
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 Evalua-
13 tions & 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 (learning
 62 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)

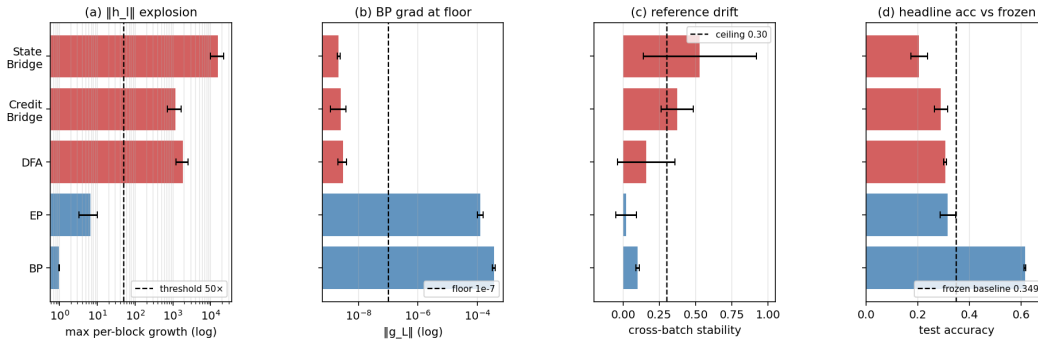


Figure 1: Five-method audit on the 4-block $d=256$ pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

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 The first failure mode is a scale pathology, not yet an alignment pathology. On the audited 4-block
 95 pre-LayerNorm ResMLP ($d=256$, CIFAR-10, 100 epochs, 3 seeds), DFA optimizes block-local
 96 objectives of the form $\langle f_l(h_l), e_T B_l^T \rangle$ with no explicit scale constraint on f_l , so the residual stream
 97 is free to inflate while still reducing the local loss [7]. In the same runs, each block’s w_1 and w_2
 98 grows by roughly $200\times$ in relative delta, their norm product reaches about 5×10^4 per block, and
 99 the terminal hidden-state norm $\|h_L\|$ rises monotonically from about 9 at random initialization to
 100 about 4×10^8 by epoch 100 (Figure 2). Most of that growth appears immediately: $\|h_L\|$ already
 101 reaches about 10^6 by epoch 5. Once the residual stream reaches this regime, the backpropagation
 102 reference vector no longer behaves like a healthy target.

103 The measurement failure occurs at the point where the hidden-layer BP gradient ceases to be a mean-
 104 ingful reference direction. In terminal-LayerNorm architectures, the LayerNorm Jacobian scales
 105 as $\partial \text{LN}(h)/\partial h \propto 1/\|h\|$ in expectation, so the same residual-stream inflation is accompanied by
 106 collapse of the hidden-layer BP reference norm: on DFA-trained ResMLP, $\|g_L\|$ falls from about
 107 9.8×10^{-4} at random initialization to about 5×10^{-10} by epoch 100, a six-order-of-magnitude drop,
 108 while the reported cosine remains mathematically defined only because `F.cosine_similarity`
 109 clamps the denominator at $\varepsilon=10^{-8}$ (Table 1; Figure 1). At that endpoint the reference norm is about
 110 $20\times$ below the clamp, so the quantity being reported is effectively $(a \cdot b)/(\|a\| \max(\|b\|, 10^{-8}))$
 111 rather than a comparison to an informative BP direction. At that point, reporting a cosine is no
 112 longer evidence about credit quality.

113 The simplest control is architectural, not theoretical. On the same ResMLP backbone, BP keeps
 114 $\|h_L\|$ near 200 and $\|g_L\|$ near 4×10^{-4} throughout training, while EP keeps $\|h_L\|$ around 5×10^3
 115 and $\|g_L\|$ around 1.3×10^{-4} , so hard optimization on CIFAR-10 by itself does not force hidden-

116 layer gradients to the numerical floor (Table 1; Figure 2). The broader cross-architecture pattern is
 117 consistent with the same interpretation: StudentNet and the BatchNorm CNN, which lack terminal
 118 LayerNorm, keep deepest BP gradients around 10^{-4} and never trigger diagnostic (b), whereas ViT-
 119 Mini with a terminal LN shows the same collapse pattern and triggers diagnostic (b) by epochs 2–3
 120 (Figure 2). The pathology therefore belongs to the evaluated FA regime, not to CIFAR-10 or the
 121 backbone alone.

122 The collapse is not a late-epoch curiosity. For vanilla DFA on the ResMLP temporal replay, $\|g_L\|$
 123 drops from 9.8×10^{-4} at epoch 0 to 1.4×10^{-6} at epoch 1, 3.1×10^{-7} at epoch 2, 1.3×10^{-7} at
 124 epoch 3, and 6.7×10^{-8} at epoch 4, so diagnostic (b) fires at epoch 3–4 across all three seeds, while
 125 the max-per-block growth detector fires slightly later at epochs 8–11 (Figure 2). Both detectors
 126 therefore fire in the first 11 epochs of a 100-epoch run, making the protocol actionable as an early-
 127 stop criterion rather than a post hoc explanation. The practical point is reinforced by accuracy: DFA
 128 is at 0.308 already at epoch 4 and ends at 0.306 by epoch 100, so the remaining training budget
 129 adds essentially nothing to the headline result once the measurement has already degenerated. Once
 130 measurement degeneracy is identified, the next question is whether poor deep credit remains even
 131 before collapse.

132 4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

133 The second failure mode appears even in the meaningful-measurement regime. At the earliest vanilla
 134 DFA checkpoints on ResMLP, the hidden backpropagated gradient at the first deep block remains
 135 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
 136 seeds, all above the 10^{-7} threshold used to distinguish measurable from collapsed gradients. Yet the
 137 corresponding deep-layer cosine values are already essentially null: across layers 1–4, all seed-level
 138 measurements at epoch 1 lie in $[-0.04, +0.02]$, with a three-seed mean of -0.008 ± 0.013 , and
 139 by epoch 2 the deep mean is still only -0.018 ± 0.018 (Table 2). This is the observational pattern
 140 predicted by low credit-direction quality rather than mere disappearance of signal: the gradient is
 141 still present enough to measure, but the directions delivered to the deep network carry little agree-
 142 ment with backpropagation, consistent with prior concerns that alternative feedback rules can fail by
 143 supplying poor credit assignments even before full collapse [8, 9, 11?]. This rules out the simplest
 144 objection that the deep-layer null result is merely a byproduct of collapse.

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

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

166 5 Intervention and Cross-Architecture Evidence

167 The penalty intervention first matters as a rescue of the measurement regime. When we add a per-
 168 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
 169 epochs on CIFAR-10, the $\lambda=10^{-2}$ setting contains the terminal hidden-state scale from $\|h_L\| \sim$
 170 4.4×10^8 under vanilla DFA to $\sim 4.0 \times 10^4$, while lifting the deepest BP reference norm from
 171 $\|g_L\| \sim 5 \times 10^{-10}$ to $\sim 9.0 \times 10^{-7}$, a roughly four-order-of-magnitude rescue on both quantities
 172 (Figure 3; Table 2). At that setting, both diagnostic (a) and diagnostic (b) pass on penalized DFA,
 173 and test accuracy rises to 0.363 ± 0.001 from 0.308 ± 0.014 for vanilla DFA. The key point is not
 174 yet that the recovered network has good deep credit, but that the deep reference vector is again large
 175 enough to function as a meaningful target direction rather than a clamp-dominated artifact. That
 176 rescue makes the second question measurable rather than hypothetical.

177 Once the reference vector is meaningful again, the deep layers no longer sit exactly at null. At
 178 $\lambda=10^{-2}$, penalized DFA reaches a three-seed deep-layer mean cosine of $+0.155 \pm 0.025$ and deep
 179 perturbation correlation of $+0.080 \pm 0.011$, whereas vanilla DFA is essentially zero on both metrics
 180 in the deep blocks, consistent with prior concerns that alternative feedback can fail by supplying
 181 poor credit directions even before full collapse [8, 9, 11?]. The null calibration rules out the inter-
 182 pretation that this recovered signal is merely measurement noise: on the same penalized checkpoint,
 183 replacing the training-time feedback matrices with 20 fresh random B_l draws gives a deep cosine
 184 of only $+0.002 \pm 0.022$, with per-layer standard deviations of 0.013–0.023, all within noise of zero
 185 (Table 2). The λ sweep sharpens the dissociation further: at $\lambda=10^{-4}$, Mode 1 is already alleviated,
 186 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
 187 rises to $+0.165$ and deep ρ to $+0.091$ (Figure 3). The improvement is real, but it is only partial.

188 A rescue intervention is only informative if its direct cost is controlled. The relevant control is BP
 189 trained under the same penalty: BP falls from 0.609 ± 0.004 without the penalty to 0.530 with
 190 $\lambda=10^{-2}$, so the penalty has a direct cost of about 8 percentage points even when credit assignment
 191 is correct, whereas DFA moves in the opposite direction, from 0.308 ± 0.014 to 0.363 ± 0.001
 192 under the same intervention (Figure 3). Relative to the frozen-blocks baseline of 0.349, BP+penalty
 193 still retains a margin of $+18.1$ points, while DFA+penalty retains only $+1.4$ points. The remaining
 194 gap, $0.530 - 0.363 = 17$ points, is therefore a lower bound on the part of DFA’s deficit that is not
 195 explained by simple penalty-induced capacity loss alone, though not a clean isolation because BP
 196 uses an end-to-end loss whereas DFA uses block-local losses. The residual gap after that control is
 197 what keeps Mode 2 substantively alive.

198 The architecture comparison sharpens the scope of the critique. In the terminal-LN architectures we
 199 audited, both diagnostics fire for DFA-trained ResMLP at $d=256$, the same pattern recurs at $d=512$
 200 with even larger max-per-block growth (about 1.5×10^4), and ViT-Mini with a class token and
 201 terminal LN shows diagnostic (a) by epoch 1 and diagnostic (b) by epochs 2–3 (Figure 2). In the non-
 202 terminal-LN controls, the pattern is different: StudentNet shows diagnostic (a) only at epochs 14–25
 203 while diagnostic (b) never fires across 100 epochs and three seeds, and the BatchNorm CNN on
 204 CIFAR-10 likewise shows strong growth under DFA, with max-per-block growth up to $237\times$, but
 205 keeps deepest BP gradients around $\|g\| \sim 10^{-3}$ and never triggers diagnostic (b) (Figure 2). BP
 206 never triggers either diagnostic in any audited architecture. This is an observational association
 207 rather than a causal identification of terminal LayerNorm as the unique mechanism, but it is enough
 208 to support a narrower claim: diagnostic (b) appears tied to the terminal-LN architectures audited
 209 here, while diagnostic (a) remains useful more broadly. This lets the paper end with a reporting rule
 210 rather than an overclaimed theory.

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

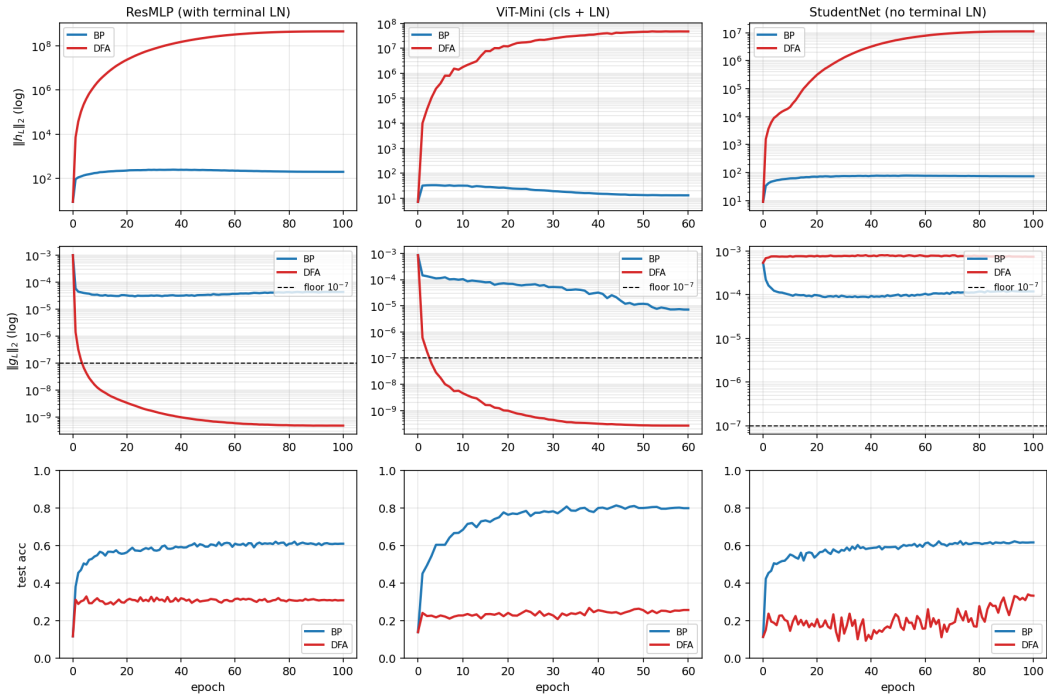


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

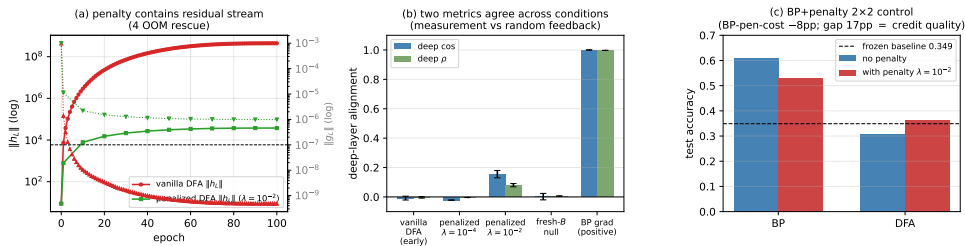


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

211 6 Recommended FA Evaluation Protocol

212 The protocol has four diagnostics because the evaluation failure is not visible from any single head-
 213 line number.

214 Diagnostics (a), (b), and (d) are independently sufficient for binary detection on the audited failures,
 215 while diagnostic (c) is primarily interpretive.

216 The decision-utility ablation is the compact empirical argument for why this protocol belongs in an
 217 E&D paper.

218 Threshold calibration is strong for diagnostics (a) and (b) and deliberately weaker for diagnostic (d),
 219 so the paper should state that asymmetry rather than oversell uniform robustness.

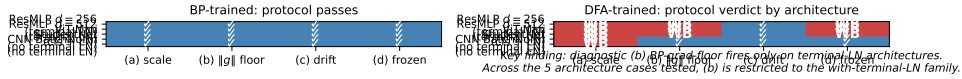


Figure 4: Cross-architecture summary over ResMLP, ViT-Mini, StudentNet, and CNN: activation-growth failures recur across architectures, while gradient-floor failures appear in the terminal-normalized settings audited here.

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

220 7 Discussion, Limits, Conclusion

221 The main recommendation of this paper is that headline accuracy and headline Γ should no longer
 222 be treated as sufficient evidence that deep local-credit learning is working on modern residual archi-
 223 tectures.

224 Our claim is deliberately scoped to the architectures and methods audited here, and especially to
 225 pre-LayerNorm residual settings where measurement degeneracy is empirically strongest.

226 Positioned against prior evaluation-methodology papers, this work contributes a failure analysis and
 227 diagnostic protocol for a mature evaluation practice rather than a new benchmark suite, dataset
 228 release, or leaderboard [3, 2, 1].

229 A reasonable conclusion for the field is therefore not that FA-like methods are categorically impossi-
 230 ble, but that future claims must report whether they have escaped both failure modes, under matched
 231 baselines and with diagnostics that remain meaningful at the layers where the scientific claim is
 232 being made.

233 References

- 234 [1] Daniel Paleka et al. Pitfalls in evaluating model behavior: measurement, reporting, and inter-
 235 pretability failures. In *International Conference on Learning Representations, 2026*.
- 236 [2] Leslie O’Bray et al. Evaluation beyond leaderboard metrics: methodology matters. In *Internat-
 237 tional Conference on Learning Representations, 2022*.
- 238 [3] Matt Jordan et al. Evaluating machine learning: tests, cases, and expectations. In *International
 239 Conference on Machine Learning, 2020*.
- 240 [4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random
 241 synaptic feedback weights support error backpropagation for deep learning. *Nature Communi-
 242 cations*, 7:13276, 2016.
- 243 [5] Arild Nøklund. Direct feedback alignment provides learning in deep neural networks. In
 244 *Advances in Neural Information Processing Systems, 2016*.
- 245 [6] Mohamad Akrouf, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B.
 246 Tweed. Deep feedback control. In *Advances in Neural Information Processing Systems, 2019*.
- 247 [7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback align-
 248 ment scales to modern deep learning tasks and architectures. In *Advances in Neural Informa-
 249 tion Processing Systems, 2020*.

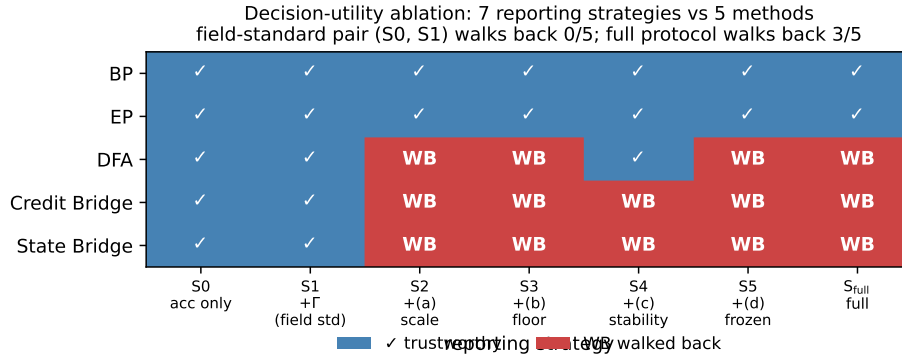


Figure 5: Decision-utility ablation comparing the field-standard reporting pair against progressively richer diagnostic strategies: accuracy only and accuracy+ Γ walk back no audited failures, while the full protocol walks back the three silent failures.

- 250 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and
251 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-
252 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 253 [9] Ted H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep con-
254 volutional networks. In *Advances in Neural Information Processing Systems*, 2018.
- 255 [10] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Florent Krzakala. Aligning residual
256 pathways: normalization, scale, and feedback in deep networks. In *International Conference*
257 *on Machine Learning*, 2023.
- 258 [11] Brian Crafton, Abhinav Parihar, Eric Gebhardt, and Arijit Raychowdhury. Backpropagation
259 through feedback alignment for deep learning in analog hardware. In *International Conference*
260 *on Acoustics, Speech, and Signal Processing*, 2019.
- 261 [12] Ruibin Xiong, Yunchang Yu, and others. On layer normalization in the transformer architecture.
262 In *International Conference on Machine Learning*, 2020.

263 A Reference Implementation

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

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

278 B Pipeline Pitfalls Catalog

279 **Pitfall 1: Layer-0 dominance hidden by global averaging.** A single global cosine can look
280 mildly positive even when all deep trainable blocks are effectively null, because the shallowest layer

281 dominates the norm budget. The protocol therefore treats layerwise inspection as mandatory and
282 interprets any aggregate headline only after checking where the signal comes from.

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

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

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

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

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

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

308 C Walk-Back Chain Methodology

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

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

322 D All Seven Validations

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

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

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

333 E Threshold Sensitivity Full Sweep

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

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

345 F Per-Architecture Detailed Audits

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

351 StudentNet and the CNN serve a different role. They test whether the protocol overgeneralizes from
 352 terminal-normalized residual architectures to settings where gradient-floor collapse is not expected.
 353 In those models, activation-growth checks can still reveal weak depth usage or poor scaling, but
 354 diagnostic (b) is not expected to fire in the same way. This asymmetry is not a weakness of the pro-

355 tocol; it is part of the empirical scoping claim of the paper and helps prevent readers from mistaking
356 a targeted evaluation standard for a universal pathology claim [12, 8].

357 **G Reproducibility**

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

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