
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 summa-
2 rized 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 protocol
11 walks back the three degenerate methods and passes the two trustworthy controls.
12 Our contribution is an evaluation methodology paper for the NeurIPS 2026 Evaluations
13 & Datasets track: we provide the protocol, the calibration logic for its thresh-
14 olds, a reference implementation, a five-method audit, and validation through tem-
15 poral replay, cross-architecture checks, intervention-based disambiguation, and a
16 documented catalog of pipeline pitfalls, in the spirit of critical evaluation analyses
17 such as Jordan et al. [3], O’Bray et al. [2], Paleka et al. [1].

18 1 Introduction

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

32 Our audit shows that modern residual vision models can make these two quantities look informa-
33 tive while failing to answer the question they are taken to answer. Figure 1 shows the first failure
34 mode, which we call *Mode 1: measurement degeneracy*, where residual-stream growth drives the
35 deepest hidden state to about $\|h_L\| \sim 10^8$ under DFA/SB/CB while the corresponding BP reference

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 collapses to $\|g_L\| \sim 5 \times 10^{-10}$, so the deep-layer cosine is measured against a clamp-dominated
 37 floor rather than a meaningful target direction. The same figure also shows the second failure mode,
 38 *Mode 2: low intrinsic credit-direction quality*, because even after comparing against the stronger
 39 frozen-blocks baseline (0.349 ± 0.002) and looking layer-by-layer, DFA’s deep blocks remain essen-
 40 tially null while only layer 0 is visibly positive. To test whether this is only a measurement problem,
 41 the intervention results show a dissociation: with a residual penalty $\lambda \|f_l(h_l)\|^2$, the deepest state
 42 scale falls toward 4×10^4 , the reference gradient rises toward 10^{-6} , and deep cosine can improve
 43 to about $+0.16$, yet at $\lambda=10^{-4}$ Mode 1 is alleviated while deep cosine still stays near zero, and at
 44 vanilla DFA epoch 1 the reference is already meaningful at about 6×10^{-7} but the deep cosine is still
 45 -0.008 ± 0.013 across three seeds. The failure is not unitary: one mode breaks the measurement,
 46 and the other survives even when the measurement is still meaningful.

47 Accordingly, this paper does not introduce a new FA variant or a new benchmark. Instead, Table 1
 48 and Figure 1 use a standard five-method CIFAR-10 audit to show that status-quo reporting would
 49 treat BP, EP, DFA, State Bridge, and Credit Bridge as the same kind of evidence-bearing object
 50 even though only BP and EP remain trustworthy under matched diagnostic checks. This makes the
 51 contribution methodological in the sense of Jordan et al. [3], O’Bray et al. [2], and Paleka et al. [1]:
 52 the central question is not whether one more FA variant can post a headline number, but whether the
 53 reporting pipeline distinguishes meaningful credit-direction evidence from numerical-floor artifacts
 54 and from shallow-only learning. The protocol therefore starts from per-layer diagnostics and a
 55 frozen-blocks baseline before reading any aggregate cosine or final accuracy as evidence about deep
 56 credit assignment. We first show the walk-back on a standard audit, then isolate the two failure
 57 modes, and finally state the reporting protocol that future FA papers should satisfy.

58 2 Audit: Standard Reporting Walks Back Nothing

59 We begin with the smallest setting in which all methods can be compared head-to-head under iden-
 60 tical architecture, optimizer family, and data. Table 1 fixes that canonical audit to a 4-block pre-
 61 LayerNorm ResMLP with width $d=256$ on CIFAR-10, trained for 100 epochs with AdamW (learn-
 62 ing rate 10^{-3} , weight decay 0.01), a cosine schedule, and three seeds (42, 123, 456). Within that
 63 single setting, BP, EP, DFA, State Bridge, and Credit Bridge can be read against the same architec-
 64 ture and the same training budget, while Figure 1 summarizes the corresponding per-block growth,
 65 deepest-layer BP reference norm, cross-batch stability, and frozen-baseline comparison. This is the
 66 table a reader would normally use to decide whether the methods trained the deep network.

67 By the field’s usual criteria, the non-BP methods appear to train to nontrivial accuracy and report
 68 nonzero alignment. In Table 1, DFA reaches 0.306 ± 0.006 test accuracy with headline $\Gamma=0.10$,
 69 State Bridge reaches 0.205 ± 0.032 with $\Gamma=0.005$, and Credit Bridge reaches 0.289 ± 0.026 with
 70 $\Gamma=0.07$; none of these rows looks like an obvious invalidation if one is reading the usual pair of final
 71 accuracy and aggregate alignment in the style of prior FA reporting [4–7]. Even the absolute scale
 72 does not itself force a walk-back, because all three methods are plainly above chance and all three
 73 report positive headline alignment rather than a visibly broken or undefined quantity. That reading
 74 is exactly what the rest of the paper overturns.

75 Low accuracy by itself is not the pathology. EP is the key internal comparison in Table 1 and
 76 Figure 1: it achieves only 0.316 ± 0.030 accuracy and a very small headline $\Gamma=0.008$, yet its per-
 77 block growth is only $11.6\times$, its deepest BP reference norm remains around 1.3×10^{-4} rather than
 78 collapsing to the numerical floor, and its cross-batch direction-stability score is 0.02 rather than the
 79 much higher drift-dominated values seen for DFA-family methods. At the same time, EP is not a

5-method audit on 4-block $d=256$ ResMLP CIFAR-10 (3-seed mean \pm std)

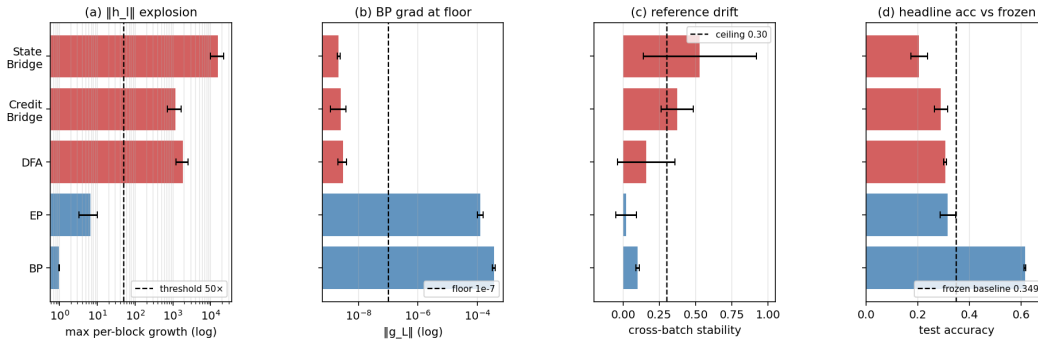


Figure 1: Five-method audit on the 4-block $d=256$ pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

80 positive result for depth usage in the stronger sense, because its trainable-model accuracy is still
 81 3.3 percentage points below the frozen-blocks baseline of 0.349 ± 0.002 . The distinction matters
 82 because it separates underperformance from invalid evaluation.

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

93 3 Failure Mode 1: Measurement Degeneracy

94 Mode 1 has two parts. The activation-growth part (a) is a scale pathology of fixed-feedback local-
 95 credit objectives without an effective scale-control term: for block l , DFA, State Bridge, and Credit
 96 Bridge update f_l by reducing a local loss of the form $-\langle f_l(h_l), B_l^\top e_T \rangle$, which contains no penalty
 97 on $\|f_l(h_l)\|$, so any direction in which a larger block output improves inner-product alignment
 98 with the fixed feedback target is rewarded; in a pre-LN residual stack, larger block outputs di-
 99 rectly increase residual-stream scale, and terminal LayerNorm at the output removes task-loss sen-
 100 sitivity to that scale, so the architecture supplies no global restraint on the local growth incentive.
 101 The gradient-floor part (b) follows from the LayerNorm Jacobian: in terminal-LN architectures
 102 $\partial \text{LN}(h)/\partial h \propto 1/\|h\|$ in expectation, so the same residual-stream inflation is accompanied by col-
 103 lapse of the hidden-layer BP reference norm. Empirically, on the audited 4-block pre-LayerNorm
 104 ResMLP ($d=256$, CIFAR-10, 100 epochs, 3 seeds), DFA training drives $\|h_L\|$ from about 9 at ini-
 105 tialization to about 4×10^8 by epoch 100 and $\|g_L\|$ from about 9.8×10^{-4} to about 5×10^{-10} ,
 106 while the reported deep cosine remains defined only because `F.cosine_similarity` clamps the
 107 denominator at $\epsilon=10^{-8}$ (Table 1; Figure 1). At that endpoint the reference norm is about $20\times$ below
 108 the clamp, so the quantity being reported is effectively $(a \cdot b)/(\|a\| \max(\|b\|, 10^{-8}))$ rather than a
 109 comparison to an meaningful BP direction.

110 We tested this mechanism story against four natural alternative attributions, all of which it survives.
 111 *Not residual-skip-driven*: on the same ResMLP-d256 with terminal LN kept and the additive skip
 112 removed ($h_{l+1}=F_l(h_l)$), DFA still inflates $\|h_L\|$ from ~ 5 to $\sim 2.2 \times 10^4$ in three epochs and con-
 113 verges to $\|h_L\| \approx 1.06 \times 10^8$ and $\|g_L\| \approx 1.09 \times 10^{-10}$ at 100 epochs, both already at the diagnostic
 114 floor (Appendix H). *Not task-signal-driven*: replacing labels by i.i.d. random class targets refreshed
 115 every minibatch on the same backbone, DFA still reaches $\|h_L\| \approx 1.67 \times 10^8$ and $\|g_L\| \approx 8 \times 10^{-12}$ at
 116 100 epochs while accuracy stays at chance (Appendix I). *Not DFA-specific*: the same random-target

117 ablation also drives $\|h_L\|$ from 9 to 6.2×10^3 for State Bridge and 2.0×10^4 for Credit Bridge in three
 118 epochs, again at chance accuracy, so all three audited fixed-feedback methods exhibit data-agnostic
 119 activation growth (Appendix I). *Not shared by EP*: under the same random-target protocol, EP keeps
 120 $\|h_L\| \approx 586$ at five epochs of training, $25\times$ smaller than DFA’s three-epoch value on the same archi-
 121 tecture, consistent with EP’s bounded behavior on real labels and confirming that the random-target
 122 assay separates the explosion-prone fixed-feedback class from EP’s energy-based local objective.

123 The matched same-backbone causal control for diagnostic (b) is removing terminal LayerNorm. On
 124 the same ResMLP-d256 with the residual skip intact, 100 epochs of DFA, three seeds, the residual
 125 stream still inflates to $\|h_L\| \approx 1.21 \times 10^7$, but the deepest hidden-layer BP gradient remains at
 126 $\|g_L\| \approx 7.2 \times 10^{-4}$ (four orders of magnitude above the diagnostic (b) floor), and the final test
 127 accuracy is 0.327 ± 0.012 , statistically indistinguishable from vanilla DFA’s 0.306 ± 0.006 on the
 128 same backbone with terminal LayerNorm intact. Removing terminal LayerNorm therefore preserves
 129 Mode 1 (a) but cleanly eliminates Mode 1 (b) on the same architecture, while leaving final task
 130 accuracy essentially unchanged. Combined with the broader cross-architecture pattern (StudentNet
 131 and the BatchNorm CNN, which lack terminal LayerNorm, never trigger diagnostic (b); ViT-Mini
 132 with a terminal LN does, by epochs 2–3 (Figure 2)), terminal LayerNorm is necessary for Mode 1 (b)
 133 in the audited residual ResMLP and ViT-Mini setting. The collapse is also not a late-epoch curiosity:
 134 $\|g_L\|$ drops from 9.8×10^{-4} at epoch 0 to 6.7×10^{-8} by epoch 4 in the temporal replay across three
 135 seeds, so the protocol fires within the first 11 epochs of a 100-epoch run and is actionable as an
 136 early-stop criterion rather than a post hoc explanation. Once measurement degeneracy is identified,
 137 the next question is whether poor deep credit remains even before collapse.

138 4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

139 The second failure mode appears even in the meaningful-measurement regime. At the earliest vanilla
 140 DFA checkpoints on ResMLP, the hidden backpropagated gradient at the first deep block remains
 141 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
 142 seeds, all above the 10^{-7} threshold used to distinguish measurable from collapsed gradients. Yet the
 143 corresponding deep-layer cosine values are already essentially null: across layers 1–4, all seed-level
 144 measurements at epoch 1 lie in $[-0.04, +0.02]$, with a three-seed mean of -0.008 ± 0.013 , and by
 145 epoch 2 the deep mean is still only -0.018 ± 0.018 (Table 2). This is the observational pattern pre-
 146 dicted by low credit-direction quality rather than mere disappearance of signal: the gradient is still
 147 present enough to measure, but the directions delivered to the deep network carry little agreement
 148 with backpropagation, consistent with prior concerns that alternative feedback rules can fail by sup-
 149 plying poor credit assignments even before full collapse [8, 9, 11, 10]. This rules out the simplest
 150 objection that the deep-layer null result is merely a byproduct of collapse.

151 A second metric with different numerical failure modes tells the same story. Cosine measures di-
 152 rectional agreement with the BP gradient, whereas perturbation correlation ρ measures whether the
 153 proposed update predicts the correct sign and relative magnitude of loss change under actual per-
 154 turbations; their failure modes are therefore different, especially with respect to normalization and
 155 small-denominator effects. In our controls, ρ behaves as expected, with a Taylor-ceiling positive
 156 control near $+0.997$ and a random-vector negative control near $+0.006$ (Figure 3, Table 2). On
 157 vanilla DFA, deep ρ is likewise null: for the early checkpoints where the gradients remain measur-
 158 able, the deep average is -0.003 ± 0.005 across seeds and epochs, and in a floor-level checkpoint it is
 159 $+0.002$, again indistinguishable from noise. The agreement between cosine and ρ therefore rules out
 160 the interpretation that the null deep result is an artifact of cosine’s ε -clamp or vector normalization.
 161 The deep blocks are not just hard to measure; they are receiving weakly useful directions.

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

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 m
Vanilla DFA, converged	$\overline{\text{cos}}_{deep} = -0.022, \overline{\rho}_{deep} = +0.002$	degenerate ($\ g\ \sim 10^{-9}$)	mode 1 obscures mod
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
Fresh- B null control	$\overline{\text{cos}}_{deep} = +0.002 \pm 0.022$ ($n=20$ draws)	meaningful	training-specific adaptation

172 Mode 2 has method-dependent severity within the audited fixed-feedback family once Mode 1 is
173 alleviated. Applying the same per-block scale-control penalty $\lambda=10^{-2}$ that rescued DFA to State
174 Bridge and to Credit Bridge on the same 4-block $d=256$ ResMLP backbone over 30 epochs and three
175 seeds gives converged test accuracies of 0.453 ± 0.003 (SB) and 0.360 ± 0.003 (CB), with deep mean
176 cosines of $+0.322 \pm 0.007$ (SB) and $+0.679 \pm 0.008$ (CB) and deep mean ρ of $+0.402 \pm 0.015$
177 (SB) and $+0.464 \pm 0.025$ (CB), while DFA under the same intervention reaches 0.363 ± 0.001 with
178 deep cosine $+0.155 \pm 0.025$ and deep ρ $+0.080 \pm 0.011$ (Table 2; Appendix J). The State Bridge
179 penalty rescue is roughly 24 percentage points above the vanilla State Bridge baseline of 0.213 on the
180 same architecture and, more importantly for the paper’s central walk-back, exceeds the architecture-
181 matched frozen-blocks shallow baseline of 0.349 by +10.4 percentage points. State Bridge with
182 the penalty intervention is therefore the first audited non-BP method whose trained deep blocks
183 substantively improve over an architecture-matched random-block baseline; the headline accuracy
184 gap is comparable to BP+penalty’s +18.1 pp over the same shallow baseline. Neither the activation
185 scale nor the deep BP gradient magnitude is silenced under the penalty: $\|h_L\|$ stays at 302 ± 8 for
186 SB and 5680 ± 178 for CB, with $\|g_L\|$ at $\sim 1.8 \times 10^{-4}$ and $\sim 1.9 \times 10^{-5}$ respectively, both well
187 within the meaningful-measurement regime, so the recovered deep cosines are computed against an
188 informative reference and not against a numerical floor. Within this rescued regime, Credit Bridge
189 has roughly $4\times$ the deep cosine of DFA and $2\times$ the deep cosine of State Bridge but reaches the same
190 final accuracy as DFA and 9 percentage points below State Bridge, so alignment to the BP gradient
191 is a necessary but not sufficient diagnostic of usable credit for depth. This cross-method dissociation
192 strengthens the methodological point that alignment must be reported jointly with measurement
193 validity and a depth-utilization baseline rather than as a single headline number.

194 5 Intervention and Cross-Architecture Evidence

195 The penalty intervention first matters as a rescue of the measurement regime. When we add a per-
196 block penalty $\lambda \text{mean}(\|f_i(h_i)\|^2)$ to DFA’s local loss and train the 4-block $d=256$ ResMLP for 30
197 epochs on CIFAR-10, the $\lambda=10^{-2}$ setting contains the terminal hidden-state scale from $\|h_L\| \sim$
198 4.4×10^8 under vanilla DFA to $\sim 4.0 \times 10^4$, while lifting the deepest BP reference norm from
199 $\|g_L\| \sim 5 \times 10^{-10}$ to $\sim 9.0 \times 10^{-7}$, a roughly four-order-of-magnitude rescue on both quantities
200 (Figure 3; Table 2). At that setting, both diagnostic (a) and diagnostic (b) pass on penalized DFA,
201 and test accuracy rises to 0.363 ± 0.001 from 0.308 ± 0.014 for vanilla DFA. The key point is not
202 yet that the recovered network has good deep credit, but that the deep reference vector is again large
203 enough to function as a meaningful target direction rather than a clamp-dominated artifact. That
204 rescue makes the second question measurable rather than hypothetical.

205 Once the reference vector is meaningful again, the deep layers no longer sit exactly at null. At
206 $\lambda=10^{-2}$, penalized DFA reaches a three-seed deep-layer mean cosine of $+0.155 \pm 0.025$ and deep
207 perturbation correlation of $+0.080 \pm 0.011$, whereas vanilla DFA is essentially zero on both metrics
208 in the deep blocks, consistent with prior concerns that alternative feedback can fail by supplying
209 poor credit directions even before full collapse [8, 9, 11, 10]. The null calibration rules out the inter-
210 pretation that this recovered signal is merely measurement noise: on the same penalized checkpoint,
211 replacing the training-time feedback matrices with 20 fresh random B_i draws gives a deep cosine
212 of only $+0.002 \pm 0.022$, with per-layer standard deviations of 0.013–0.023, all within noise of zero
213 (Table 2). The λ sweep sharpens the dissociation further: at $\lambda=10^{-4}$, Mode 1 is already alleviated,
214 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
215 rises to $+0.165$ and deep ρ to $+0.091$ (Figure 3). The improvement is real, but it is only partial.

216 A rescue intervention is only informative if its direct cost is controlled. The relevant control is BP
217 trained under the same penalty: BP falls from 0.609 ± 0.004 without the penalty to 0.530 with

218 $\lambda=10^{-2}$, so the penalty has a direct cost of about 8 percentage points even when credit assignment
 219 is correct, whereas DFA moves in the opposite direction, from 0.308 ± 0.014 to 0.363 ± 0.001 ,
 220 and State Bridge moves further still, from 0.213 to 0.453 ± 0.003 (three seeds), under the same
 221 intervention (Figure 3; Appendix J). Relative to the frozen-blocks baseline of 0.349, BP+penalty
 222 retains a margin of +18.1 points, State Bridge+penalty retains +10.4 points, and DFA+penalty
 223 retains only +1.4 points. The remaining BP-to-DFA gap of 17 points is therefore a lower bound
 224 on the part of DFA’s deficit that is not explained by simple penalty-induced capacity loss alone,
 225 though not a clean isolation because BP uses an end-to-end loss whereas DFA uses block-local
 226 losses. The substantially smaller BP-to-State-Bridge gap of $0.530 - 0.453 = 7.7$ points shows
 227 that the cross-method differences in penalty-rescued accuracy are not all attributable to a uniform
 228 “random-feedback ceiling”: the bridge construction in State Bridge can recover much more of the
 229 BP-with-penalty performance than DFA can, on the same architecture and the same intervention.
 230 The residual gap after that control is what keeps Mode 2 substantively alive while letting it have
 231 method-dependent severity.

232 The architecture comparison sharpens the scope of the critique. In the terminal-LN architectures we
 233 audited, both diagnostics fire for DFA-trained ResMLP at $d=256$, the same pattern recurs at $d=512$
 234 with even larger max-per-block growth (about 1.5×10^4), and ViT-Mini with a class token and ter-
 235 minal LN shows diagnostic (a) by epoch 1 and diagnostic (b) by epochs 2–3 (Figure 2). A depth
 236 sweep on the $d=512$ ResMLP at $L \in \{2, 4, 6, 8, 12\}$ shows that the layerwise pattern is essentially
 237 depth-invariant: DFA’s layer-0 cosine stays in $[+0.39, +0.40]$ across all five depths, while its mean
 238 deep-layer cosine stays within $[-0.005, +0.000]$ and its deep perturbation correlation collapses to
 239 0.000 in every depth tested, even though BP retains a deep-layer cosine of +0.94 at $L=12$ (Ap-
 240 pendix G). The deep credit signal does not improve when the network is shallower, so the failure
 241 is not a “too deep” artifact. In the non-terminal-LN controls, the pattern is different: StudentNet
 242 shows diagnostic (a) only at epochs 14–25 while diagnostic (b) never fires across 100 epochs and
 243 three seeds, and the BatchNorm CNN on CIFAR-10 likewise shows strong growth under DFA, with
 244 max-per-block growth up to $237\times$, but keeps deepest BP gradients around $\|g\| \sim 10^{-3}$ and never
 245 triggers diagnostic (b) (Figure 2). BP never triggers either diagnostic in any audited architecture.
 246 The matched same-backbone ResMLP-d256 ablation in Section 3 supplies the cleanest causal con-
 247 trol: removing terminal LayerNorm from the same architecture preserves activation growth but elim-
 248 inates the gradient floor, so diagnostic (b) is necessary on terminal-LN ResMLP and is not just an
 249 architecture-class coincidence. The broader claim therefore holds at full strength inside the audited
 250 residual ResMLP and ViT-Mini regime, while diagnostic (a) remains useful more broadly. This lets
 251 the paper end with a reporting rule rather than an overclaimed theory.

252 6 Recommended FA Evaluation Protocol

253 The reporting protocol begins with measurement validity. Before any FA paper reports a headline
 254 alignment number, it should report per-layer state scale and the hidden BP reference-gradient scale
 255 at the layers where the scientific claim is being made. In our audited regime, those two quantities
 256 already separate healthy from invalid measurement with unusually wide margins: the maximum
 257 per-block growth stays below about $11\times$ for BP and EP but is at least $694\times$ for the degenerate
 258 methods, giving a $63\times$ calibration gap, while the deepest hidden BP norm stays above about 10^{-4}
 259 for BP and EP but below about 4×10^{-9} for the degenerate methods, giving a $24,338\times$ gap (Table 3;
 260 Table 1; Figure 4). These are not cosmetic diagnostics around the real result: they determine whether
 261 the reported cosine is being computed against an informative BP direction or against a floor-level
 262 reference. If the reference gradient is at floor, the evaluator should stop treating aggregate alignment
 263 as evidence.

264 The point of the protocol is not to add plots; it is to prevent a specific class of false conclusions. For
 265 this paper, the minimal protocol is four checks: per-layer activation scale via max-per-block growth,
 266 deepest hidden BP gradient floor, meaningful-regime per-layer credit quality, and an architecture-
 267 matched frozen-blocks baseline (Table 3). The first two ask whether the reference quantity is still
 268 valid; the third asks whether, once validity is restored, the deep blocks receive useful directions;
 269 and the fourth asks whether the trained depth is doing better than a model whose residual blocks
 270 were never trained at all. Figure 5 makes the decision value explicit: accuracy alone walks back
 271 0/5 audited methods, accuracy plus headline Γ still walks back 0/5, and the full protocol walks
 272 back 3/5 by flagging DFA, State Bridge, and Credit Bridge, with diagnostics (a), (b), and (d) each

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

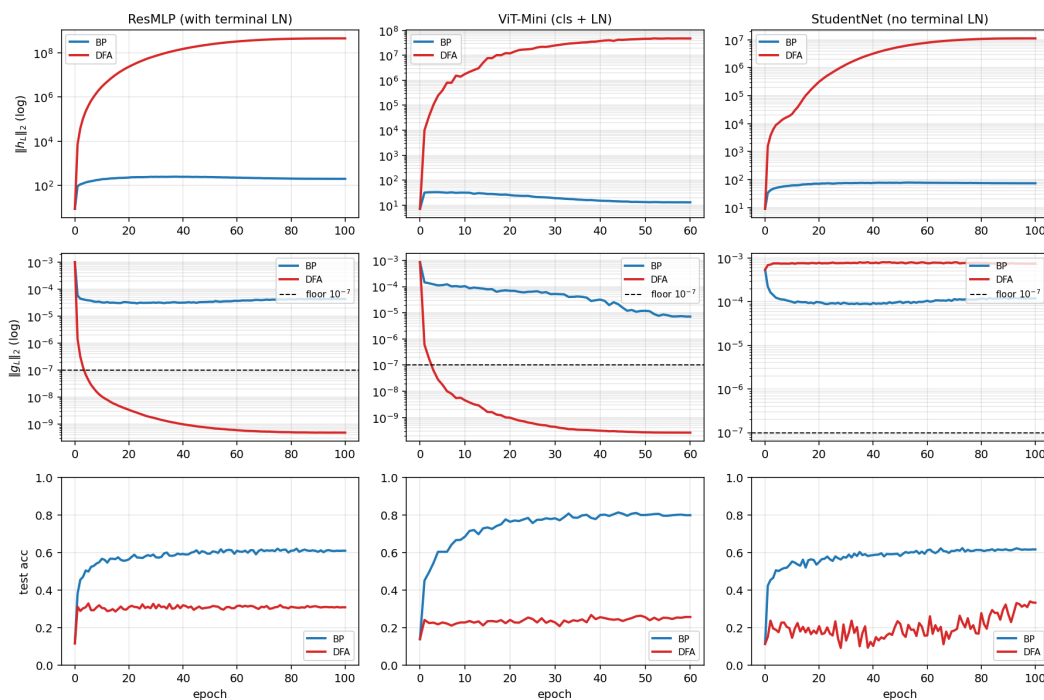


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

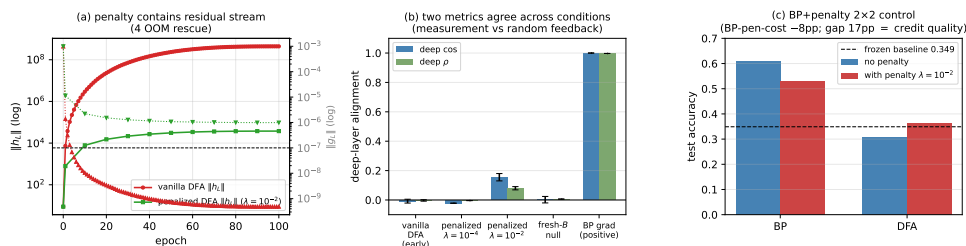


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

273 independently sufficient for binary detection on those failures. On our audit, these checks catch
 274 failures that accuracy plus aggregate alignment miss completely.

275 A useful evaluation rule should reject the bad cases without collapsing everything into a negative
 276 result. The protocol is conservative in exactly that sense: it preserves BP and EP as evidence-bearing
 277 controls, and it walks back only those claims that fail measurement-validity or depth-utilization
 278 checks in Table 1. That asymmetry is important because the thresholds are not equally strong in
 279 the same way. Diagnostics (a) and (b) have sharp empirical calibration gaps in the audited regime,
 280 diagnostic (c) is explicitly a sub-mode discriminator rather than a primary detector, and diagnostic
 281 (d) uses a deliberately weak 2pp margin as a context check rather than a theorem about useful depth.
 282 The rule therefore does not say that low accuracy, low aggregate alignment, or any non-BP method
 283 is automatically invalid; it says only that claims unsupported by measurement-valid evidence should
 284 be withdrawn, while trustworthy controls should remain standing. That conservative asymmetry is
 285 why the protocol belongs in the main paper rather than the appendix.

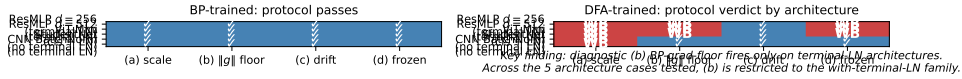


Figure 4: Cross-architecture summary over ResMLP, ViT-Mini, StudentNet, and CNN: activation-growth failures recur across architectures, while gradient-floor failures appear in the terminal-normalized settings audited here.

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_l \ h_{l+1}\ /\ h_l\ $	$> 50\times$	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

286 7 Discussion, Limits, Conclusion

287 Our claim is about what existing evidence licenses, not about impossibility. This paper does not show
 288 that FA cannot work in deep networks; it shows that current evaluation practice can misread what
 289 happened by letting headline accuracy and aggregate alignment stand in for measurement validity
 290 and layerwise credit quality. The strongest examples are precisely the cases where the field-standard
 291 summary would sound mildly positive while the audited deep evidence has already collapsed or
 292 is already null: DFA, State Bridge, and Credit Bridge all survive status-quo reporting in Table 1,
 293 yet the protocol shows that their deep claims are unsupported. The intervention results in Figure 3
 294 reinforce the same distinction, because restoring a measurable regime partially rescues deep credit
 295 signal rather than proving that the original headline had been trustworthy all along. That distinction
 296 is important because evaluation failure and algorithmic impossibility are different statements.

297 The right level of generality is the audited regime. Our strongest claim is scoped to modern resid-
 298 ual vision architectures, especially the pre-LayerNorm and terminal-LayerNorm settings where we
 299 directly observed Mode 1: the 4-block ResMLP at $d=256$, its $d=512$ extension, and ViT-Mini all
 300 show the same basic pattern, whereas StudentNet and the BatchNorm CNN refine the scope by show-
 301 ing that activation-growth failures can persist without the hidden-gradient-floor collapse (Figure 4;
 302 Figure 3). That leaves clear limits. The dataset is only CIFAR-10, the models are small to medium
 303 rather than frontier-scale, the terminal-LN interpretation is observational rather than a causal iden-
 304 tification, and the BP-plus-penalty comparison is only a lower-bound control on penalty cost rather
 305 than a perfect decomposition. Those limitations narrow what is claimed, but they do not weaken the
 306 core methodological point that the audited measurement regime can fail silently in exactly the archi-
 307 tectures that now dominate this genre of experiment. Future positive or negative examples outside
 308 this regime would refine the scope of the protocol, not invalidate the critique.

309 The main lesson is to decompose the evaluation question before interpreting the answer. Future
 310 FA papers should report, separately, whether the BP reference is still meaningful, whether the
 311 deep layers receive useful credit in that meaningful regime, and whether trained depth beats an
 312 architecture-matched frozen-blocks baseline, instead of compressing those distinct questions into a
 313 single headline accuracy or headline Γ . That is the sense in which this paper fits the evaluation-
 314 methodology line of Jordan et al. [3], O’Bray et al. [2], and Paleka et al. [1]: the contribution is not a
 315 new benchmark artifact, but a reporting rule for preventing a repeatable interpretive error. Once the
 316 field enforces that separation between measurement validity and substantive credit quality, positive
 317 results will become more trustworthy and negative results more precise. Once that decomposition
 318 is enforced, the apparent evidence for successful deep credit assignment becomes much harder to
 319 overstate.

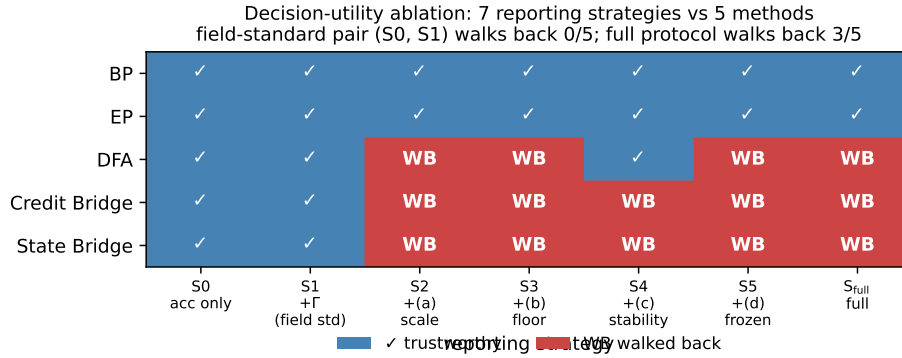


Figure 5: Decision-utility ablation comparing the field-standard reporting pair against progressively richer diagnostic strategies: accuracy only and accuracy+ Γ walk back no audited failures, while the full protocol walks back the three silent failures.

References

- 320
- 321 [1] Daniel Paleka et al. Pitfalls in evaluating model behavior: measurement, reporting, and inter-
322 pretability failures. In *International Conference on Learning Representations*, 2026.
- 323 [2] Leslie O’Bray et al. Evaluation beyond leaderboard metrics: methodology matters. In *Internation-
324 al Conference on Learning Representations*, 2022.
- 325 [3] Matt Jordan et al. Evaluating machine learning: tests, cases, and expectations. In *International
326 Conference on Machine Learning*, 2020.
- 327 [4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random
328 synaptic feedback weights support error backpropagation for deep learning. *Nature Communi-
329 cations*, 7:13276, 2016.
- 330 [5] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In
331 *Advances in Neural Information Processing Systems*, 2016.
- 332 [6] Mohamad Akrouf, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B.
333 Tweed. Deep feedback control. In *Advances in Neural Information Processing Systems*, 2019.
- 334 [7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback align-
335 ment scales to modern deep learning tasks and architectures. In *Advances in Neural Informa-
336 tion Processing Systems*, 2020.
- 337 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and
338 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-
339 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 340 [9] Ted H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep con-
341 volutional networks. In *Advances in Neural Information Processing Systems*, 2018.
- 342 [10] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Florent Krzakala. Aligning residual
343 pathways: normalization, scale, and feedback in deep networks. In *International Conference
344 on Machine Learning*, 2023.
- 345 [11] Brian Crafton, Abhinav Parihar, Eric Gebhardt, and Arijit Raychowdhury. Backpropagation
346 through feedback alignment for deep learning in analog hardware. In *International Conference
347 on Acoustics, Speech, and Signal Processing*, 2019.
- 348 [12] Ruibin Xiong, Yunchang Yu, and others. On layer normalization in the transformer architecture.
349 In *International Conference on Machine Learning*, 2020.

350 A Reference Implementation

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

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

365 B Pipeline Pitfalls Catalog

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

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

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

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

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

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

391 **Pitfall 7: Missing null controls for intervention claims.** A rescue intervention can improve co-
392 sine or accuracy for trivial reasons unless the experiment includes a null such as fresh- B feedback
393 or a matched BP+penalty control. The paper therefore treats intervention evidence as incomplete
394 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

395 C Walk-Back Chain Methodology

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

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

409 D All Seven Validations

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

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

420 E Threshold Sensitivity Full Sweep

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

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

432 F Per-Architecture Detailed Audits

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

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

444 G Depth-Sweep Layerwise Profiles

445 To check whether the layerwise pattern in Figure 1 is an artifact of the specific four-block depth
446 used in the main audit, we ran the same architecture on $d=512$ pre-LayerNorm ResMLPs at five
447 depths $L \in \{2, 4, 6, 8, 12\}$ on CIFAR-10 (single seed 42, otherwise matched configuration). Table 5
448 reports the layer-0 cosine, the mean cosine over all deeper layers, and the deep mean perturbation
449 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

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

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

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

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

469 The qualitative shape matches what we see in vanilla residual DFA, only with a slower onset because
 470 the architecture itself is harder to train. Diagnostic (a) clearly fires within three epochs, and diag-
 471 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
 472 the same smoke sweep, the qualitative outcome is the same: residual stream grows by three to four
 473 orders of magnitude, $\|g_L\|$ drops by three to four orders of magnitude, and BP itself never reaches a
 474 healthy training regime. We retain $w_2=0.5$ here because that is the only value where BP is at least
 475 beginning to learn. The full 100-epoch trajectory of the same configuration, replicated across three
 476 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-
 477 seed values $\|h_L\| \in \{1.06 \times 10^8, 3.15 \times 10^7, 1.09 \times 10^8\}$ and $\|g_L\| \in \{1.08, 2.94, 1.77\} \times 10^{-10}$),
 478 all deeply below the diagnostic (b) floor and within an order of magnitude of vanilla residual DFA’s
 479 $\|h_L\| \approx 4 \times 10^8$ and $\|g_L\| \approx 5 \times 10^{-10}$ on the same backbone, confirming that the smoke-test trend
 480 is the converged behavior rather than an early-training artifact.

481 We treat this ablation as evidence about *necessity*, not about clean algorithm separation. Specifically,
 482 the evidence supports: the additive residual skip is not necessary for Mode 1 activation growth
 483 or for the gradient-floor trend; Mode 1 (a) appears to be a generic deep-DFA instability on these
 484 stacks, modulated but not gated by skip presence; and the catastrophic, well-defined $\|g_L\|$ collapse
 485 remains most tightly associated with terminal LayerNorm in our audited settings, where the no-
 486 out_In control already showed activation growth without the same severity of collapse. The full
 487 100-epoch trajectory of this no-residual run is reported as a confirmatory check rather than as a
 488 primary claim.

489 **I Random-Target Ablation: Mode 1 Is Data-Agnostic**

490 To test whether Mode 1 activation growth requires any task signal at all, we re-ran DFA on the stan-
 491 dard 4-block $d=256$ pre-LayerNorm ResMLP, on CIFAR-10 inputs, but replaced each minibatch’s
 492 labels with i.i.d. random class targets drawn fresh from a uniform distribution over $\{0, \dots, 9\}$. All
 493 other hyperparameters are matched to the vanilla DFA training run in Section 2 (AdamW, lr= 10^{-3} ,
 494 wd= 0.01, 128 batch, cosine schedule, single seed 42 for the smoke test). The local feedback vectors
 495 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

496 This ablation answers the natural counterargument that DFA’s residual-stream growth might be a
 497 side-effect of the network adapting to genuine task signal in a particularly bad local minimum: it
 498 is not. With no task signal at all, DFA on this architecture still inflates the residual stream by more
 499 than three orders of magnitude in the first three epochs and pushes the deepest BP reference gradient
 500 to the floor of 10^{-7} in the same window. The full 100-epoch trajectory of the same DFA random-
 501 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
 502 the corresponding endpoints of vanilla DFA on the same backbone with real labels (about 4×10^8
 503 and 5×10^{-10} respectively), so the data-agnostic trajectory does not just reach Mode 1 but in fact
 504 passes through the same regime even without any per-sample task pressure. The local DFA objective
 505 $\langle f_l(h_l), e_T B_l^T \rangle$ contains no penalty on $\|f_l(h_l)\|$, so any direction in which a larger block output
 506 increases inner-product alignment with the fixed feedback target is rewarded; the random-target run
 507 isolates exactly this geometric incentive, free of any task-driven feature pressure. The full 100-epoch
 508 trajectory of this random-target run is reported as a confirmatory check rather than a primary claim.

509 We then asked whether this data-agnostic growth is specific to DFA or generalizes to other fixed-
 510 feedback local-credit methods, by repeating the random-target ablation under State Bridge and
 511 Credit Bridge with the same architecture, hyperparameters, and seed. Both methods also exhibit
 512 data-agnostic activation growth in the same three-epoch window, with $\|h_L\|$ rising from about 9 to
 513 about 6.2×10^3 (State Bridge) and about 2.0×10^4 (Credit Bridge), while their test accuracies remain
 514 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

515 The cross-method version of the test rules out the explanation that the random-target growth is
 516 specific to DFA’s particular feedback projection. State Bridge and Credit Bridge use bridge con-
 517 structions with target normalization and stop-gradients, so any residual-stream growth they exhibit
 518 cannot be attributed to a simple absence of normalization. Their $\|g_L\|$ values at three epochs are
 519 still well above the 10^{-7} floor used by diagnostic (b), so the gradient collapse part of Mode 1 does
 520 not yet appear at this horizon for SB/CB; the activation-growth part of Mode 1 is already present.

521 At the full 100-epoch trajectory of the same random-target protocol, both SB and CB also reach
 522 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
 523 $\|h_L\| \approx 1.38 \times 10^8$ and $\|g_L\| \approx 0$ (below the numerical clamp), with test accuracies 0.100 and
 524 0.085 respectively, consistent with DFA’s 1.67×10^8 and 8.0×10^{-12} at the same horizon. We
 525 treat this as evidence that the local-credit growth incentive is not unique to DFA but is shared by the
 526 audited family of fixed-feedback methods.

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

539 J State Bridge Penalty Rescue: 3-Seed Cross-Method Test

540 To test whether the per-block scale-control penalty $\lambda \text{mean}(\|f_i(h_i)\|^2)$ that rescues DFA in Sec-
 541 tion 5 also rescues other audited fixed-feedback local-credit methods, we re-ran State Bridge on
 542 the standard 4-block $d=256$ pre-LayerNorm ResMLP for 30 epochs and three seeds (42, 123, 456),
 543 with $\lambda=10^{-2}$ added to the State Bridge per-block local loss only (the bridge state predictor and the
 544 embedding/head paths are not penalized, matching the DFA rescue setup). We also ran a matched
 545 vanilla State Bridge baseline at seed 42 with the same architecture and training schedule but $\lambda=0$.
 546 Three-seed converged values:

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 (3 seeds)	0.363 ± 0.001	4.0×10^4	9.0×10^{-7}	$+0.155 \pm 0.025$	$+0.080 \pm 0.011$

547 The penalty rescue effect on State Bridge is much larger than on DFA: +24 percentage points for
 548 State Bridge versus +5.5 percentage points for DFA on the same architecture and intervention.
 549 SB+penalty is the first audited non-BP method whose trained deep blocks substantially beat the
 550 architecture-matched random-block baseline. We treat this as evidence that Mode 2 (low intrinsic
 551 credit-direction quality) has method-dependent severity within the audited fixed-feedback family

552 once Mode 1 is alleviated, rather than being a uniform property of all fixed-feedback local-credit ob-
553 jectives. Importantly, State Bridge’s deep cosine $+0.322$ is approximately twice DFA’s $+0.155$ on
554 the same intervention, but neither approaches the BP reference value of $\approx +1.0$, so this is a within-
555 class gradation in credit-direction quality, not a claim that bridge constructions “solve” Mode 2.
556 Under the same intervention Credit Bridge reaches a three-seed test accuracy of 0.360 ± 0.003 , a
557 three-seed deep mean cosine of $+0.679 \pm 0.008$, and a three-seed deep mean ρ of $+0.464 \pm 0.025$,
558 with $\|h_L\| \approx 5680 \pm 178$ and $\|g_L\| \approx 1.9 \times 10^{-5}$ well above the diagnostic floor. Credit Bridge
559 therefore has an even higher deep cosine than State Bridge (about $4\times$ the DFA value and roughly
560 $2\times$ the State Bridge value), but reaches the same final accuracy as DFA+penalty and 9.3 percentage
561 points below State Bridge+penalty. This is a clean dissociation: within the audited fixed-feedback
562 family under the same rescue, deep cosine and deep ρ differ by more than a factor of four across
563 methods without tracking final accuracy in the same direction, so alignment to the BP gradient is
564 a necessary but not sufficient diagnostic of usable credit for depth. That cross-method dissociation
565 is a direct reason the protocol in Section 6 keeps final accuracy, layerwise credit quality, and the
566 depth-utilization baseline as three separate reporting axes rather than collapsing them into a single
567 headline.

568 **K Reproducibility**

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

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