with a per-block scale-control penalty further reveals method-
13 dependent severity within the audited fixed-feedback family: State Bridge then
14 exceeds the architecture-matched frozen-blocks baseline by about 10 percentage
15 points, while Credit Bridge attains much higher deep BP cosine than DFA at the
16 same final accuracy, a dissociation that motivates reporting layerwise credit quality
17 jointly with a depth-utilization baseline. Our contribution is an evaluation method-
18 ology paper for the NeurIPS 2026 Evaluations & Datasets track: we provide the
19 protocol, the calibration logic for its thresholds, a reference implementation, a five-
20 method audit, and validation through temporal replay, cross-architecture checks,
21 intervention-based disambiguation, and a documented catalog of pipeline pitfalls,
22 in the spirit of critical evaluation analyses such as Jordan et al. [3], O’Bray et al.
23 [2], Paleka et al. [1].

24 1 Introduction

25 Feedback-alignment papers are usually judged by two numbers: task accuracy and an aggregate
26 similarity between the method’s local credit signal and the backpropagation gradient [4–7]. On
27 the audited 4-block $d=256$ ResMLP, however, Table 1 already shows that this pair is not a validity
28 check: DFA reaches only 0.306 ± 0.006 test accuracy, below the architecture-matched frozen-blocks
29 baseline of 0.349 ± 0.002 , while still looking superficially comparable to other non-BP methods.
30 Figure 1 further shows that the apparent cosine evidence is concentrated at the shallowest block,
31 with DFA at seed 42 reaching about $+0.42$ at layer 0 but approximately -0.03 to 0 on layers 1–4, so
32 the aggregate obscures where credit direction is and is not present. At the same time, the deepest BP
33 reference norm is only about 5×10^{-10} for DFA, State Bridge, and Credit Bridge, below the 10^{-8}
34 clamp used by `F.cosine_similarity`, whereas BP remains around 4×10^{-4} , so the reported deep
35 cosine is partly computed against a numerical-floor reference rather than an informative gradient

Table 1: Main audit table for the 4-block $d=256$ pre-LayerNorm ResMLP on CIFAR-10. The row and column structure is fixed here; fill from the three-seed audit output.

Method	Test acc.	Headline Γ	Status-quo verdict	Protocol verdict
BP	0.615 ± 0.003	≈ 1.0	trustworthy	trustworthy
EP	0.316 ± 0.030	0.008	trustworthy	trustworthy
DFA	0.306 ± 0.006	0.10	trustworthy	walked back
State Bridge	0.205 ± 0.032	0.005	trustworthy	walked back
Credit Bridge	0.289 ± 0.026	0.07	trustworthy	walked back

36 direction (Figure 1; Table 1). Those numbers can be useful, but only if the measurement regime
37 itself is valid.

38 Our audit shows that modern residual vision models can make these two quantities look informative
39 while failing to answer the question they are taken to answer. Figure 1 shows the first failure mode,
40 which we call *Mode 1: measurement degeneracy*, where residual-stream growth drives the deepest
41 hidden state to about $\|h_L\| \sim 10^8$ under DFA/SB/CB while the corresponding BP reference col-
42 lapses to $\|g_L\| \sim 5 \times 10^{-10}$, so the deep-layer cosine is measured against a clamp-dominated floor
43 rather than a meaningful target direction. The same figure also shows the second failure mode, *Mode*
44 *2: low intrinsic credit-direction quality*, because even after comparing against the stronger frozen-
45 blocks baseline (0.349 ± 0.002) and looking layer-by-layer, DFA’s deep blocks remain essentially
46 null while only layer 0 is visibly positive. Intervention sharpens both modes. Adding a per-block
47 residual penalty $\lambda \|f_i(h_i)\|^2$ to DFA at $\lambda=10^{-2}$ contains $\|h_L\|$ to about 4×10^4 and lifts the deep BP
48 reference to about 10^{-6} , but DFA’s rescued deep cosine is only about $+0.16$; State Bridge under the
49 same intervention reaches a three-seed deep cosine of $+0.32$ and, unlike DFA, exceeds the frozen-
50 blocks baseline by $+10$ points in final accuracy; Credit Bridge reaches a deep cosine near $+0.68$
51 yet matches only the DFA accuracy, so Mode 2 has method-dependent severity and deep cosine is
52 not a sufficient predictor of final accuracy across methods. At the same time, at $\lambda=10^{-4}$ Mode 1 is
53 alleviated while the DFA deep cosine still stays near zero, and at vanilla DFA epoch 1 the reference
54 is already meaningful at about 6×10^{-7} but the deep cosine is still -0.008 ± 0.013 across three
55 seeds. The failure is therefore neither unitary nor uniform: Mode 1 and Mode 2 are observationally
56 separable, and within the audited fixed-feedback family, the severity of each mode varies by method.

57 Accordingly, this paper does not introduce a new FA variant or a new benchmark. Of the five
58 methods we audit, BP, EP, and DFA are established baselines from the published literature; the
59 remaining two, which we call *State Bridge* and *Credit Bridge*, are diagnostic probes we construct
60 in this paper to directly learn the two targets that different strands of the BP-free literature argue
61 should produce good per-layer credit (formal definitions and citations in Section 2). Instead, Table 1
62 and Figure 1 use a standard five-method CIFAR-10 audit to show that status-quo reporting would
63 treat BP, EP, DFA, State Bridge, and Credit Bridge as the same kind of evidence-bearing object
64 even though only BP and EP remain trustworthy under matched diagnostic checks. This makes the
65 contribution methodological in the sense of Jordan et al. [3], O’Bray et al. [2], and Paleka et al. [1]:
66 the central question is not whether one more FA variant can post a headline number, but whether the
67 reporting pipeline distinguishes meaningful credit-direction evidence from numerical-floor artifacts
68 and from shallow-only learning. The protocol therefore starts from per-layer diagnostics and a
69 frozen-blocks baseline before reading any aggregate cosine or final accuracy as evidence about deep
70 credit assignment. We first show the walk-back on a standard audit, then isolate the two failure
71 modes, and finally state the reporting protocol that future FA papers should satisfy.

72 2 Audit: Standard Reporting Walks Back Nothing

73 Table 1 fixes the canonical audit to a 4-block pre-LayerNorm ResMLP with width $d=256$ on CIFAR-
74 10, trained for 100 epochs with AdamW (learning rate 10^{-3} , weight decay 0.01), a cosine schedule,
75 and three seeds (42, 123, 456); all five methods are read against the same architecture, optimizer,
76 and training budget, and Figure 1 summarizes the corresponding per-block growth, deepest-layer
77 BP reference norm, cross-batch stability, and frozen-baseline comparison.

78 Two rows in Table 1, *State Bridge* (SB) and *Credit Bridge* (CB), are diagnostic probes we
 79 construct in this paper, not prior FA variants. Each directly learns a target that a different
 80 strand of the BP-free literature argues should produce good per-layer credit, and each uses the
 81 same block local loss $-\langle f_l(h_l), a_l \rangle$ as DFA but with a different a_l . SB instantiates the target-
 82 propagation view that accurate prediction of a downstream hidden state yields a usable credit
 83 signal [13, 14]: an auxiliary $G_\psi(h_l, t_l, s)$ is fit by MSE to predict h_L from $(h_l, t_l=l/L, s=e_T)$,
 84 and $a_l^{\text{SB}} = \nabla_{h_l} \text{CE}(W_{\text{out}} \text{LN}(G_\psi(h_l, t_l, s)), y)$. CB instantiates the synthetic-gradient view that a
 85 learned value network, if its input-gradient approximates the BP gradient, can stand in for it [15]:
 86 $V_\phi(h_l, t_l, s)$ is fit via a bridge residual against an EMA target, and $a_l^{\text{CB}} = \nabla_{h_l} V_\phi(h_l, t_l, s)$. Both
 87 auxiliaries are trained on detached hidden states. We use SB and CB as controls that populate different
 88 points in the (angular agreement with BP, functional usefulness) plane; that is what makes the
 89 cross-method cosine-versus-accuracy dissociation in Section 4 visible.

90 By the field’s usual criteria, the non-BP methods appear to train to nontrivial accuracy and report
 91 nonzero alignment. In Table 1, DFA reaches 0.306 ± 0.006 test accuracy with headline $\Gamma=0.10$,
 92 State Bridge reaches 0.205 ± 0.032 with $\Gamma=0.005$, and Credit Bridge reaches 0.289 ± 0.026 with
 93 $\Gamma=0.07$; none of these rows looks like an obvious invalidation if one is reading the usual pair of final
 94 accuracy and aggregate alignment in the style of prior FA reporting [4–7]. Even the absolute scale
 95 does not itself force a walk-back, because all three methods are plainly above chance and all three
 96 report positive headline alignment rather than a visibly broken or undefined quantity. That reading
 97 is exactly what the rest of the paper overturns.

98 Low accuracy by itself is not the pathology. Equilibrium Propagation (EP), a contrastive energy-
 99 based alternative to BP that updates weights from the difference between a free-phase and a nudged-
 100 phase hidden trajectory, is the key internal comparison in Table 1 and Figure 1: it achieves only
 101 0.316 ± 0.030 accuracy and a very small headline $\Gamma=0.008$, yet its per-block growth is only $11.6\times$,
 102 its deepest BP reference norm remains around 1.3×10^{-4} rather than collapsing to the numerical
 103 floor, and its cross-batch direction-stability score is 0.02 rather than the much higher drift-dominated
 104 values seen for DFA-family methods. At the same time, EP is not a positive result for depth usage
 105 in the stronger sense, because its trainable-model accuracy is still 3.3 percentage points below the
 106 frozen-blocks baseline of 0.349 ± 0.002 . The distinction matters because it separates underperform-
 107 ance from invalid evaluation.

108 When we compare each method to a frozen-blocks baseline matched to the same architecture, the
 109 headline interpretation changes immediately. The frozen-blocks model, which trains only the em-
 110 bedding, LayerNorm, and head while holding the residual blocks fixed, reaches 0.349 ± 0.002 across
 111 the same three seeds; against that baseline, BP is higher by 26.6 points, but DFA is lower by 4.3
 112 points, State Bridge by 14.4 points, Credit Bridge by 6.0 points, and even EP by 3.3 points. Fig-
 113 ure 1 shows that this accuracy comparison lines up with the diagnostic split: DFA, State Bridge, and
 114 Credit Bridge also combine extreme per-block growth ($237\times$, $12000\times$, and $96\times$), deepest-layer BP
 115 norms around 10^{-9} , and high cross-batch instability (0.16, 0.53, and 0.37), so their deep blocks are
 116 at best passengers and in practice often harmful. This establishes the audit question the rest of the
 117 paper must answer: why do the standard signals fail so badly?

118 3 Failure Mode 1: Measurement Degeneracy

119 Mode 1 has two parts. The activation-growth part (a) is a scale pathology of fixed-feedback local-
 120 credit objectives without an effective scale-control term: for block l , DFA, State Bridge, and Credit
 121 Bridge each update f_l by reducing a local loss of the form $-\langle f_l(h_l), a_l \rangle$, where the per-layer credit
 122 vector a_l is the method-specific projection of the output error (for DFA, $a_l = B_l^\top e_T$ with a fixed
 123 random B_l ; for State Bridge, a_l is the gradient of a cross-entropy loss measured through a learned
 124 state predictor $G_\psi(h_l, t_l, s)$ that estimates h_L ; for Credit Bridge, a_l is the gradient of a learned
 125 value network $V(h_l, t_l, s)$). None of these three local losses contains a penalty on $\|f_l(h_l)\|$, so any
 126 direction in which a larger block output improves inner-product alignment with the method’s fixed
 127 or learned credit target is rewarded; in a pre-LN residual stack, larger block outputs directly increase
 128 residual-stream scale, and terminal LayerNorm at the output removes task-loss sensitivity to that
 129 scale, so the architecture supplies no global restraint on the local growth incentive. The gradient-
 130 floor part (b) follows from the LayerNorm Jacobian: in terminal-LN architectures $\partial \text{LN}(h)/\partial h \propto$
 131 $1/\|h\|$ in expectation, so the same residual-stream inflation is accompanied by collapse of the hidden-
 132 layer BP reference norm. Empirically, on the audited 4-block pre-LayerNorm ResMLP ($d=256$,

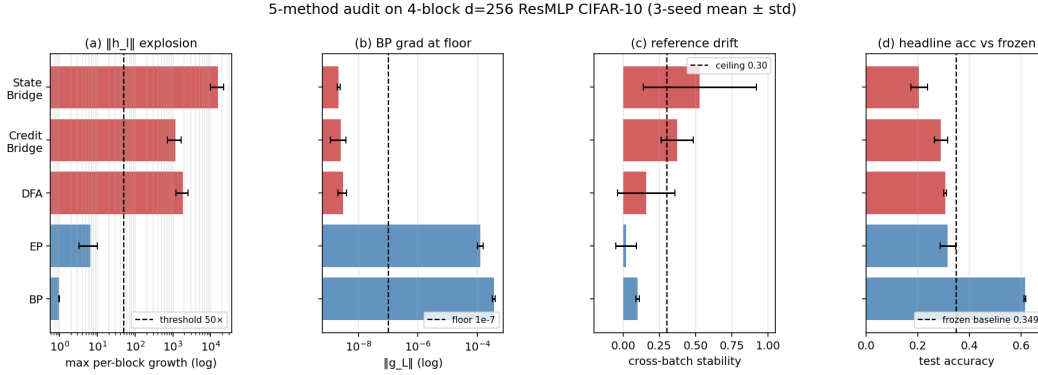


Figure 1: Five-method audit on the 4-block $d=256$ pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

133 CIFAR-10, 100 epochs, 3 seeds), DFA training drives $\|h_L\|$ from about 9 at initialization to about
 134 4×10^8 by epoch 100 and $\|g_L\|$ from about 9.8×10^{-4} to about 5×10^{-10} , while the reported deep
 135 cosine remains defined only because `F.cosine_similarity` clamps the denominator at $\varepsilon=10^{-8}$
 136 (Table 1; Figure 1). At that endpoint the reference norm is about $20\times$ below the clamp, so the
 137 quantity being reported is effectively $(a \cdot b)/(\|a\| \max(\|b\|, 10^{-8}))$ rather than a comparison to a
 138 meaningful BP direction.

139 We tested this mechanism story against four natural alternative attributions, all of which it survives.
 140 *Not residual-skip-driven*: with terminal LN kept and the additive skip removed ($h_{l+1}=F_l(h_l)$), DFA
 141 still converges to $\|h_L\| \approx 1.06 \times 10^8$ and $\|g_L\| \approx 1.09 \times 10^{-10}$ at 100 epochs, both at the diagnostic
 142 floor (Appendix H). *Not task-signal-driven*: under i.i.d. random class targets per minibatch, DFA
 143 still reaches $\|h_L\| \approx 1.67 \times 10^8$ and $\|g_L\| \approx 8 \times 10^{-12}$ while accuracy stays at chance (Appendix I). *Not*
 144 *DFA-specific*: the same random-target ablation drives $\|h_L\|$ to 6.2×10^3 for SB and 2.0×10^4 for CB
 145 in three epochs, so all three audited fixed-feedback methods exhibit data-agnostic activation growth.
 146 *Not shared by EP*: under the same protocol, EP keeps $\|h_L\| \approx 586$ at five epochs, $25\times$ smaller than
 147 DFA’s three-epoch value, confirming that the random-target assay separates the explosion-prone
 148 fixed-feedback class from EP’s energy-based objective.

149 The matched same-backbone causal control for diagnostic (b) is removing terminal LayerNorm. On
 150 the same ResMLP- $d=256$ with the residual skip intact, 100 epochs of DFA, three seeds, the residual
 151 stream still inflates to $\|h_L\| \approx 1.21 \times 10^7$, but the deepest hidden-layer BP gradient remains at
 152 $\|g_L\| \approx 7.2 \times 10^{-4}$ (four orders of magnitude above the diagnostic (b) floor), and the final test
 153 accuracy is 0.327 ± 0.012 , statistically indistinguishable from vanilla DFA’s 0.306 ± 0.006 on the
 154 same backbone with terminal LayerNorm intact. Removing terminal LayerNorm therefore preserves
 155 Mode 1 (a) but cleanly eliminates Mode 1 (b) on the same architecture, while leaving final task
 156 accuracy essentially unchanged. Combined with the broader cross-architecture pattern (StudentNet
 157 and the BatchNorm CNN, which lack terminal LayerNorm, never trigger diagnostic (b); ViT-Mini
 158 with a terminal LN does, by epochs 2–3 (Figure 2)), terminal LayerNorm is necessary for Mode 1 (b)
 159 in the audited residual ResMLP and ViT-Mini setting. The collapse is also not a late-epoch curiosity:
 160 $\|g_L\|$ drops from 9.8×10^{-4} at epoch 0 to 6.7×10^{-8} by epoch 4 in the temporal replay across three
 161 seeds, so the protocol fires within the first 11 epochs of a 100-epoch run and is actionable as an
 162 early-stop criterion rather than a post hoc explanation. Once measurement degeneracy is identified,
 163 the next question is whether poor deep credit remains even before collapse.

164 4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

165 The second failure mode appears even in the meaningful-measurement regime. At the earliest vanilla
 166 DFA checkpoints on ResMLP, the hidden backpropagated gradient at the first deep block remains
 167 above the numerical floor: at epoch 1, $\|g_2\|$ is 6.7×10^{-7} , 6.5×10^{-7} , and 3.9×10^{-7} across the three
 168 seeds, all above the 10^{-7} threshold used to distinguish measurable from collapsed gradients. Yet the

169 corresponding deep-layer cosine values are already essentially null: across layers 1–4, all seed-level
 170 measurements at epoch 1 lie in $[-0.04, +0.02]$, with a three-seed mean of -0.008 ± 0.013 , and by
 171 epoch 2 the deep mean is still only -0.018 ± 0.018 (Table 2). This is the observational pattern pre-
 172 dicted by low credit-direction quality rather than mere disappearance of signal: the gradient is still
 173 present enough to measure, but the directions delivered to the deep network carry little agreement
 174 with backpropagation, consistent with prior concerns that alternative feedback rules can fail by sup-
 175 plying poor credit assignments even before full collapse [8, 9, 11, 10]. This rules out the simplest
 176 objection that the deep-layer null result is merely a byproduct of collapse.

177 A second metric with different numerical failure modes tells the same story. Cosine measures di-
 178 rectional agreement with the BP gradient, whereas perturbation correlation ρ measures whether the
 179 proposed update predicts the correct sign and relative magnitude of loss change under actual per-
 180 turbations; their failure modes are therefore different, especially with respect to normalization and
 181 small-denominator effects. In our controls, ρ behaves as expected, with a Taylor-ceiling positive
 182 control near $+0.997$ and a random-vector negative control near $+0.006$ (Figure 3, Table 2). On
 183 vanilla DFA, deep ρ is likewise null: for the early checkpoints where the gradients remain measur-
 184 able, the deep average is -0.003 ± 0.005 across seeds and epochs, and in a floor-level checkpoint it is
 185 $+0.002$, again indistinguishable from noise. The agreement between cosine and ρ therefore rules out
 186 the interpretation that the null deep result is an artifact of cosine’s ε -clamp or vector normalization.
 187 The deep blocks are not just hard to measure; they are receiving weakly useful directions.

188 Per-layer reporting is therefore not cosmetic. In ResMLP under vanilla DFA, the headline aggregate
 189 alignment $\Gamma \approx 0.07$ – 0.10 can look mildly positive only because layer 0 remains strongly aligned
 190 while the deep network is not: at the same early checkpoints where layers 1–4 are essentially zero,
 191 layer 0 has cosine $+0.42$, $+0.45$, and $+0.39$ across seeds (Table 2). The resulting average can there-
 192 fore be driven by the embedding layer even when the interior blocks are effectively unaligned, so
 193 aggregate reporting obscures the very distinction needed to separate “measurement collapse” from
 194 “poor credit direction.” This layer-0 dominance is specific to the ResMLP DFA setting; on ViT-Mini
 195 DFA, all layers are near zero, which strengthens the broader methodological point that alignment
 196 should be reported per layer rather than only in aggregate. With the two modes separated observa-
 197 tionally, the remaining question is whether intervention can move them independently.

198 Mode 2 has method-dependent severity within the audited fixed-feedback family once Mode 1 is
 199 alleviated. Applying the same per-block scale-control penalty $\lambda=10^{-2}$ that rescued DFA to State
 200 Bridge and to Credit Bridge on the same 4-block $d=256$ ResMLP backbone over 30 epochs and three
 201 seeds gives converged test accuracies of 0.453 ± 0.003 (SB) and 0.360 ± 0.003 (CB), with deep mean
 202 cosines of $+0.322 \pm 0.007$ (SB) and $+0.679 \pm 0.008$ (CB) and deep mean ρ of $+0.402 \pm 0.015$
 203 (SB) and $+0.464 \pm 0.025$ (CB), while DFA under the same intervention reaches 0.363 ± 0.001
 204 with deep cosine $+0.155 \pm 0.025$ and deep ρ $+0.080 \pm 0.011$ (Table 2; Appendix J). The State
 205 Bridge penalty rescue is roughly 24 percentage points above the vanilla State Bridge baseline of
 206 0.213 on the same architecture and, more importantly for the paper’s central walk-back, exceeds
 207 the architecture-matched frozen-blocks shallow baseline of 0.349 by $+10.4$ percentage points. State
 208 Bridge with the penalty intervention is therefore the first audited non-BP method whose trained deep
 209 blocks substantively improve over an architecture-matched random-block baseline; the headline ac-
 210 curacy gap is comparable to BP+penalty’s $+18.1$ pp over the same shallow baseline. Neither the
 211 activation scale nor the deep BP gradient magnitude is silenced under the penalty: $\|h_L\|$ stays at
 212 302 ± 8 for SB and 5680 ± 178 for CB, with $\|g_L\|$ at $\sim 1.8 \times 10^{-4}$ and $\sim 1.9 \times 10^{-5}$ respectively,
 213 both well within the meaningful-measurement regime, so the recovered deep cosines are computed
 214 against an informative reference and not against a numerical floor. Within this rescued regime, the
 215 three methods reveal a clean cosine-versus-accuracy dissociation. Credit Bridge achieves roughly
 216 $4\times$ the deep cosine of DFA and $2\times$ that of State Bridge, yet its final accuracy matches DFA’s and
 217 is 9 percentage points below State Bridge’s. We therefore frame the Mode 2 reading as a three-part
 218 proposition. *Observation*: under the same intervention and matched training budget, CB and DFA
 219 reach the same accuracy despite a $4\times$ deep-cosine gap, while SB is the best accuracy with interme-
 220 diate cosine. *Inference*: layerwise cosine to the BP gradient is necessary to rule out grossly wrong
 221 credit signals (it distinguishes the rescued regime from the clamp-dominated vanilla regime), but
 222 it is not sufficient to certify that the supplied signal is useful credit for depth. *Mechanism hypoth-*
 223 *esis*: usefulness depends on whether the local update induces useful forward-state change across
 224 blocks, not merely whether its direction is close to the BP gradient in angle. Under this reading, CB
 225 supplies a gradient-direction surrogate that aligns with BP in angle but does not translate to a coor-

Table 2: Two-mode validation table built around the intervention and disambiguation results.

Condition	Deep-layer alignment signal	Measurement regime	Interpretation
Vanilla DFA, early epoch	$\overline{\text{cos}}_{deep} = -0.008 \pm 0.013, \overline{\rho}_{deep} = -0.003 \pm 0.005$	meaningful ($\ g\ \sim 10^{-6}$)	mode 2 present without mode 1
Vanilla DFA, converged	$\overline{\text{cos}}_{deep} = -0.022, \overline{\rho}_{deep} = +0.002$	degenerate ($\ g\ \sim 10^{-9}$)	mode 1 obscures mode 2
Penalized DFA, $\lambda = 10^{-2}$	$\overline{\text{cos}}_{deep} = +0.155 \pm 0.025, \overline{\rho}_{deep} = +0.080 \pm 0.011$	meaningful ($\ g\ \sim 10^{-6}$)	partial alleviation of both modes
Fresh- B null control	$\overline{\text{cos}}_{deep} = +0.002 \pm 0.022$ ($n=20$ draws)	meaningful	training-specific adaptation check

226 dinated forward-state improvement, while State Bridge supplies a state-level downstream teaching
 227 signal that preserves aspects of useful credit which layerwise cosine does not measure. We state this
 228 as a mechanism hypothesis rather than a theorem because we have measured the angle-to-accuracy
 229 gap but not the full functional-credit content; the reporting rule that follows is robust to either inter-
 230 pretation. This cross-method dissociation strengthens the methodological point that alignment must
 231 be reported jointly with measurement validity and a depth-utilization baseline rather than as a single
 232 headline number.

233 5 Intervention and Cross-Architecture Evidence

234 The penalty intervention first matters as a rescue of the measurement regime. When we add a per-
 235 block penalty $\lambda \text{mean}(\|f_l(h_l)\|^2)$ to DFA’s local loss and train the 4-block $d=256$ ResMLP for 30
 236 epochs on CIFAR-10, the $\lambda=10^{-2}$ setting contains the terminal hidden-state scale from $\|h_L\| \sim$
 237 4.4×10^8 under vanilla DFA to $\sim 4.0 \times 10^4$, while lifting the deepest BP reference norm from
 238 $\|g_L\| \sim 5 \times 10^{-10}$ to $\sim 9.0 \times 10^{-7}$, a roughly four-order-of-magnitude rescue on both quantities
 239 (Figure 3; Table 2). At that setting, both diagnostic (a) and diagnostic (b) pass on penalized DFA,
 240 and test accuracy rises to 0.363 ± 0.001 from 0.308 ± 0.014 for vanilla DFA. The key point is not
 241 yet that the recovered network has good deep credit, but that the deep reference vector is again large
 242 enough to function as a meaningful target direction rather than a clamp-dominated artifact. That
 243 rescue makes the second question measurable rather than hypothetical.

244 Once the reference vector is meaningful again, the deep layers no longer sit exactly at null. At
 245 $\lambda=10^{-2}$, penalized DFA reaches a three-seed deep-layer mean cosine of $+0.155 \pm 0.025$ and deep
 246 perturbation correlation of $+0.080 \pm 0.011$, whereas vanilla DFA is essentially zero on both metrics
 247 in the deep blocks, consistent with prior concerns that alternative feedback can fail by supplying
 248 poor credit directions even before full collapse [8, 9, 11, 10]. The null calibration rules out the inter-
 249 pretation that this recovered signal is merely measurement noise: on the same penalized checkpoint,
 250 replacing the training-time feedback matrices with 20 fresh random B_l draws gives a deep cosine
 251 of only $+0.002 \pm 0.022$, with per-layer standard deviations of 0.013–0.023, all within noise of zero
 252 (Table 2). The λ sweep sharpens the dissociation further: at $\lambda=10^{-4}$, Mode 1 is already alleviated,
 253 with $\|h_L\|=2.4 \times 10^4$ and $\|g_L\|=6.3 \times 10^{-7}$, but deep cosine remains -0.022 , while at $\lambda=10^{-2}$ it
 254 rises to $+0.165$ and deep ρ to $+0.091$ (Figure 3). The improvement is real, but it is only partial.

255 A rescue intervention is only informative if its direct cost is controlled. The relevant control is BP
 256 trained under the same penalty: BP falls from 0.609 ± 0.004 without the penalty to 0.530 with
 257 $\lambda=10^{-2}$, so the penalty has a direct cost of about 8 percentage points even when credit assignment
 258 is correct, whereas DFA moves in the opposite direction, from 0.308 ± 0.014 to 0.363 ± 0.001 ,
 259 and State Bridge moves further still, from 0.213 to 0.453 ± 0.003 (three seeds), under the same
 260 intervention (Figure 3; Appendix J). Relative to the frozen-blocks baseline of 0.349 , BP+penalty
 261 retains a margin of $+18.1$ points, State Bridge+penalty retains $+10.4$ points, and DFA+penalty
 262 retains only $+1.4$ points. The remaining BP-to-DFA gap of 17 points is therefore a lower bound
 263 on the part of DFA’s deficit that is not explained by simple penalty-induced capacity loss alone,
 264 though not a clean isolation because BP uses an end-to-end loss whereas DFA uses block-local
 265 losses. The substantially smaller BP-to-State-Bridge gap of $0.530 - 0.453 = 7.7$ points shows
 266 that the cross-method differences in penalty-rescued accuracy are not all attributable to a uniform
 267 “random-feedback ceiling”: the bridge construction in State Bridge can recover much more of the
 268 BP-with-penalty performance than DFA can, on the same architecture and the same intervention.
 269 The residual gap after that control is what keeps Mode 2 substantively alive while letting it have
 270 method-dependent severity.

271 The architecture comparison sharpens the scope of the critique. In the terminal-LN architectures we
 272 audited, both diagnostics fire for DFA-trained ResMLP at $d=256$, the same pattern recurs at $d=512$

Cross-architecture temporal evolution of FA diagnostics (seed 42)

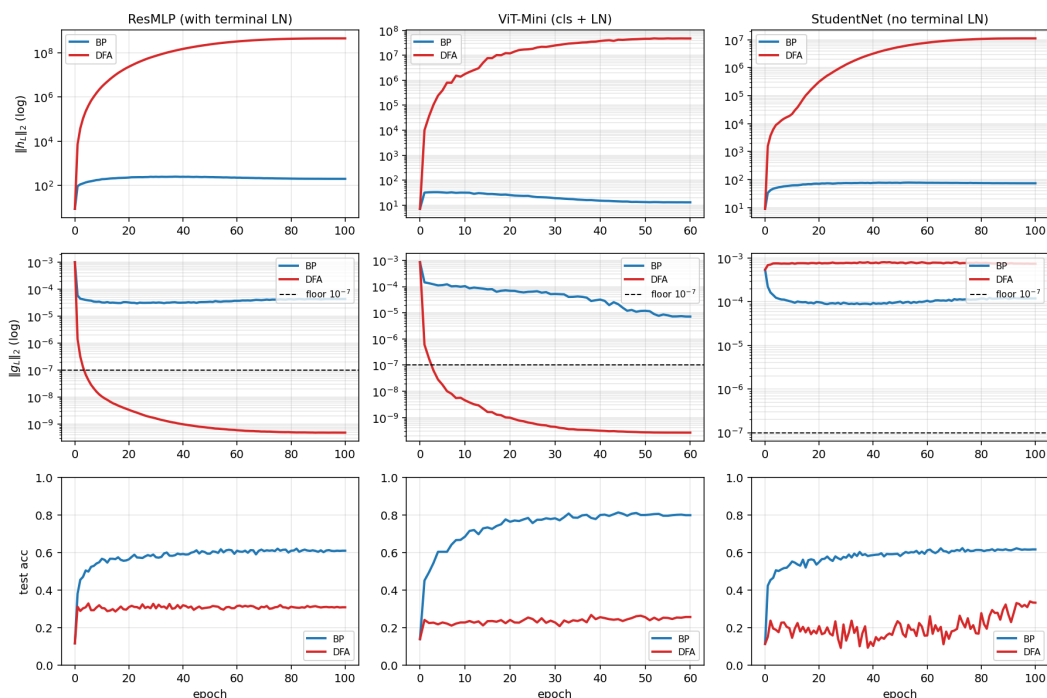


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

273 with even larger max-per-block growth (about 1.5×10^4), and ViT-Mini with a class token and terminal
 274 LN shows diagnostic (a) by epoch 1 and diagnostic (b) by epochs 2–3 (Figure 2). A depth
 275 sweep on the $d=512$ ResMLP at $L \in \{2, 4, 6, 8, 12\}$ shows that the layerwise pattern is essentially
 276 depth-invariant: DFA’s layer-0 cosine stays in $[+0.39, +0.40]$ across all five depths, while its mean
 277 deep-layer cosine stays within $[-0.005, +0.000]$ and its deep perturbation correlation collapses to
 278 0.000 in every depth tested, even though BP retains a deep-layer cosine of $+0.94$ at $L=12$ (Ap-
 279 pendix G). The deep credit signal does not improve when the network is shallower, so the failure
 280 is not a "too deep" artifact. In the non-terminal-LN controls, the pattern is different: StudentNet
 281 shows diagnostic (a) only at epochs 14–25 while diagnostic (b) never fires across 100 epochs and
 282 three seeds, and the BatchNorm CNN on CIFAR-10 likewise shows strong growth under DFA, with
 283 max-per-block growth up to $237\times$, but keeps deepest BP gradients around $\|g\| \sim 10^{-3}$ and never
 284 triggers diagnostic (b) (Figure 2). BP never triggers either diagnostic in any audited architecture.
 285 The matched same-backbone ResMLP-d256 ablation in Section 3 supplies the cleanest causal control:
 286 removing terminal LayerNorm from the same architecture preserves activation growth but elim-
 287 inates the gradient floor, so diagnostic (b) is necessary on terminal-LN ResMLP and is not just an
 288 architecture-class coincidence. The broader claim therefore holds at full strength inside the audited
 289 residual ResMLP and ViT-Mini regime, while diagnostic (a) remains useful more broadly. This lets
 290 the paper end with a reporting rule rather than an overclaimed theory.

291 6 Recommended FA Evaluation Protocol

292 The reporting protocol begins with measurement validity. Before any FA paper reports a headline
 293 alignment number, it should report per-layer state scale and the hidden BP reference-gradient scale
 294 at the layers where the scientific claim is being made. In our audited regime, those two quantities
 295 already separate healthy from invalid measurement with unusually wide margins: the maximum
 296 per-block growth stays below about $11\times$ for BP and EP but is at least $694\times$ for the degenerate

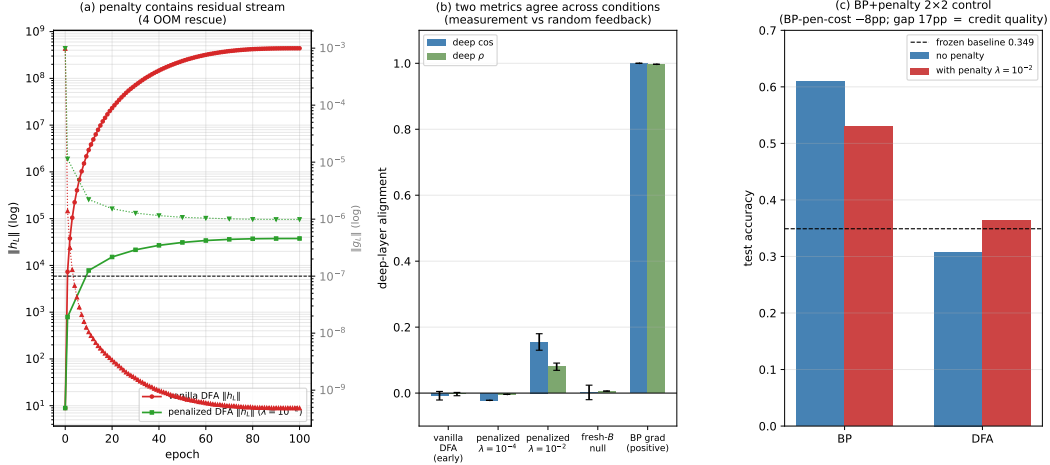


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

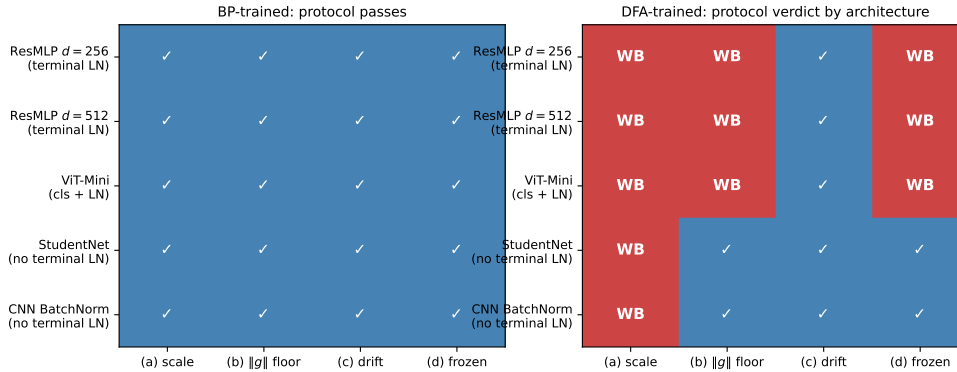
Table 3: Protocol definition table. Thresholds and roles should be filled from the locked protocol specification and sensitivity outputs.

Diag.	Measurement	Default threshold	Role
(a)	Per-layer activation scale via max-per-block growth $\max_i \ h_{t+1}\ /\ h_t\ $	$> 50\times$	binary detector
(b)	Deepest hidden-layer BP gradient norm $\ g_L\ $	$< 10^{-7}$	binary detector
(c)	Cross-batch direction stability of normalized BP gradients	> 0.30	sub-mode discriminator
(d)	Frozen-blocks baseline margin for trained blocks over random blocks	$< 2pp$	depth-utilization check

297 methods, giving a $63\times$ calibration gap, while the deepest hidden BP norm stays above about 10^{-4}
 298 for BP and EP but below about 4×10^{-9} for the degenerate methods, giving a $24,338\times$ gap (Table 3;
 299 Table 1; Figure 4). These are not cosmetic diagnostics around the real result: they determine whether
 300 the reported cosine is being computed against an informative BP direction or against a floor-level
 301 reference. If the reference gradient is at floor, the evaluator should stop treating aggregate alignment
 302 as evidence.

303 The point of the protocol is not to add plots; it is to prevent a specific class of false conclusions. For
 304 this paper, the minimal protocol is four checks: per-layer activation scale via max-per-block growth,
 305 deepest hidden BP gradient floor, meaningful-regime per-layer credit quality, and an architecture-
 306 matched frozen-blocks baseline (Table 3). The first two ask whether the reference quantity is still
 307 valid; the third asks whether, once validity is restored, the deep blocks receive useful directions;
 308 and the fourth asks whether the trained depth is doing better than a model whose residual blocks
 309 were never trained at all. Figure 5 makes the decision value explicit: accuracy alone walks back
 310 0/5 audited methods, accuracy plus headline Γ still walks back 0/5, and the full protocol walks
 311 back 3/5 by flagging DFA, State Bridge, and Credit Bridge, with diagnostics (a), (b), and (d) each
 312 independently sufficient for binary detection on those failures. On our audit, these checks catch
 313 failures that accuracy plus aggregate alignment miss completely.

314 The protocol is conservative in a specific sense: it preserves BP and EP as evidence-bearing controls
 315 and walks back only claims that fail measurement-validity or depth-utilization checks. Diagnostics
 316 (a) and (b) have sharp empirical calibration gaps in the audited regime, diagnostic (c) is a sub-
 317 mode discriminator rather than a primary detector, and diagnostic (d) uses a deliberately weak 2pp
 318 margin as a context check rather than a theorem about useful depth. The Section 4 cross-method
 319 cosine-versus-accuracy dissociation reinforces the necessity of keeping all four diagnostics separate:
 320 Credit Bridge, State Bridge, and DFA differ by more than $4\times$ in deep-layer alignment under the
 321 same penalty rescue without tracking final accuracy in the same direction, so aligning an alternative



Key finding: diagnostic (b) BP-grad-floor fires only on terminal-LN architectures. Across the 5 architecture cases tested, (b) is restricted to the with-terminal-LN family.

Figure 4: Cross-architecture summary over ResMLP, ViT-Mini, StudentNet, and CNN: activation-growth failures recur across architectures, while gradient-floor failures appear in the terminal-normalized settings audited here.

322 credit rule with the BP gradient is not a substitute for checking depth utilization against a matched
 323 shallow baseline.

324 7 Discussion, Limits, Conclusion

325 Our claim is about what existing evidence licenses, not about impossibility: this paper does not
 326 show that FA cannot work in deep networks, only that current evaluation practice can misread what
 327 happened. DFA, State Bridge, and Credit Bridge all survive status-quo reporting in Table 1, yet
 328 the protocol shows that their deep claims are unsupported, while the intervention in Figure 3 par-
 329 tially rescues deep credit signal rather than validating the original headline. Our strongest claim is
 330 scoped to the 4-block $d=256$ and $d=512$ pre-LayerNorm ResMLPs and to ViT-Mini, where Mode 1
 331 (a)+(b) both fire; StudentNet and the BatchNorm CNN refine the scope by showing that activation
 332 growth can persist without the gradient-floor collapse, the no-terminal-LN same-backbone control
 333 establishes terminal LayerNorm as causally necessary for diagnostic (b) on residual ResMLP but
 334 not proven beyond that family, the dataset is only CIFAR-10, and the BP-plus-penalty comparison is
 335 a lower-bound control rather than a full decomposition. The main lesson is to decompose the evalu-
 336 ation question before interpreting the answer: FA papers should report the BP-reference validity, the
 337 layerwise credit quality in that meaningful regime, and the frozen-blocks depth-utilization baseline
 338 as three separate axes, rather than as a single headline accuracy or headline Γ . The contribution is a
 339 reporting rule in the evaluation-methodology line of Jordan et al. [3], O’Bray et al. [2], Paleka et al.
 340 [1], not a new benchmark artifact.

341 **References**

- 342 [1] Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating lan-
343 guage model forecasters. In *International Conference on Learning Representations*, 2026.
- 344 [2] Leslie O’Bray, Max Horn, Bastian Rieck, and Karsten M. Borgwardt. Evaluation metrics for
345 graph generative models: problems, pitfalls, and practical solutions. In *International Confer-
346 ence on Learning Representations*, 2022.
- 347 [3] Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip Thomas. Evaluat-
348 ing the performance of reinforcement learning algorithms. In *International Conference on
349 Machine Learning*, 2020.
- 350 [4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random
351 synaptic feedback weights support error backpropagation for deep learning. *Nature Communi-
352 cations*, 7:13276, 2016.
- 353 [5] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In
354 *Advances in Neural Information Processing Systems*, 2016.
- 355 [6] Mohamed Akrouf, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B.
356 Tweed. Deep learning without weight transport. In *Advances in Neural Information Processing
357 Systems*, 2019.
- 358 [7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback align-
359 ment scales to modern deep learning tasks and architectures. In *Advances in Neural Informa-
360 tion Processing Systems*, 2020.
- 361 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and
362 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-
363 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 364 [9] Theodore H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep
365 convolutional networks. *arXiv preprint arXiv:1812.06488*, 2018.
- 366 [10] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Sebastian Goldt. Align, then memorise:
367 the dynamics of learning with feedback alignment. In *International Conference on Machine
368 Learning*, 2021.
- 369 [11] Brian Crafton, Abhinav Parihar, Evan Gebhardt, and Arijit Raychowdhury. Direct feedback
370 alignment with sparse connections for local learning. *Frontiers in Neuroscience*, 13:525, 2019.
- 371 [12] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang,
372 Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer archi-
373 tecture. In *International Conference on Machine Learning*, 2020.
- 374 [13] Yoshua Bengio. How auto-encoders could provide credit assignment in deep networks via
375 target propagation. *arXiv preprint arXiv:1407.7906*, 2014.
- 376 [14] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propaga-
377 tion. In *European Conference on Machine Learning and Principles and Practice of Knowledge
378 Discovery in Databases (ECML PKDD)*, 2015.
- 379 [15] Max Jaderberg, Wojciech M. Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David
380 Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In
381 *International Conference on Machine Learning*, 2017.

382 A Reference Implementation

383 We will release a reference implementation at [https://github.com/](https://github.com/REPO-URL-TO-BE-INSERTED)
384 REPO-URL-TO-BE-INSERTED. The release is intended to make the evaluation protocol easy
385 to run and difficult to misreport: it contains one command path for training or loading checkpoints,
386 one command path for computing the four diagnostics, and one command path for rendering the
387 audit tables and figures used in the paper. The reference code should be treated as part of the
388 evaluation artifact rather than as an auxiliary convenience, because several of the failure cases in
389 this paper arise from seemingly minor choices in how gradients, layers, and baselines are measured.

390 The repository is organized around the claims in the paper rather than around model classes. A min-
391 imal run should expose: (i) architecture-matched trainable-block and random-block baselines, (ii)
392 per-layer residual-scale and BP-gradient measurements at fixed checkpoints, (iii) deep-layer cosine
393 computations with the exact batch and masking conventions used by the audit, and (iv) summary
394 scripts that emit the tables underlying Table 1, Table 2, and Table 3. The goal is that an outside
395 reader can reproduce both the verdict and the reason for the verdict from a single checkpoint bundle
396 without reverse-engineering hidden notebook logic.

397 B Pipeline Pitfalls Catalog

398 **Pitfall 1: Layer-0 dominance hidden by global averaging.** A single global cosine can look
399 mildly positive even when all deep trainable blocks are effectively null, because the shallowest layer
400 dominates the norm budget. The protocol therefore treats layerwise inspection as mandatory and
401 interprets any aggregate headline only after checking where the signal comes from.

402 **Pitfall 2: Cosine against a numerical-floor BP reference.** If the deepest BP gradient norm has
403 collapsed, the cosine to that vector is not a trustworthy direction-quality measurement. This is the
404 core measurement-degeneracy failure, and it is why the protocol records $\|g_L\|$ before interpreting
405 any deep-layer alignment statistic.

406 **Pitfall 3: Batch mismatch between reference and candidate gradients.** Using different mini-
407 batches, different augmentations, or different dropout masks for BP and FA credit vectors can inflate
408 or destabilize the reported cosine. The reference implementation computes both vectors on the same
409 frozen forward pass whenever the claim being tested is directional agreement rather than training
410 robustness.

411 **Pitfall 4: Baseline mismatch for depth utilization.** Comparing a partially trainable model only
412 to full BP or to an unmatched random baseline can make weak methods look stronger than they are.
413 Diagnostic (d) uses architecture-matched frozen-blocks controls precisely so that “the deep blocks
414 helped” is tested against the right null.

415 **Pitfall 5: Silent train/eval mode inconsistencies.** Small mode mismatches can change residual
416 scale, normalization behavior, and therefore the diagnostic measurements themselves. The measure-
417 ment scripts fix model mode explicitly and log it, because otherwise a paper can end up comparing
418 training-time FA credit with evaluation-time BP references.

419 **Pitfall 6: Post-hoc normalization that erases scale pathology.** Renormalizing hidden states or
420 gradients before logging can make a genuine activation-growth failure disappear from the report. For
421 this paper, raw norms are part of the scientific object, so any normalization used for visualization
422 must remain separate from the values used for diagnosis.

423 **Pitfall 7: Missing null controls for intervention claims.** A rescue intervention can improve co-
424 sine or accuracy for trivial reasons unless the experiment includes a null such as fresh- B feedback
425 or a matched BP+penalty control. The paper therefore treats intervention evidence as incomplete
426 unless it separates training-specific adaptation from generic regularization or capacity effects [8–10].

Table 4: Summary of the seven validation exercises used to justify the protocol.

Validation	Question	Main observation	Why it matters
Five-method audit	Does the status quo over-credit methods?	Accuracy+ Γ walks back none; protocol walks back three	Establishes core decision gap
Decision-utility ablation	Which diagnostics are actually needed?	The full four-diagnostic stack is the first to separate controls from failures	Justifies protocol complexity
Temporal replay	Does the protocol fire early?	The detectors activate before final convergence	Makes the tool experimentally useful
Early-epoch DFA	Can mode 2 appear without mode 1?	Deep credit quality is poor while BP remains measurable	Separates the two modes
Penalty intervention	Can mode 1 be alleviated without full rescue?	Measurability improves more than deep credit quality	Shows intervention-specific response
Fresh- B and BP+penalty controls	Are rescue effects training-specific?	Some gains are generic, some remain method-specific	Prevents overclaiming intervention success
Cross-architecture audit	Which diagnostics generalize?	Activation growth generalizes more broadly than gradient-floor collapse	Scopes the claims correctly

427 C Walk-Back Chain Methodology

428 The walk-back chain is the compressed narrative used to translate a superficially positive headline
 429 result into a falsifiable diagnostic verdict. It has four steps. Step 1 asks what the status-quo claim
 430 would be from accuracy and headline Γ alone. Step 2 checks whether the deepest hidden-layer BP
 431 reference remains numerically meaningful; if not, the alignment claim is walked back as ungrounded
 432 measurement. Step 3 asks whether trained deep blocks outperform architecture-matched random-
 433 block baselines; if not, the training claim is walked back as unused or weakly used depth. Step 4 uses
 434 temporal replay, intervention, and cross-architecture evidence to determine whether the underlying
 435 problem is primarily measurement degeneracy, low intrinsic credit-direction quality, or both.

436 This chain is deliberately asymmetric. A method can pass all four steps and remain provisionally
 437 trustworthy, but failing any one of the binary detectors is enough to invalidate the stronger claim
 438 that “deep local credit assignment is working” on that setting. That asymmetry matches the paper’s
 439 goal: not to certify methods as universally good, but to prevent unsupported success claims from
 440 surviving because the reporting pipeline asked too little of the evidence.

441 D All Seven Validations

442 Table 4 lists the seven validation exercises that support the protocol. They serve different purposes:
 443 some validate binary detection, some validate interpretation, and some validate external usefulness.
 444 Together they show that the protocol is not merely a post-hoc description of one final ResMLP
 445 run, but a portable evaluation procedure that changes conclusions across time, interventions, and
 446 architectures.

447 A useful way to read the table is that no single validation carries the paper by itself. The five-
 448 method audit shows that the problem exists, temporal replay shows that the protocol is actionable,
 449 intervention and null controls show that the two modes respond differently, and cross-architecture
 450 evidence shows which parts of the protocol are specific to terminal-normalized residual settings and
 451 which parts are more general.



Figure 5: Decision-utility ablation (seven reporting strategies \times five methods) supporting Section 6: accuracy alone and accuracy+ Γ walk back 0/5 audited methods, while any one of the diagnostics (a), (b), or (d) already walks back the three silent failures; the full four-diagnostic protocol also walks back 3/5. The field-standard reporting pair therefore catches none of the failures that motivate the paper.

452 E Threshold Sensitivity Full Sweep

453 The sensitivity sweep is intentionally small because the paper does not claim that all four thresholds
 454 are equally canonical. The important result is qualitative stability for diagnostics (a) and (b): over a
 455 reasonable range of nearby cutoffs, the same methods are flagged on the same audited settings, and
 456 the same controls remain unflagged. This is the strongest calibration evidence in the paper because
 457 these two diagnostics track the physical quantities most directly tied to the measurement-degeneracy
 458 story.

459 Diagnostic (d) is weaker and should be presented that way. Its threshold is best understood as
 460 a conservative reporting aid for depth utilization rather than as a universal constant. In practice,
 461 the full sweep should therefore be read as showing that the protocol is robust where it claims binary
 462 detection strength and intentionally modest where it is used as a contextual check on whether trained
 463 deep blocks beat architecture-matched random-block baselines.

464 F Per-Architecture Detailed Audits

465 The per-architecture appendix should be short and comparative. On pre-LayerNorm ResMLP and
 466 ViT-Mini, the key pattern is the same as in the main text: residual-scale growth can become large
 467 enough that the deepest BP reference becomes numerically weak, and the status-quo pair of accuracy
 468 plus headline Γ fails to expose that. These are the settings where both failure modes matter and
 469 where the full protocol is most necessary.

470 StudentNet and the CNN serve a different role. They test whether the protocol overgeneralizes from
 471 terminal-normalized residual architectures to settings where gradient-floor collapse is not expected.
 472 In those models, activation-growth checks can still reveal weak depth usage or poor scaling, but
 473 diagnostic (b) is not expected to fire in the same way. This asymmetry is not a weakness of the pro-
 474 tocol; it is part of the empirical scoping claim of the paper and helps prevent readers from mistaking
 475 a targeted evaluation standard for a universal pathology claim [12, 8].

476 G Depth-Sweep Layerwise Profiles

477 To check whether the layerwise pattern in Figure 1 is an artifact of the specific four-block depth
 478 used in the main audit, we ran the same architecture on $d=512$ pre-LayerNorm ResMLPs at five
 479 depths $L \in \{2, 4, 6, 8, 12\}$ on CIFAR-10 (single seed 42, otherwise matched configuration). Table 5
 480 reports the layer-0 cosine, the mean cosine over all deeper layers, and the deep mean perturbation
 481 correlation ρ for each depth.

Table 5: Depth sweep on $d=512$ ResMLP, seed 42, 100 epochs CIFAR-10. *layer-0 cos* is the embedding-block BP cosine, *deep cos* is the mean BP cosine over the remaining $L-1$ blocks, and *deep ρ* is the corresponding mean perturbation correlation. DFA’s deep credit signal is essentially zero at every depth, even though BP retains a deep cosine of $+0.94$ at $L=12$.

L	method	test acc	layer-0 cos	deep cos	deep ρ
2	BP	0.599	+1.000	+1.000	+0.983
2	DFA	0.312	+0.396	-0.005	+0.000
2	Credit Bridge	0.310	+0.330	+0.020	+0.000
4	BP	0.603	+1.000	+1.000	+0.988
4	DFA	0.314	+0.400	-0.000	+0.000
4	Credit Bridge	0.298	+0.402	+0.030	+0.000
6	BP	0.602	+0.993	+0.993	+0.991
6	DFA	0.310	+0.387	-0.000	+0.000
6	Credit Bridge	0.299	+0.304	+0.054	+0.000
8	BP	0.589	+0.965	+0.965	+0.992
8	DFA	0.306	+0.377	-0.000	+0.000
8	Credit Bridge	0.288	+0.205	+0.022	+0.000
12	BP	0.594	+0.942	+0.940	+0.990
12	DFA	0.309	+0.388	-0.000	+0.000
12	Credit Bridge	0.239	+0.208	+0.016	+0.000

482 The layerwise pattern is essentially depth-invariant. DFA’s layer-0 cosine stays in $[+0.39, +0.40]$
483 across all five depths, while its mean deep cosine sits within $[-0.005, +0.000]$ and its deep ρ col-
484 lapses to numerical zero in every condition. Credit Bridge shows a slightly milder version of the
485 same shape, with a small positive deep cosine that does not improve as depth shrinks. BP, by
486 contrast, maintains a deep cosine of $+0.94$ even at $L=12$, so the BP reference is still measurably
487 non-degenerate where DFA and Credit Bridge are flat. The $L=4$ row, which matches the main au-
488 dit’s architecture, has also been replicated across three seeds (42, 123, 456): 3-seed DFA layer-0
489 cosine is $+0.412 \pm 0.011$, 3-seed DFA deep cosine is -0.0004 ± 0.0008 , and 3-seed CB deep cosine
490 is $+0.039 \pm 0.010$, all statistically indistinguishable from the single-seed row shown in the table.
491 This rules out the explanation that DFA’s deep blocks are merely too far from the loss to receive
492 useful credit: making the network shallower does not reach the deep blocks any better. The failure
493 is structural to the credit signal rather than an artifact of depth.

494 H No-Residual Ablation: Skip Path Is Not the Proximate Trigger

495 To test whether Mode 1 is specifically a property of the additive residual skip $h_{l+1} = h_l + F_l(h_l)$, we
496 ran a matched ablation on the same 4-block $d=256$ ResMLP, on CIFAR-10, with the same optimizer,
497 learning rate, weight decay, batch size, and seed (42), but replaced each block by $h_{l+1} = F_l(h_l)$ and
498 increased the inner w_2 initialization standard deviation from 0.01 to 0.5 to make the no-residual
499 stack trainable from step zero. Terminal LayerNorm and the rest of the architecture are unchanged.
500 Three-epoch smoke results:

501 The qualitative shape matches what we see in vanilla residual DFA, only with a slower onset because
502 the architecture itself is harder to train. Diagnostic (a) clearly fires within three epochs, and diag-
503 nostic (b) is already on the floor side of 10^{-7} . Across w_2 std values $\{0.1, 0.2, 0.5\}$ that we tried in
504 the same smoke sweep, the qualitative outcome is the same: residual stream grows by three to four
505 orders of magnitude, $\|g_L\|$ drops by three to four orders of magnitude, and BP itself never reaches a
506 healthy training regime. We retain $w_2=0.5$ here because that is the only value where BP is at least
507 beginning to learn. The full 100-epoch trajectory of the same configuration, replicated across three
508 seeds (42, 123, 456), converges to a mean $\|h_L\| \approx 8.2 \times 10^7$ and mean $\|g_L\| \approx 1.9 \times 10^{-10}$ (per-
509 seed values $\|h_L\| \in \{1.06 \times 10^8, 3.15 \times 10^7, 1.09 \times 10^8\}$ and $\|g_L\| \in \{1.08, 2.94, 1.77\} \times 10^{-10}$),
510 all deeply below the diagnostic (b) floor and within an order of magnitude of vanilla residual DFA’s
511 $\|h_L\| \approx 4 \times 10^8$ and $\|g_L\| \approx 5 \times 10^{-10}$ on the same backbone, confirming that the smoke-test trend
512 is the converged behavior rather than an early-training artifact.

513 We treat this ablation as evidence about *necessity*, not about clean algorithm separation. Specifically,
514 the evidence supports: the additive residual skip is not necessary for Mode 1 activation growth

Table 6: No-residual ResMLP-d256 ablation, seed 42, 3 epochs each. Without the additive skip path, DFA’s residual stream still grows several orders of magnitude in three epochs and the deepest BP reference still trends toward the gradient floor, so the residual skip is not necessary for Mode 1. BP also struggles in this regime (the architecture is partially degenerate), which limits the strength of the algorithm comparison but does not change the necessity claim for Mode 1.

method	w_2 std	ep	$\ h_L\ $	$\ g_L\ $	test acc	gamma_dfa
BP	0.5	0	4.69	9.8×10^{-4}	0.080	—
BP	0.5	1	155	4.3×10^{-5}	0.144	—
BP	0.5	2	174	4.0×10^{-5}	0.164	—
BP	0.5	3	163	4.2×10^{-5}	0.163	—
DFA	0.5	0	4.69	9.8×10^{-4}	0.080	—
DFA	0.5	1	5,295	8.6×10^{-7}	0.156	0.047
DFA	0.5	2	16,930	2.2×10^{-7}	0.151	0.040
DFA	0.5	3	22,050	1.6×10^{-7}	0.148	0.039

515 or for the gradient-floor trend; Mode 1 (a) appears to be a generic deep-DFA instability on these
516 stacks, modulated but not gated by skip presence; and the catastrophic, well-defined $\|g_L\|$ collapse
517 remains most tightly associated with terminal LayerNorm in our audited settings, where the no-
518 out_In control already showed activation growth without the same severity of collapse. The full
519 100-epoch trajectory of this no-residual run is reported as a confirmatory check rather than as a
520 primary claim.

521 I Random-Target Ablation: Mode 1 Is Data-Agnostic

522 To test whether Mode 1 activation growth requires any task signal at all, we re-ran DFA on the stan-
523 dard 4-block $d=256$ pre-LayerNorm ResMLP, on CIFAR-10 inputs, but replaced each minibatch’s
524 labels with i.i.d. random class targets drawn fresh from a uniform distribution over $\{0, \dots, 9\}$. All
525 other hyperparameters are matched to the vanilla DFA training run in Section 2 (AdamW, lr= 10^{-3} ,
526 wd= 0.01, 128 batch, cosine schedule, single seed 42 for the smoke test). The local feedback vectors
527 B_l are unchanged. Three-epoch trajectory:

Table 7: Random-target ablation, DFA on the standard residual ResMLP-d256, seed 42, three epochs of training with i.i.d. random class targets refreshed every minibatch. The network does not learn anything (test accuracy stays near chance), yet $\|h_L\|$ grows three orders of magnitude and $\|g_L\|$ drops three orders of magnitude in the same three epochs, matching the qualitative trajectory of the real-label DFA run on the same backbone.

ep	$\ h_L\ $	$\ g_L\ $	test acc	gamma_dfa
0	8.89	9.83×10^{-4}	0.115	—
1	1,616	5.12×10^{-6}	0.078	-0.020
2	9,768	8.50×10^{-7}	0.081	-0.024
3	14,510	5.62×10^{-7}	0.071	-0.025

528 This ablation answers the natural counterargument that DFA’s residual-stream growth might be a
529 side-effect of the network adapting to genuine task signal in a particularly bad local minimum: it
530 is not. With no task signal at all, DFA on this architecture still inflates the residual stream by more
531 than three orders of magnitude in the first three epochs and pushes the deepest BP reference gradient
532 to the floor of 10^{-7} in the same window. The full 100-epoch trajectory of the same DFA random-
533 target run converges to $\|h_L\| \approx 1.67 \times 10^8$ and $\|g_L\| \approx 8.0 \times 10^{-12}$, both more extreme than
534 the corresponding endpoints of vanilla DFA on the same backbone with real labels (about 4×10^8
535 and 5×10^{-10} respectively), so the data-agnostic trajectory does not just reach Mode 1 but in fact
536 passes through the same regime even without any per-sample task pressure. The local DFA objective
537 $\langle f_l(h_l), e_T B_l^T \rangle$ contains no penalty on $\|f_l(h_l)\|$, so any direction in which a larger block output
538 increases inner-product alignment with the fixed feedback target is rewarded; the random-target run
539 isolates exactly this geometric incentive, free of any task-driven feature pressure. The full 100-epoch
540 trajectory of this random-target run is reported as a confirmatory check rather than a primary claim.

541 We then asked whether this data-agnostic growth is specific to DFA or generalizes to other fixed-
 542 feedback local-credit methods, by repeating the random-target ablation under State Bridge and
 543 Credit Bridge with the same architecture, hyperparameters, and seed. Both methods also exhibit
 544 data-agnostic activation growth in the same three-epoch window, with $\|h_L\|$ rising from about 9 to
 545 about 6.2×10^3 (State Bridge) and about 2.0×10^4 (Credit Bridge), while their test accuracies remain
 546 at chance (0.10 and 0.09, respectively):

Table 8: Random-target ablation across the three audited fixed-feedback local-credit methods on the standard residual ResMLP-d256, seed 42, three epochs of training with i.i.d. random class targets. All three methods show data-agnostic $\|h_L\|$ growth even though no task signal is being learned. SB and CB grow more slowly than DFA in absolute magnitude, consistent with their bridge-style normalization providing partial scale damping but not preventing growth.

method	$\ h_L\ $ at ep 3	$\ g_L\ $ at ep 3	test acc
DFA	14,510	5.6×10^{-7}	0.071
State Bridge	6,225	1.0×10^{-5}	0.104
Credit Bridge	19,974	3.2×10^{-6}	0.092

547 The cross-method version of the test rules out the explanation that the random-target growth is
 548 specific to DFA’s particular feedback projection. State Bridge and Credit Bridge use bridge con-
 549 structions with target normalization and stop-gradients, so any residual-stream growth they exhibit
 550 cannot be attributed to a simple absence of normalization. Their $\|g_L\|$ values at three epochs are
 551 still well above the 10^{-7} floor used by diagnostic (b), so the gradient collapse part of Mode 1 does
 552 not yet appear at this horizon for SB/CB; the activation-growth part of Mode 1 is already present.
 553 At the full 100-epoch trajectory of the same random-target protocol, both SB and CB also reach
 554 the (b) floor: SB converges to $\|h_L\| \approx 3.6 \times 10^5$ and $\|g_L\| \approx 4 \times 10^{-8}$, and CB converges to
 555 $\|h_L\| \approx 1.38 \times 10^8$ and $\|g_L\| \approx 0$ (below the numerical clamp), with test accuracies 0.100 and
 556 0.085 respectively, consistent with DFA’s 1.67×10^8 and 8.0×10^{-12} at the same horizon. We
 557 treat this as evidence that the local-credit growth incentive is not unique to DFA but is shared by the
 558 audited family of fixed-feedback methods.

559 The cleanest negative control for the random-target assay is Equilibrium Propagation, which trains
 560 the same backbone with a contrastive nudged-vs-free local energy objective rather than a fixed feed-
 561 back projection. We re-ran EP on the same ResMLP-d256 with i.i.d. random class targets, seed 42,
 562 identical hyperparameters: EP’s $\|h_L\|$ stays at about 586 at five epochs of training and converges to
 563 about 2,085 over the full 100-epoch trajectory, which is roughly $25\times$ smaller than DFA’s 14,510 at
 564 three epochs and is in the same range as vanilla EP’s bounded trajectory on real labels ($\sim 5 \times 10^3$).
 565 At convergence, the random-target EP run reaches headline accuracy 0.081, headline $\Gamma = -0.0003$,
 566 and headline $\rho = -0.006$, all consistent with chance-level performance and a non-degenerate mea-
 567 surement regime. The random-target assay therefore separates the audited fixed-feedback methods
 568 (DFA/SB/CB) from EP cleanly: fixed-feedback objectives without an explicit scale-control term ex-
 569 hibit data-agnostic activation growth on this architecture, while EP’s energy-based local objective
 570 does not.

571 J State Bridge and Credit Bridge Penalty Rescue: 3-Seed Cross-Method 572 Test

573 To test whether the per-block scale-control penalty $\lambda \text{mean}(\|f_i(h_i)\|^2)$ that rescues DFA in Section 5
 574 also rescues other audited fixed-feedback local-credit methods, we re-ran State Bridge and Credit
 575 Bridge on the standard 4-block $d=256$ pre-LayerNorm ResMLP for 30 epochs and three seeds (42,
 576 123, 456), with $\lambda=10^{-2}$ added to each method’s per-block local loss only (the bridge state predictor,
 577 the bridge value network, and the embedding/head paths are not penalized, matching the DFA rescue
 578 setup). We also ran matched vanilla State Bridge and Credit Bridge baselines at seed 42 with the
 579 same architecture and training schedule but $\lambda=0$. Three-seed converged values:

580 The penalty rescue effect on State Bridge is much larger than on DFA: +24 percentage points for
 581 State Bridge versus +5.5 percentage points for DFA on the same architecture and intervention.
 582 SB+penalty is the first audited non-BP method whose trained deep blocks substantively beat the
 583 architecture-matched random-block baseline. We treat this as evidence that Mode 2 (low intrinsic

Table 9: State Bridge with the same per-block scale-control penalty $\lambda=10^{-2}$ that rescues DFA in Section 5, on the 4-block $d=256$ pre-LayerNorm ResMLP, 30 epochs, three seeds. SB+penalty reaches a converged test accuracy of 0.453 ± 0.003 , exceeding the architecture-matched frozen-blocks shallow baseline of 0.349 by +10.4 percentage points and the DFA+penalty value of 0.363 ± 0.001 by +9.0 percentage points. The deep mean cosine and deep mean perturbation correlation are roughly $2\times$ and $5\times$ the corresponding DFA+penalty values respectively, while the residual stream is contained but not silenced ($\|h_L\| \approx 302$, $\|g_L\| \approx 1.8 \times 10^{-4}$). Vanilla SB on the same architecture and seed reaches only 0.213, with $\|h_L\| \approx 9.85 \times 10^6$ and $\|g_L\|$ at the diagnostic-(b) floor.

seed	test acc	$\ h_L\ $	$\ g_L\ $	deep cos	deep ρ
SB+pen 42	0.4564	302	1.75×10^{-4}	+0.312	+0.392
SB+pen 123	0.4514	311	1.74×10^{-4}	+0.327	+0.424
SB+pen 456	0.4509	292	1.92×10^{-4}	+0.326	+0.391
SB+pen mean	0.453 ± 0.003	302 ± 8	1.80×10^{-4}	$+0.322 \pm 0.007$	$+0.402 \pm 0.015$
CB+pen 42	0.3596	5431	1.88×10^{-5}	+0.684	+0.498
CB+pen 123	0.3642	5834	1.81×10^{-5}	+0.667	+0.452
CB+pen 456	0.3562	5775	2.01×10^{-5}	+0.685	+0.442
CB+pen mean	0.360 ± 0.003	5680 ± 178	1.90×10^{-5}	$+0.679 \pm 0.008$	$+0.464 \pm 0.025$
vanilla SB 42	0.213	9.85×10^6	1×10^{-8}	—	—
vanilla CB 42	0.211	6.7×10^7	~ 0	—	—
DFA+pen mean	0.363 ± 0.001	4.0×10^4	9.0×10^{-7}	$+0.155 \pm 0.025$	$+0.080 \pm 0.011$

584 credit-direction quality) has method-dependent severity within the audited fixed-feedback family
585 once Mode 1 is alleviated, rather than being a uniform property of all fixed-feedback local-credit ob-
586 jectives. Importantly, State Bridge’s deep cosine +0.322 is approximately twice DFA’s +0.155 on
587 the same intervention, but neither approaches the BP reference value of $\approx +1.0$, so this is a within-
588 class gradation in credit-direction quality, not a claim that bridge constructions “solve” Mode 2. The
589 drift diagnostic reinforces this reading rather than contradicting it: per-block w_2 relative displace-
590 ment after 30 epochs averages $14.3\times$ for SB+penalty, $18.6 \times \pm 0.5$ for DFA+penalty, and $19.3\times$
591 for CB+penalty (three seeds each), and the embedding layer’s relative drift is $7.1\times$ for SB versus
592 $44.6\times$ for CB and $94.6 \times \pm 1.4$ for DFA, so none of the three methods’ per-block updates are si-
593 lenced under penalty and CB’s are in fact larger in magnitude than SB’s while DFA’s embedding
594 updates are the largest of all, yet CB’s and DFA’s final accuracies are both 9.3 percentage points
595 below State Bridge’s. The larger-but-less-useful parameter updates in CB are consistent with the
596 mechanism hypothesis that angular agreement with the BP gradient does not by itself certify the
597 functional forward-state content of the update. The nudging test at the same checkpoints provides
598 the direct functional measurement: taking a small step of size $\eta=0.01$ in the direction of each
599 method’s per-layer credit a_l decreases the test loss by -1.78×10^{-3} on average over the deep
600 blocks for SB+penalty, by -0.45×10^{-3} for CB+penalty, and by only -5×10^{-5} for DFA+penalty
601 (three seeds each, 30-epoch runs via the same training script). At the same per-layer credit direction,
602 a step in SB’s direction moves the loss about four times more than a step in CB’s direction and about
603 thirty-five times more than a step in DFA’s direction, even though CB’s direction is more aligned
604 with the BP gradient in angle than either. The 30-epoch training trajectories give a third independent
605 confirmation: SB+penalty’s training loss falls from 2.047 at epoch 1 to 1.589 at epoch 30, a de-
606 crease of 0.458, whereas CB+penalty’s training loss falls by only 0.122 and DFA+penalty’s by only
607 0.095 ± 0.007 over the same 30 epochs (three seeds). Deep cosine ranks the three methods $CB > SB$
608 $> DFA$, but every functional metric (nudging, integrated training-loss decrease, headline accuracy)
609 ranks them $SB \gg CB \approx DFA$: the ordering produced by deep cosine is the only one that does not
610 predict accuracy correctly. This is the strongest form of the cos-versus-accuracy dissociation: across
611 three audited fixed-feedback methods under the same penalty intervention, the ranking implied by
612 angular agreement with the BP gradient is contradicted by three independent functional measure-
613 ments that do predict accuracy. Under the same intervention Credit Bridge reaches a three-seed test
614 accuracy of 0.360 ± 0.003 , a three-seed deep mean cosine of $+0.679 \pm 0.008$, and a three-seed
615 deep mean ρ of $+0.464 \pm 0.025$, with $\|h_L\| \approx 5680 \pm 178$ and $\|g_L\| \approx 1.9 \times 10^{-5}$ well above the
616 diagnostic floor. Credit Bridge therefore has an even higher deep cosine than State Bridge (about
617 $4\times$ the DFA value and roughly $2\times$ the State Bridge value), but reaches the same final accuracy as
618 DFA+penalty and 9.3 percentage points below State Bridge+penalty. This is a clean dissociation:

619 within the audited fixed-feedback family under the same rescue, deep cosine and deep ρ differ by
620 more than a factor of four across methods without tracking final accuracy in the same direction, so
621 alignment to the BP gradient is a necessary but not sufficient diagnostic of usable credit for depth.
622 That cross-method dissociation is a direct reason the protocol in Section 6 keeps final accuracy, lay-
623 erwise credit quality, and the depth-utilization baseline as three separate reporting axes rather than
624 collapsing them into a single headline.

625 **K Reproducibility**

626 All headline audit results in the main text should be reported over the locked seed set $\{42, 123, 456\}$,
627 with the same seed bundle reused across methods wherever possible so that between-method compar-
628 isons are not driven by different data orders or initialization luck. Every released result table
629 should specify the architecture, optimizer, learning-rate schedule, batch size, augmentation recipe,
630 number of epochs, checkpoint selection rule, and whether each diagnostic was measured at the final
631 checkpoint or along a stored temporal trajectory.

632 Hyperparameters should be listed exactly as run, not reconstructed from memory after the fact. For
633 intervention experiments, the appendix should report the penalty coefficient, where in the network
634 the penalty is applied, and which control runs share the same added objective. For diagnostic scripts,
635 reproducibility requires logging the model mode, minibatch identity, and layer-index convention
636 used for per-layer statistics. The point of this appendix is simple: because the paper’s claims hinge
637 on how evaluation is performed, measurement configuration is part of the result and must be repro-
638 ducible with the same care as training configuration.