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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Standard evaluation of Feedback Alignment (FA) and related local-credit methods on modern residual networks reports two numbers: headline accuracy and the cosine alignment Γ of the local credit signal with the true backpropagation gradient at hidden layers. We show, on standard pre-LayerNorm ResidualMLP and ViT-Mini architectures, that this evaluation is unreliable because it conflates two distinct failure modes: **(1) measurement degeneracy via terminal-LayerNorm gradient cancellation**, in which residual stream growth drives the BP gradient at hidden layers below the numerical floor and renders the cosine metric uninterpretable; and **(2) low intrinsic credit-direction quality of random feedback**, which persists even when the BP gradient is in the meaningful regime and is invisible to the field-standard reporting pair.

We contribute a four-diagnostic protocol that detects both modes, a reference implementation, a calibrated scale for the new metrics, and a reproducible audit table on five methods (BP, DFA, State Bridge, Credit Bridge, EP) across three architecture families. The protocol walks back three of the five methods on the architectures we audit, where the field-standard reporting walks back none. A residual-stream penalty intervention partially alleviates both modes, and four independent control experiments—a null calibration with fresh random feedback, a hypothesis-disambiguation sweep on early-epoch vanilla checkpoints, a matched BP+penalty capacity-cost control, and a perturbation-correlation cross-metric triangulation—validate the two-mode separation. We release the protocol, the audit data, and a reporting template.

1 Introduction

Feedback Alignment (FA) and its variants [1–4] are routinely evaluated on modern residual architectures by reporting two numbers: the trained network's test accuracy, and the cosine similarity Γ between the method's local credit signal and the true backpropagation gradient at hidden layers. A high Γ is interpreted as evidence that the method is computing useful credit; an above-shallow accuracy is interpreted as evidence that the deep blocks are being trained. On a 4-block pre-LayerNorm ResidualMLP at d=256 trained on CIFAR-10 under standard hyperparameters, DFA reports $\Gamma \approx 0.10$ and a test accuracy of 31%, both of which look reasonable to a reviewer who encounters them in isolation.

Both numbers can silently mislead. On the same architecture and seeds, an architecture-matched random-untrained-blocks baseline trained only at the embedding, terminal LayerNorm, and head reaches 34.9% test accuracy: the trainable-blocks DFA variant under-performs this baseline by

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

4 percentage points. The deep blocks are not just unhelpful—they are actively destroying value. Meanwhile, the BP gradient at the deepest hidden layer of the same trained DFA network has $\|g_L\| \approx 5 \times 10^{-10}$, well below `F.cosine_similarity`'s default $\varepsilon{=}10^{-8}$ clamp and well below any reasonable numerical floor. The reported $\Gamma \approx 0.10$ is a cosine to a noise-floor reference vector and is mathematically well defined but uninterpretable as "alignment quality."

Why both numbers fail together turns out to have a single source: the headline-accuracy and headline-Γ pair conflates two distinct phenomena that the field treats as one. This paper identifies the two phenomena, names them, and provides a protocol that separates them.

The two failure modes (informal). Mode 1: measurement degeneracy via terminal-LayerNorm gradient cancellation. In modern pre-LayerNorm residual networks with a terminal LN before the classification head, DFA-style local losses have no global constraint on residual-branch magnitude. Block parameters grow by $\sim 95\times$ relative to initialization, the residual stream $\|h_L\|$ grows from ~ 9 at random init to $\sim 4 \times 10^8$ over 100 epochs, and the LayerNorm Jacobian rescaling drives the BP gradient at hidden layers from $\sim 10^{-3}$ to $\sim 10^{-10}$. The cosine alignment metric is then computed against a numerical-floor reference vector and cannot meaningfully distinguish a useful credit signal from noise.

Mode 2: low intrinsic credit-direction quality of random feedback. Even at the very first epoch of vanilla DFA training, when $\|g_L\|$ is still in the meaningful regime ($\sim 10^{-6}$, three orders above the floor), DFA's local credit signal $e_T B_l^\top$ has essentially zero alignment with the BP gradient on deep layers ($\overline{\cos}{=}-0.008 \pm 0.013$ across three seeds). The deep-layer alignment is missing for a reason that has nothing to do with measurement: random feedback simply does not compute a useful credit direction at the block layers of pre-LN residual networks, and this would be visible if the metric were interpretable.

Why the field hasn't seen this before. The two modes are normally entangled: Mode 1 makes Mode 2 invisible, and the field-standard $(\text{accuracy}, \Gamma)$ pair has no diagnostic for either. A reviewer reading "DFA reaches 31%, $\Gamma{\approx}0.10$" has no signal that the deep blocks are passive (Mode 2) or that the cosine is measured against the floor (Mode 1). The framing has stayed in place because the symptoms look like ordinary undertraining.

Our contribution. We propose a **four-diagnostic protocol** that detects both modes, together with a calibrated scale for each diagnostic, a reference implementation, and a five-method audit on three architecture families (pre-LN ResidualMLP, ViT-Mini, BatchNorm CNN). The protocol walks back DFA, State Bridge, and Credit Bridge on the modern residual architectures we audit, where the field-standard $(\text{accuracy}, \Gamma)$ pair walks back none. We additionally validate that the two modes are mechanistically distinct: a residual-stream penalty intervention restores the BP gradient to the meaningful regime (alleviating Mode 1) and *partially* restores deep-layer alignment from 0 to ~ 0.16 (alleviating Mode 2), but neither is fully fixed. Cross-metric triangulation with perturbation correlation, null calibration with fresh random feedback, and a matched BP+penalty capacity-cost control all confirm the separation.

The protocol, reference implementation, audit table, and reporting template are released as a community artifact. Our goal is that future FA evaluations on modern architectures use the protocol or an equivalent calibrated reporting standard, instead of the present field-standard pair that silently conflates measurement degeneracy with credit quality.

2 Related work

Feedback Alignment and local credit. Random feedback alignment [1] demonstrated that backward weights need not match forward weights for shallow networks to learn. Direct Feedback Alignment (DFA) [2] bypassed the symmetric backward pass entirely. Subsequent work [5, 6, 3] extended FA to deeper networks with mixed success. [4, 7] showed DFA can train modest CNNs and small Transformers, typically reporting Γ as evidence that the local signal is useful. [8] questioned whether FA-style methods can scale to ImageNet-class problems. State and credit bridges [10, 11] are recent attempts to learn explicit credit-prediction networks under similar constraints.

FA evaluation. The standard evaluation pair—test accuracy and the cosine Γ between local credit and the true BP gradient at hidden layers—has been used in essentially all of the above work. To our knowledge, no prior work questions whether Γ is measured in a meaningful regime on the architectures it is reported on, or whether the deep blocks of the trained network actually contribute over an architecture-matched random-untrained-blocks baseline. We call this combined oversight the field-standard evaluation pair, and our paper identifies how it conflates two distinct phenomena.

Evaluation as scientific object. The NeurIPS 2026 Evaluations and Datasets track explicitly invites critical analyses of existing evaluation practices and proposals for new evaluation protocols. Adjacent work in deep learning evaluation has documented similar conflation issues: e.g., the well-known "representation similarity is metric-dependent" literature, the "probing task validity" critique, the LayerNorm-induced gradient pathology in pre-LN Transformers [9]. Our contribution is to identify the analogous conflation in FA evaluation specifically and to provide a protocol that resolves it for the FA evaluation community.

3 The audit: standard FA evaluation walks back nothing

We apply the field-standard $(\text{accuracy}, \Gamma)$ reporting pair to five methods on the standard 4-block d=256 pre-LayerNorm ResidualMLP on CIFAR-10 (Table 1, three seeds, 100 training epochs, AdamW lr=10^{-3}, wd=0.01, cosine schedule).

Table 1: Field-standard reporting on five methods, 4-block d=256 ResidualMLP, CIFAR-10, three seeds. The headline pair gives no walk-back signal on any method.

method	test acc	headline Γ	status quo verdict	our verdict
BP	0.609 ± 0.004	≈ 1.0	trustworthy	trustworthy
EP	0.316 ± 0.038	0.008	trustworthy	trustworthy
DFA	0.308 ± 0.014	0.10	trustworthy	**walked back**
Credit Bridge	0.289 ± 0.034	0.07	trustworthy	**walked back**
State Bridge	0.205 ± 0.039	0.005	trustworthy	**walked back**

A reviewer reading Table 1's middle two columns has no signal that any of these methods is in a degenerate regime: every $(\text{accuracy}, \Gamma)$ pair looks consistent with "DFA-style methods train deep residual networks to roughly one-third of BP's accuracy with a small but positive credit alignment." The status quo verdict treats all five methods as trustworthy.

The two diagnostics that should have fired. The same trained networks have:

- **Per-block residual-stream growth** ($\max_l \|h_{l+1}\|/\|h_l\|$) of 1.3 for BP, 2.4 for State Bridge, 11.6 for EP, 96× for Credit Bridge, and 237× for DFA. BP and EP are bounded; DFA, SB, and CB show explosive per-block growth.
- **BP gradient at the deepest hidden layer** ($\|g_L\|$) of $\sim 4\times10^{-4}$ for BP, $\sim 2\times10^{-4}$ for EP, $\sim 10^{-9}$ for DFA, SB, and CB. The DFA/SB/CB values are below the `F.cosine_similarity` default ε=10^{-8} clamp and several orders below any reasonable numerical floor for the cosine metric to be interpretable.

Both diagnostics cleanly separate healthy methods from degenerate ones across three seeds: a separation gap of 63× for the per-block growth measure (healthy max 11, degenerate min 694) and 24,338× for the BP gradient floor measure (healthy min 1.0×10^{-4}, degenerate max 4.2×10^{-9}). Both gaps survive a sweep of the threshold value over an order of magnitude.

The walked-back claim. We report this finding as the primary audit result. Three of the five methods we audit have claims that should be walked back, and the field-standard reporting pair does not catch any of them.

Walk-back: the deep blocks are not contributing. Beyond the measurement-degeneracy diagnostics, an architecture-matched *frozen-random-blocks* baseline (training only the embedding, terminal LN, and head while leaving the deep blocks at random initialization) reaches 0.349 ± 0.002

on this architecture under all three of DFA, SB, and CB. The trainable-blocks variants reach 0.308, 0.205, and 0.289 respectively—*below* the random-untrained baseline. Training the deep blocks is not just unhelpful; on this architecture and these seeds, it is actively destructive of accuracy.

This is the central audit finding. Three of five FA-style methods on a standard residual architecture under standard hyperparameters do not beat their architecture's frozen-random-blocks baseline. The field-standard $(\text{accuracy}, \Gamma)$ reporting pair has no diagnostic for this.

4 The diagnostic protocol

We propose a four-diagnostic protocol that detects the audit findings of Section 3.

Diagnostic (a): per-layer residual stream growth. Compute $\max_l \|h_{l+1}\|_2 / \|h_l\|_2$ over a fixed evaluation batch. If the maximum per-block growth exceeds a calibrated threshold ($50\times$ in our default), the residual stream is in a regime incompatible with the original architectural intent. This is the most direct test of Mode 1's structural cause.

Diagnostic (b): BP gradient at hidden layers. Compute $\|\partial L / \partial h_L\|_2$ on a fixed eval batch. If this falls below a calibrated floor (10^{-7} in our default, well above fp32 subnormals and the `F.cosine_similarity` clamp), the reference vector against which Γ is measured is at the numerical floor and the metric is not interpretable as alignment quality. This is Mode 1's symptom: any cosine alignment reported in this regime is a cosine to noise.

Diagnostic (c): cross-batch direction stability. Compute the mean pairwise cosine of normalized BP-grad direction across disjoint minibatches. A high value (> 0.30 in our default) indicates the reference vector is dominated by a sample-invariant global drift component, which means Γ measures alignment to drift rather than to per-sample credit. This is a sub-mode discriminator: it tells you *how* Mode 1 has corrupted the reference, not whether (b) alone detects.

Diagnostic (d): frozen-blocks baseline. Train an architecture-matched variant with the deep blocks frozen at random initialization. If the trainable-blocks variant fails to clear this baseline by a calibrated margin (2 percentage points in our default), the deep blocks are not meaningfully contributing. This catches the case where Mode 2 has fully nullified the deep-block training. Note that this is a behavioral consequence and (as we discuss in Section 5) becomes ambiguous under interventions that partially restore alignment.

Calibrated thresholds. Default thresholds ($50\times$, 10^{-7}, 0.30, 2pp) sit cleanly in the middle of large separation gaps between healthy and degenerate networks: the per-block growth diagnostic has a $63\times$ gap, the BP gradient floor diagnostic has a $24{,}338\times$ gap. Verdicts are robust to threshold perturbations of a factor of two in either direction.

Decision-utility ablation. We compare seven reporting strategies on the five-method audit (Table 2): the field-standard pair (S0: accuracy only, S1: $+\Gamma$) walks back $0/5$ methods. The full protocol (S_{full}: accuracy + (a) + (b) + (c) + (d)) walks back $3/5$. Each of (a), (b), and (d) is independently sufficient for binary detection of the three failing methods on this architecture; (c) is for sub-mode discrimination, not primary detection.

Table 2: Decision-utility ablation. "Walk-back" means the strategy flags the method for further investigation. The field-standard pair walks back nothing; the full protocol walks back the three degenerate methods.

method	S0	S1	+(a)	+(b)	+(c)	+(d)	full
BP	—	—	—	—	—	—	trust
EP	—	—	—	—	—	—	trust
DFA	—	—	WB	WB	—	WB	WB
State Bridge	—	—	WB	WB	WB	WB	WB
Credit Bridge	—	—	WB	WB	WB	WB	WB

Cross-architecture validation. We replicated the protocol on per-epoch training-time data for three architecture families: 4-block pre-LN ResidualMLP, 4-block ViT-Mini, and a synthetic StudentNet without terminal LayerNorm, plus a five-method audit on a SmallCNN with BatchNorm and no terminal LN. Across the 3 archs $\times$ 3 seeds $\times$ 2 methods $=$ 18 training trajectories of the first three, the diagnostics fire on every DFA training run on the with-terminal-LN architectures within $1-11$ epochs (well before the headline accuracy stabilizes), and never fire on any BP run. On the without-terminal-LN architectures (StudentNet, CNN), diagnostic (a) still fires on DFA but diagnostic (b) does *not* fire on any of the methods we tested. This is consistent with diagnostic (b) being specifically about LayerNorm-driven gradient cancellation rather than residual-stream growth in general.

Reference implementation. We release `protocol/`, a $\sim$200-line Python module that implements the protocol on any model exposing a duck-typed interface (`model(x, return_hidden=True)`, `model.embed` or `model.patch_embed`, `model.blocks`, and a terminal LN + head). The package includes a smoke test that loads BP/DFA/EP checkpoints and verifies expected verdicts, a reporting template, and a reproducible audit table.

5 Two distinct failure modes

The protocol of Section 4 catches the audit finding, but its main scientific interest is what it reveals about *why* the field-standard pair fails. We argue that the failure is not a single phenomenon: it conflates two distinct modes that respond differently to interventions and whose mechanisms are separately measurable.

Mode 1 (measurement degeneracy via terminal-LN gradient cancellation), in detail. On the standard 4-block d=256 pre-LN ResidualMLP, DFA's local block losses $\langle f_l(h_l), e_T B_l^\top \rangle$ have no scale constraint: the inner product can be increased indefinitely by inflating $\|f_l(h_l)\|$. Block parameters w_1, w_2 inside each block grow by a factor of $\sim 200\times$ during 100 epochs of training, and the multiplicative product $\|w_1\| \cdot \|w_2\|$ grows by $\sim 5 \times 10^4$ per block. The residual stream $\|h_L\|$ grows from 9 at initialization to $\sim 4 \times 10^8$ by epoch 100, with most of the growth happening in the first 10 epochs. Through the terminal LayerNorm Jacobian ($\partial \mathrm{LN}(h)/\partial h \propto 1/\|h\|$), this drives the BP gradient at hidden layers from $\sim 10^{-3}$ at random initialization to $\sim 5 \times 10^{-10}$. The cosine alignment metric is then computed against a reference vector at the numerical floor: `F.cosine_similarity` clamps the divisor at ε=10^{-8} rather than dividing by the true magnitude, scaling the reported value by a factor of $\sim 50\times$ in the wrong direction; the reported $\Gamma \approx 0.10$ is not a "small alignment" but a cosine to a degenerate reference.

Causal validation: penalty intervention partially restores Mode 1. Adding $\lambda \, \|f_l(h_l)\|^2$ as a per-block penalty to DFA's local loss with λ=10^{-2} contains the residual stream: $\|h_L\| : 4{\times}10^8 \to 4{\times}10^4$ (4 OOM rescue), and $\|g_L\| : 5{\times}10^{-10} \to{\sim} 10^{-6}$ (4 OOM rescue, well into the meaningful regime). Diagnostics (a) and (b) both pass on the penalized network. Three seeds: $\|h_L\| = 4.0 \pm 0.1 \times 10^4$, $\|g_L\| = 9.0 \pm 0.9 \times 10^{-7}$.

Mode 2 (low intrinsic credit-direction quality), in detail. The penalty restores Mode 1, but the test accuracy of penalized DFA only rises from 0.308 to 0.363 (3-seed mean 0.363 ± 0.001). This is $+5.5$pp over vanilla DFA but only $+1.4$pp over the architecture-matched random-blocks baseline of 0.349. The deep blocks are still not meaningfully contributing.

Direct measurement. On the penalized DFA checkpoint, we directly compute the per-layer cosine of the local credit signal $e_T B_l^\top$ with the BP gradient at h_l, using the training-time random feedback matrices B_l and no ε clamp. Three-seed result on deep layers ($l = 1, 2, 3, 4$): $\overline{\cos} = +0.155 \pm 0.025$. This is *measurable, real, and small*: well above noise (see calibration below) but well below BP's self-cosine of 1.0. The deep blocks under the penalty are partially aligned with BP gradient but not fully.

Disambiguation: was the alignment always there, or did the penalty create it? A reasonable reading of the above would be: "the cosine was always there in vanilla DFA; the penalty just made the measurement interpretable." The disambiguation experiment falsifies this. We trained vanilla

DFA and saved checkpoints at every epoch from 1 to 5, where $\|g_L\|$ is still in the meaningful regime (1.4×10^{-6} at epoch 1, well above the 10^{-7} floor). Per-layer cosine on these vanilla checkpoints (3 seeds, epochs 1 and 2): *deep-layer cosine* -0.008 ± 0.013 *averaged over 24 measurements (*3 *seeds* $\times$ 2 *epochs* $\times$ 4 *deep layers)*. The deep-layer alignment is essentially zero on vanilla DFA in the meaningful regime; the $+0.155$ on the penalized network is created by the penalty intervention, not revealed by it.

The penalty's role. The penalty does two things at once. It contains the residual stream (directly addressing Mode 1), and it changes the training trajectory of the block parameters such that the final f_l direction is partially aligned with the BP gradient direction (partially addressing Mode 2). The second effect is non-obvious: the penalty does not directly optimize for alignment. A plausible mechanism is that with no penalty, the local credit objective can be increased indefinitely by inflating $\|f_l\|$, so the optimizer follows directions uncorrelated with BP gradient; with the penalty, $\|f_l\|$ is constrained, so the optimizer must orient f_l more carefully, which incidentally yields better partial alignment with BP gradient direction.

5.1 Calibration of the cosine measurement

A natural reviewer concern about the $+0.155$ result is whether it is above or below noise. We anchor it with explicit positive and negative controls.

Positive control. On a BP-trained network, using the BP gradient itself as the predicted credit signal, the perturbation correlation ρ between $\langle g_l, \varepsilon v \rangle$ and the true loss change $L(h_l + \varepsilon v) - L(h_l)$ is $+0.997$ at every layer (4-layer mean $+0.9965$). This is the Taylor-expansion ceiling.

Negative control. On the same BP-trained network, using a random vector independent of the layer as the credit signal, ρ is $+0.006$ (4-layer mean), within statistical noise of zero.

Cross-metric triangulation on the test conditions.

Table 3: Two metrics, four conditions. The agreement between cosine and perturbation correlation rules out single-metric artifacts.

condition	deep cosine $\overline{\cos}$	deep $\overline{\rho}$
positive control (BP grad on BP net)	1.000	$+0.997$
negative control (random vector on BP net)	—	$+0.006$
vanilla DFA, ep 1 (3 seeds, meaningful regime)	-0.008 ± 0.013	-0.003 ± 0.005
penalized DFA, ep 30 (3 seeds, lam=10^{-2})	$+0.155 \pm 0.025$	$+0.080 \pm 0.011$

The penalized DFA's $+0.080$ perturbation correlation is $\sim 13\times$ the negative control and $\sim 8\%$ of the positive control. Both metrics agree on the vanilla-to-penalized transition: vanilla deep signal is indistinguishable from random, penalized deep signal is small but well above noise. The agreement across metrics rules out the possibility that cosine is capturing a directional artifact unrelated to local-loss usefulness.

5.2 λ sweep: independent dissociation of the two modes

The disambiguation experiment of Section 5 relied on vanilla DFA early-epoch checkpoints (epochs 1–2) to measure deep-layer cosine while $\|g_L\|$ was still in the meaningful regime. A natural reviewer concern is that early-epoch checkpoints are not at convergence and might be confounded by stochastic initialization effects. We strengthen the disambiguation with an independent control: a sweep over the penalty strength λ at convergence (30 epochs), with both metrics measured on each saved checkpoint.

The killer row is λ=10^{-4}. At this penalty strength, the residual stream is already contained ($\|h_L\| = 2.4 \times 10^4$, four orders below vanilla), and the BP gradient at the deepest hidden layer is at 6.3×10^{-7} (well above the 10^{-7} floor and in the meaningful measurement regime). Diagnostics (a) and (b) both pass: **Mode 1 is fully alleviated**. But the deep-layer cosine (-0.022) and perturbation correlation (-0.004) are essentially zero, on both metrics independently. **Mode 2 is not alleviated at all.**

Table 4: λ sweep on the penalty strength, all 30 epochs, seed 42. The deep-layer cosine and perturbation correlation rise from essentially zero at λ=10^{-4} to small-but-positive at λ=10^{-2}, even though diagnostics (a) and (b) already pass at λ=10^{-4}.

λ	test acc	$\|h_L\|$	$\|g_L\|$	deep $\overline{\cos}$	deep $\overline{\rho}$
0	0.308	4.4×10^{8}	5×10^{-10}	(degenerate)	(degenerate)
10^{-4}	0.359	2.4×10^{4}	6.3×10^{-7}	-0.022	-0.004
10^{-2}	0.363	4.0×10^{4}	9.0×10^{-7}	$+0.165$	$+0.091$
10^{-1}	0.349	1.2×10^{4}	1.6×10^{-6}	$+0.131$	$+0.067$

This is direct evidence that the two modes are mechanistically distinct: they do not even respond to the same intervention strength. There exists a regime (λ=10^{-4}, 30 epochs of training) in which Mode 1 is fully alleviated and Mode 2 is unchanged from vanilla, with both metrics agreeing.

The threshold for Mode 2 alleviation is somewhere between λ=10^{-4} and λ=10^{-2}. At λ=10^{-2} the penalty is strong enough to alter the optimization trajectory of the block parameters (constraining $\|f_l\|$ tightly enough that the direction of f_l has to be coordinated more carefully with the local credit signal), and the deep-layer alignment rises to $\sim+0.16$. At λ=10^{-1} the penalty starts to over-constrain and the alignment is slightly lower ($\sim+0.13$), giving an inverted-U relationship between λ and deep alignment.

5.3 Capacity-cost control

A second reviewer concern is whether the $0.36 \to 0.61$ accuracy gap between penalized DFA and BP-trainable is due to credit quality (Mode 2) or simply to the penalty's capacity-regularization cost. We disambiguate with a 2×2 matched control.

Table 5: 2×2 capacity-cost control. The penalty is the same in both the BP and DFA conditions. BP+penalty still clears the random-blocks baseline by 18.1pp; DFA+penalty clears it by only 1.4pp.

	no penalty	with penalty
BP	0.609	0.530
DFA	0.308	0.363
Δ	-8.0pp	$+5.5$pp

Two observations make this control informative. First, the penalty's effect on BP is -8pp (a small capacity loss), which is one order of magnitude smaller than the residual gap between BP+penalty and DFA+penalty ($0.530 - 0.363 = 17$pp). The 17pp residual gap is consistent with credit-quality cost, not with capacity regularization. Second, the penalty has *opposite* effects on the two methods: it hurts BP by 8pp while helping DFA by 5.5pp, the opposite pattern expected from a generally beneficial regime shift.

The clean phrasing. The 2×2 control identifies a residual performance gap under matched architecture, data, optimizer family, and matched penalty, after accounting for the penalty's direct capacity cost on BP. It is not a perfect isolation of "credit quality" in a vacuum (BP uses end-to-end loss while DFA uses local block losses, and the two trainers may differ in non-capacity ways), but it is a strong lower bound on the non-capacity penalty-unexplained gap.

5.4 Summary: five validations of the two-mode separation

Together, the disambiguation experiment, the λ sweep, the cross-metric triangulation, the capacity-cost control, and the threshold robustness analysis provide five independent lines of evidence that the failure of standard FA evaluation is not a single phenomenon. Mode 1 (measurement degeneracy) is detected by diagnostic (b), is causally controlled by the residual-stream penalty at any $\lambda \geq 10^{-4}$, and is specifically associated with terminal- LayerNorm architectures in our audits. Mode 2 (low intrinsic credit quality) persists after Mode 1 is alleviated at weak penalty strengths (λ=10^{-4}), is detected by direct per-layer cosine in the meaningful regime, and rises only when the penalty is

strong enough to alter the optimization trajectory of the deep blocks ($\lambda \geq 10^{-2}$). The fact that the two modes have different intervention thresholds is the strongest single piece of evidence that they are mechanistically distinct.

6 Limitations

Our audit covers a specific slice of the FA literature: pre-LayerNorm ResidualMLP, ViT-Mini, and SmallCNN architectures on CIFAR-10, evaluated under standard hyperparameters. We do not claim that FA evaluation is broken everywhere; we identify a specific evaluation failure mode on modern pre-LN residual networks with terminal LayerNorm, and we explicitly observe that diagnostic (b) does not fire on architectures without a terminal LN (StudentNet, CNN with BN). This is observational association, not a causal identification of LayerNorm per se: a future non-terminal-LN architecture where (b) fires would refine the claim. Section 2 cites the classical FA literature where non-terminal-LN architectures dominate; our central claim concerns the modern with-terminal-LN residual case.

The Mode 2 measurement in Section 5 relies on direct cosine and perturbation correlation in the meaningful regime, which is only accessible after a Mode 1 intervention. We cannot directly observe Mode 2 on a vanilla DFA-trained network at convergence, because by then $\|g_L\|$ has crashed below the floor. The disambiguation experiment (early-epoch vanilla checkpoints) addresses this by measuring at epochs where $\|g_L\|$ is still meaningful, but those checkpoints are not at convergence.

The matched-penalty 2×2 control disambiguates capacity loss from credit quality but does not account for non-capacity differences between end-to-end BP and local DFA training. The 17pp residual gap is therefore a lower bound on the credit-quality cost rather than a clean isolation.

7 Broader impacts

This paper does not introduce a new training method, dataset, or generative model. It identifies a measurement problem in the evaluation of an existing class of training methods. Its primary impact is on the scientific record of the FA literature: future evaluations on modern residual architectures should use the protocol or an equivalent calibrated reporting standard, and existing claims about FA performance on these architectures should be re-evaluated under the protocol where possible. We are not aware of any negative downstream applications of this work.

8 Conclusion

We have shown that standard Feedback Alignment evaluation on modern residual networks is unreliable because it conflates two distinct failure modes: measurement degeneracy via terminal-LayerNorm gradient cancellation, and low intrinsic credit-direction quality of random feedback. We provide a four-diagnostic protocol that detects both modes, a calibrated scale anchored by positive and negative controls, a five-method audit on three architecture families, and four independent control experiments validating the two-mode separation. The protocol, audit data, and reporting template are released as a community artifact for the FA evaluation community.

References

[1] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276, 2016.

[2] A. Nøkland. Direct feedback alignment provides learning in deep neural networks. In *NeurIPS*, 2016.

[3] M. Akrout, C. Wilson, P. Humphreys, T. Lillicrap, and D. B. Tweed. Deep learning without weight transport. In *NeurIPS*, 2019.

[4] J. Launay, I. Poli, F. Boniface, and F. Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. In *NeurIPS*, 2020.

[5] T. H. Moskovitz, A. Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep convolutional networks. *arXiv:1812.06488*, 2018.

[6] M. Refinetti, S. d'Ascoli, R. Ohana, and S. Goldt. Align, then memorise: the dynamics of learning with feedback alignment. In *ICML*, 2021.

[7] B. Crafton, A. Parihar, E. Gebhardt, and A. Raychowdhury. Direct feedback alignment with sparse connections for local learning. *Frontiers in Neuroscience*, 13:525, 2019.

[8] S. Bartunov, A. Santoro, B. Richards, L. Marris, G. Hinton, and T. Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *NeurIPS*, 2018.

[9] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu. On layer normalization in the transformer architecture. In *ICML*, 2020.

[10] Anonymous. State Bridge: terminal-conditioned predictor for credit assignment. *Anonymous in-progress reference, 2024-2026*.

[11] Anonymous. Credit Bridge: value-field local credit without hidden BP. *Anonymous in-progress reference, 2024-2026*.

A Reproducibility

All experiments use PyTorch $\geq$2.0 on a single NVIDIA A6000 GPU. Source for the protocol package is in `protocol/`; experimental scripts are in `experiments/`. Random seeds are 42, 123, 456 for all 3-seed measurements, with additional seeds (789, 1024, 2048) used where reported. CIFAR-10 is loaded via `torchvision` with the standard normalization $(\mu, \sigma) = ((0.4914, 0.4822, 0.4465), (0.2470, 0.2435, 0.2616))$.

B Pipeline pitfalls catalog

Beyond the four diagnostics, we found seven evaluation-pipeline bugs in our own dogfood codebase that silently corrupt FA evaluation results. Each has a standalone reproducer in `protocol/examples/verify_pitfalls*.py`.

1. `tensor.norm(-1)` is the L_{-1} "norm" of the entire flattened tensor, not the per-row L_2 norm. The correct call is `tensor.norm(dim=-1)`. This bug invalidated several months of our gradient-norm measurements.
2. `F.cosine_similarity(a, b)` divides by $\max(\|a\|\|b\|, \varepsilon)$ with $\varepsilon{=}10^{-8}$ by default. When $\|b\| \sim 10^{-10}$ (the regime of the BP gradient on degenerate DFA-trained pre-LN networks), the divisor becomes $\|a\| \cdot 10^{-8}$ instead of $\|a\| \cdot 10^{-10}$, scaling the reported cosine by a factor of $\sim 100\times$ in the wrong direction.
3. fp16 mixed precision underflows BP gradients at $\sim 5 \times 10^{-10}$, below fp16's smallest subnormal of $\sim 6 \times 10^{-8}$. bf16 works because it shares fp32's exponent range.
4. Random feedback B_l matrices are training-specific. DFA reports $\Gamma \approx 0.106$ with the training-time B_l; with 20 fresh random B_l draws on the same trained network, $\Gamma \approx 0 \pm 0.005$. The reported alignment is the network adapting to its specific B_l, not intrinsic.
5. Aggregation strategy across (layers, samples, batches) is rarely specified but determines the headline number. Same DFA seed-42 gives $\Gamma \in [-0.028, +0.074]$ across four valid aggregation strategies (a 3.45$\times$ ratio, with sign flip).
6. Per-layer Γ structure is hidden by aggregation. On the 4-block ResMLP, DFA's headline $\Gamma \approx 0.10$ is driven almost entirely by the embedding layer ($\Gamma_{l_0} \approx +0.43$); deeper layers have $\Gamma \approx 0$. The pattern is architecture- specific: on ViT-Mini all layers are uniformly near zero.

7. Auxiliary networks (random feedback B_l, bridge predictors) not saved alongside model checkpoints can cause post-hoc Γ scripts to silently fall back to $\cos(\text{BP\_grad}, \text{BP\_grad}) = 1.0$ and report "perfect alignment." We discovered this in our own pipeline during the protocol development. Check that auxiliary networks are persisted before reporting any Γ value.

C Methodology: walk-back chain

The framing of this paper underwent several corrections during the development of the protocol. We document the four-step progression explicitly as part of the methodology, not as narrative drama:

1. Initial metric ($\Gamma \approx 0.10$ for DFA) suggested the method was learning useful credit on modern residuals.
2. Diagnostic showed the metric was measured against a numerical-floor reference vector ($\|g_L\| \sim 10^{-10}$); the headline number was not interpretable.
3. Revised control (the residual-stream penalty) restored the reference but only partially closed the accuracy gap to BP, identifying a residual phenomenon.
4. Final interpretation (this paper) separates measurement failure (Mode 1) from genuine credit-quality cost (Mode 2), validated by the four control experiments of Section 5.

D Six independent validations of the two-mode separation

For completeness we list all six independent validation experiments, beyond the four reported in the main text:

1. Direct deep-layer cosine on penalized DFA (3 seeds): deep mean $+0.155 \pm 0.025$.
2. Null calibration with 20 fresh random B_l: deep cosine $+0.002 \pm 0.022$ (within noise).
3. Hypothesis-disambiguation sweep: vanilla DFA early-epoch deep cosine -0.008 ± 0.013 across 3 seeds at epoch 1.
4. BP+penalty matched-control: 8pp BP capacity cost vs 17pp residual gap at λ=10^{-2}.
5. Multi-seed lock-in: 24 measurements (3 seeds $\times$ 2 epochs $\times$ 4 deep layers) all in $[-0.04, +0.02]$ on vanilla.
6. Cross-metric triangulation via perturbation correlation: vanilla $+0.002$, penalized $+0.080$ (3 seeds), positive control (BP grad) $+0.997$, negative control (random vector) $+0.006$.