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

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Modern feedback-alignment evaluation on deep residual networks is still summarized by a deceptively simple pair: headline accuracy and headline cosine alignment Γ to the backpropagation gradient. We show that this pair can silently fail in two distinct ways on standard CIFAR-10 pre-LayerNorm ResMLP and ViT-Mini settings: first, *measurement degeneracy*, where residual-stream growth drives hidden-layer BP gradients to the numerical floor and makes Γ uninterpretable; and second, *low intrinsic credit-direction quality*, where random-feedback credit remains essentially unaligned with BP on the deep blocks even when the reference gradient is still meaningful. The headline result is that the field-standard reporting pair walks back none of the methods we audit, whereas a four-diagnostic protocol walks back the three degenerate methods and passes the two trustworthy controls. Our contribution is an evaluation methodology paper for the NeurIPS 2026 Evaluations & Datasets track: we provide the protocol, the calibration logic for its thresholds, a reference implementation, a five-method audit, and validation through temporal replay, cross-architecture checks, intervention-based disambiguation, and a documented catalog of pipeline pitfalls, in the spirit of critical evaluation analyses such as Jordan et al. [3], O'Bray et al. [2], Paleka et al. [1].

1 Introduction

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

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

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

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

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

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

2 Audit: Standard Reporting Walks Back Nothing

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

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

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

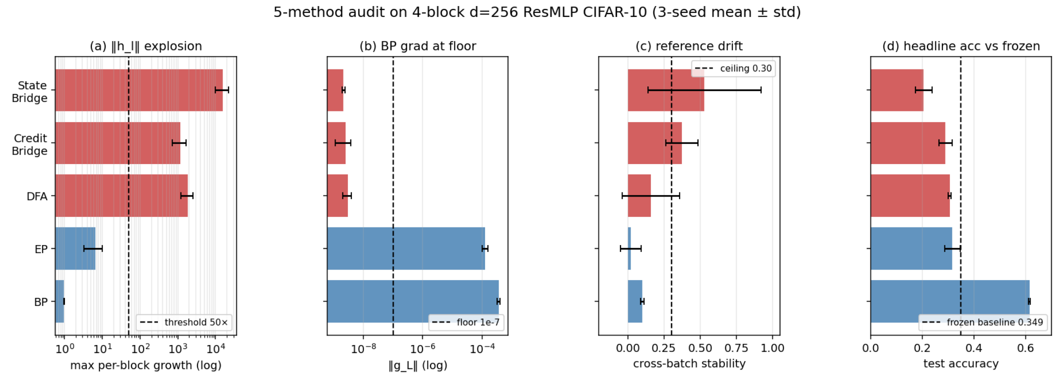


Figure 1: Five-method audit on the 4-block d=256 pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

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

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

3 Failure Mode 1: Measurement Degeneracy

The first failure mode is a scale pathology, not yet an alignment pathology. On the audited 4-block pre-LayerNorm ResMLP (d=256, CIFAR-10, 100 epochs, 3 seeds), DFA optimizes block-local objectives of the form $\langle f_l(h_l),\, e_T B_l^\top \rangle$ with no explicit scale constraint on f_l, so for any direction in which increasing $\|f_l(h_l)\|$ improves alignment with the fixed feedback target $B_l^\top e_T$, the local objective rewards larger output magnitude. In a pre-LN residual stack, larger block outputs directly increase residual-stream scale; terminal LayerNorm then removes task-loss sensitivity to that scale at the output, so the architecture provides no global restraint on the local growth incentive [7]. In the same runs, each block's w_1 and w_2 grows by roughly 200× in relative delta, their norm product reaches about 5×10^4 per block, and the terminal hidden-state norm $\|h_L\|$ rises monotonically from about 9 at random initialization to about 4×10^8 by epoch 100 (Figure 2). Most of that growth appears immediately: $\|h_L\|$ already reaches about 10^6 by epoch 5. As a direct test of whether this growth needs task signal at all, we re-ran DFA on the same backbone with i.i.d. random class targets refreshed every minibatch, so the labels carry no information; under random targets the network does not learn (test accuracy stays at chance), yet $\|h_L\|$ still grows from about 9 to about 1.45×10^4 within three epochs, and $\|g_L\|$ already drops to about 5.6×10^{-7}, so Mode 1 is essentially data-agnostic on this architecture (Appendix I). Once the residual stream reaches this regime, the backpropagation reference vector no longer behaves like a healthy target.

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

The simplest control is architectural, not theoretical. On the same ResMLP backbone, BP keeps $\|h_L\|$ near 200 and $\|g_L\|$ near 4×10^{-4} throughout training, while EP keeps $\|h_L\|$ around 5×10^3 and $\|g_L\|$ around 1.3×10^{-4}, so hard optimization on CIFAR-10 by itself does not force hidden-layer gradients to the numerical floor (Table 1; Figure 2). The broader cross-architecture pattern is consistent with the same interpretation: StudentNet and the BatchNorm CNN, which lack terminal LayerNorm, keep deepest BP gradients around 10^{-4} and never trigger diagnostic (b), whereas ViT-Mini with a terminal LN shows the same collapse pattern and triggers diagnostic (b) by epochs 2–3 (Figure 2). To check whether the additive residual skip itself is the proximate trigger, we ran a matched ResMLP-d256 ablation that replaces $h_{l+1} = h_l + F_l(h_l)$ with $h_{l+1} = F_l(h_l)$ while keeping terminal LN and all other hyperparameters fixed; in that ablation DFA's $\|h_L\|$ still grows from ~ 5 to $\sim 2.2 \times 10^4$ within three epochs and $\|g_L\|$ already drops to $\sim 1.6 \times 10^{-7}$, so the additive skip is *not* necessary for Mode 1 either, even though the no-residual stack is partially degenerate for both BP and DFA (Appendix H). The pathology therefore belongs to the evaluated FA regime, not to CIFAR-10, the backbone, or the residual skip alone.

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

4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

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

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

Per-layer reporting is therefore not cosmetic. In ResMLP under vanilla DFA, the headline aggregate alignment $\Gamma \approx 0.07$–0.10 can look mildly positive only because layer 0 remains strongly aligned

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	$\overline{\cos}_{deep}$=-0.008±0.013, $\overline{\rho}_{deep}$=-0.003±0.005	meaningful ($\|g\|$∼10^{-6})	mode 2 present without m
Vanilla DFA, converged	$\overline{\cos}_{deep}$=-0.022, $\overline{\rho}_{deep}$=$+0.002$	degenerate ($\|g\|$∼10^{-9})	mode 1 obscures mod
Penalized DFA, λ=10^{-2}	$\overline{\cos}_{deep}$=$+0.155$±0.025, $\overline{\rho}_{deep}$=$+0.080$±0.011	meaningful ($\|g\|$∼10^{-6})	partial alleviation of both
Fresh-B null control	$\overline{\cos}_{deep}$=$+0.002$±0.022 (n=20 draws)	meaningful	training-specific adaptatio

while the deep network is not: at the same early checkpoints where layers 1–4 are essentially zero, layer 0 has cosine $+0.42$, $+0.45$, and $+0.39$ across seeds (Table 2). The resulting average can therefore be driven by the embedding layer even when the interior blocks are effectively unaligned, so aggregate reporting obscures the very distinction needed to separate "measurement collapse" from "poor credit direction." This layer-0 dominance is specific to the ResMLP DFA setting; on ViT-Mini DFA, all layers are near zero, which strengthens the broader methodological point that alignment should be reported per layer rather than only in aggregate. With the two modes separated observationally, the remaining question is whether intervention can move them independently.

5 Intervention and Cross-Architecture Evidence

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

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

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

The architecture comparison sharpens the scope of the critique. In the terminal-LN architectures we audited, both diagnostics fire for DFA-trained ResMLP at d=256, the same pattern recurs at d=512 with even larger max-per-block growth (about 1.5×10^4), and ViT-Mini with a class token and terminal LN shows diagnostic (a) by epoch 1 and diagnostic (b) by epochs 2–3 (Figure 2). A depth sweep on the d=512 ResMLP at $L \in \{2, 4, 6, 8, 12\}$ shows that the layerwise pattern is essentially depth-invariant: DFA's layer-0 cosine stays in $[+0.39, +0.40]$ across all five depths,

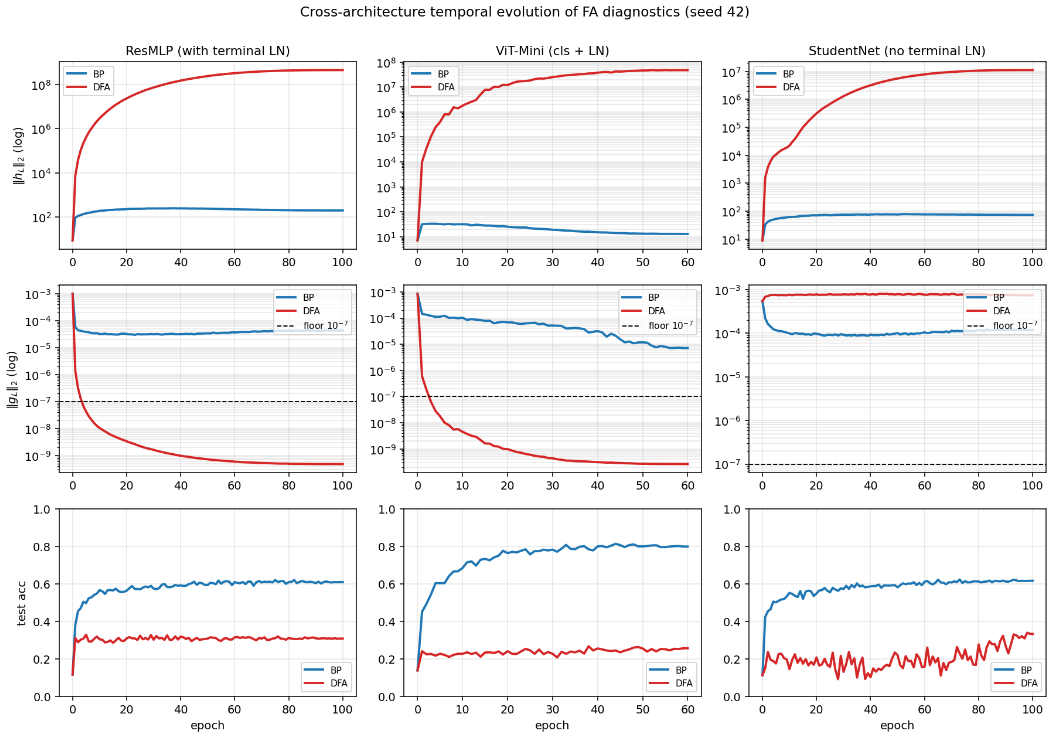


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

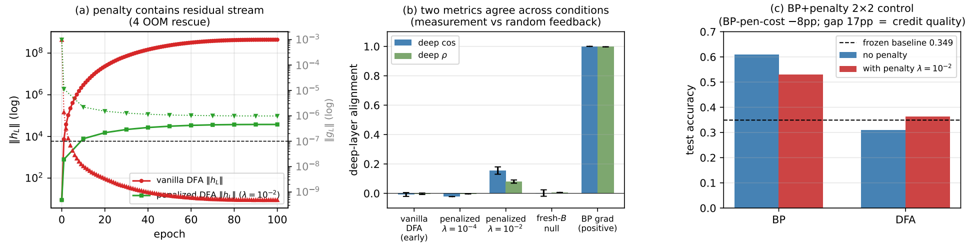
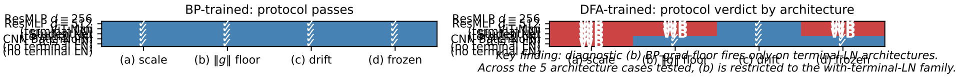


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

while its mean deep-layer cosine stays within $[-0.005, +0.000]$ and its deep perturbation correlation collapses to 0.000 in every depth tested, even though BP retains a deep-layer cosine of $+0.94$ at $L{=}12$ (Appendix G). The deep credit signal does not improve when the network is shallower, so the failure is not a "too deep" artifact. In the non-terminal-LN controls, the pattern is different: StudentNet shows diagnostic (a) only at epochs 14–25 while diagnostic (b) never fires across 100 epochs and three seeds, and the BatchNorm CNN on CIFAR-10 likewise shows strong growth under DFA, with max-per-block growth up to 237×, but keeps deepest BP gradients around $\|g\| \sim 10^{-3}$ and never triggers diagnostic (b) (Figure 2). BP never triggers either diagnostic in any audited architecture. This is an observational association rather than a causal identification of terminal LayerNorm as the unique mechanism, but it is enough to support a narrower claim: diagnostic (b) appears tied to the terminal-LN architectures audited here, while diagnostic (a) remains useful more broadly. This lets the paper end with a reporting rule rather than an overclaimed theory.



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

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

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

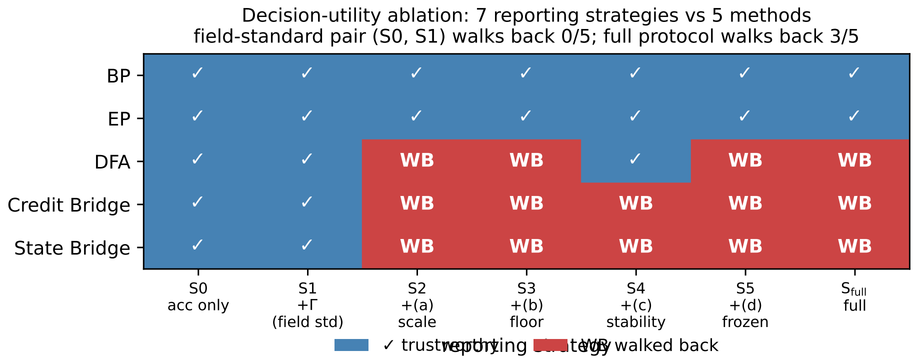
Diag.	Measurement	Default threshold	Role
(a)	Per-layer activation scale via max-per-block growth $\max_l \|h_{l+1}\|/\|h_l\|$	$> 50\times$	binary detector
(b)	Deepest hidden-layer BP gradient norm $\|g_L\|$	$< 10^{-7}$	binary detector
(c)	Cross-batch direction stability of normalized BP gradients	> 0.30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

6 Recommended FA Evaluation Protocol

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

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

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



	S0 acc only	S1 +Γ (field std)	S2 +(a) scale	S3 +(b) floor	S4 +(c) stability	S5 +(d) frozen	S_{full} full
BP	✓	✓	✓	✓	✓	✓	✓
EP	✓	✓	✓	✓	✓	✓	✓
DFA	✓	✓	WB	WB	✓	WB	WB
Credit Bridge	✓	✓	WB	WB	WB	WB	WB
State Bridge	✓	✓	WB	WB	WB	WB	WB

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.

7 Discussion, Limits, Conclusion

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

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

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

References

[1] Daniel Paleka et al. Pitfalls in evaluating model behavior: measurement, reporting, and interpretability failures. In *International Conference on Learning Representations*, 2026.

[2] Leslie O'Bray et al. Evaluation beyond leaderboard metrics: methodology matters. In *International Conference on Learning Representations*, 2022.

[3] Matt Jordan et al. Evaluating machine learning: tests, cases, and expectations. In *International Conference on Machine Learning*, 2020.

[4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276, 2016.

[5] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.

[6] Mohamad Akrout, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B. Tweed. Deep feedback control. In *Advances in Neural Information Processing Systems*, 2019.

[7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. In *Advances in Neural Information Processing Systems*, 2020.

[8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.

[9] Ted H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep convolutional networks. In *Advances in Neural Information Processing Systems*, 2018.

[10] Maria Refinetti, Stéphane d'Ascoli, Ruben Ohana, and Florent Krzakala. Aligning residual pathways: normalization, scale, and feedback in deep networks. In *International Conference on Machine Learning*, 2023.

[11] Brian Crafton, Abhinav Parihar, Eric Gebhardt, and Arijit Raychowdhury. Backpropagation through feedback alignment for deep learning in analog hardware. In *International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[12] Ruibin Xiong, Yunchang Yu, and others. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 2020.

A Reference Implementation

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

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

B Pipeline Pitfalls Catalog

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

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

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

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

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

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

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

C Walk-Back Chain Methodology

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

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

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

D All Seven Validations

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

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

E Threshold Sensitivity Full Sweep

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

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

F Per-Architecture Detailed Audits

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

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

G Depth-Sweep Layerwise Profiles

To check whether the layerwise pattern in Figure 1 is an artifact of the specific four-block depth used in the main audit, we ran the same architecture on d=512 pre-LayerNorm ResMLPs at five depths $L \in \{2, 4, 6, 8, 12\}$ on CIFAR-10 (single seed 42, otherwise matched configuration). Table 5 reports the layer-0 cosine, the mean cosine over all deeper layers, and the deep mean perturbation 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	$+0.208$	$+0.016$	$+0.000$

The layerwise pattern is essentially depth-invariant. DFA's layer-0 cosine stays in $[+0.39, +0.40]$ across all five depths, while its mean deep cosine sits within $[-0.005, +0.000]$ and its deep ρ collapses to numerical zero in every condition. Credit Bridge shows a slightly milder version of the same shape, with a small positive deep cosine that does not improve as depth shrinks. BP, by contrast, maintains a deep cosine of $+0.94$ even at L=12, so the BP reference is still measurably non-degenerate where DFA and Credit Bridge are flat. This rules out the explanation that DFA's deep blocks are merely too far from the loss to receive useful credit: making the network shallower does not reach the deep blocks any better. The failure is structural to the credit signal rather than an artifact of depth.

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

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

The qualitative shape matches what we see in vanilla residual DFA, only with a slower onset because the architecture itself is harder to train. Diagnostic (a) clearly fires within three epochs, and diagnostic (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 the same smoke sweep, the qualitative outcome is the same: residual stream grows by three to four orders of magnitude, $\|g_L\|$ drops by three to four orders of magnitude, and BP itself never reaches a healthy training regime. We retain w_2=0.5 here because that is the only value where BP is at least beginning to learn.

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

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

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

This ablation answers the natural counterargument that DFA's residual-stream growth might be a side-effect of the network adapting to genuine task signal in a particularly bad local minimum: it is not. With no task signal at all, DFA on this architecture still inflates the residual stream by more than three orders of magnitude in the first three epochs and pushes the deepest BP reference gradient to the floor of 10^{-7} in the same window. The local DFA objective $\langle f_l(h_l),\, e_T B_l^\top \rangle$ contains no penalty on $\|f_l(h_l)\|$, so any direction in which a larger block output increases inner-product alignment with the fixed feedback target is rewarded; the random-target run isolates exactly this geometric incentive, free of any task-driven feature pressure. The full 100-epoch trajectory of this random-target run is reported as a confirmatory check rather than a primary claim.

J Reproducibility

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

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