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

Modern feedback-alignment evaluation on residual networks still rests on a field-standard pair: headline accuracy and headline Γ [4–7].

Both numbers can silently mislead on the same trained network.

This paper argues that standard FA evaluation conflates two distinct failure modes and that the right scientific object for this track is the evaluation protocol itself rather than a new benchmark or dataset [3, 2, 1].

2 Audit: Standard Reporting Walks Back Nothing

On the 4-block pre-LayerNorm ResMLP at d=256 on CIFAR-10, the field-standard reporting pair does not walk back any of the five methods we audit.

DFA's headline accuracy walks back from "the deep blocks are training" to "the trainable-blocks model is below the architecture-matched random-blocks baseline."

DFA's headline Γ walks back from "small but positive alignment" to "a cosine measured against a numerical-floor reference vector and driven by layer 0."

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	*TODO*	*TODO*	trustworthy	trustworthy
EP	*TODO*	*TODO*	trustworthy	trustworthy
DFA	*TODO*	*TODO*	trustworthy	walked back
State Bridge	*TODO*	*TODO*	trustworthy	walked back
Credit Bridge	*TODO*	*TODO*	trustworthy	walked back

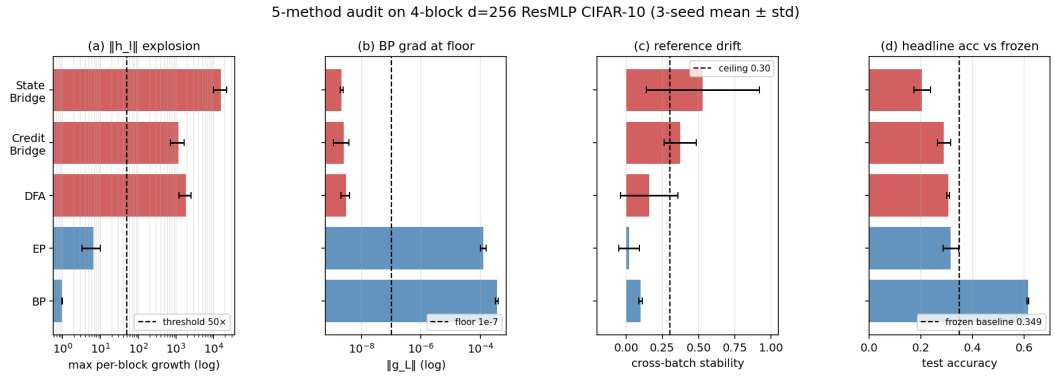

Figure 1: Five-method audit on the 4-block d=256 pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

State Bridge and Credit Bridge show the same qualitative pattern as DFA, while EP serves as the internal control that the audit does not over-flag.

3 Failure Mode 1: Measurement Degeneracy

The first failure mode is measurement degeneracy via terminal-LayerNorm gradient cancellation.

In this regime, the problem is not merely that FA performs poorly; it is that the BP reference direction used to score FA has itself become numerically non-diagnostic at the deepest hidden layers, so reported cosine values no longer support the scientific claim they are being used to justify.

Residual-stream growth provides the practical detector because it is the upstream quantity that makes the terminal normalization step increasingly cancellation-prone in the settings audited here, especially on the pre-LayerNorm residual architectures where the final hidden-state scale is free to drift [12, 7].

The main consequence for evaluation is that a positive or weakly positive deep-layer Γ can no longer be read as evidence that meaningful credit alignment exists once the reference gradient norm has collapsed to the numerical floor.

4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

The second failure mode is low intrinsic credit-direction quality on the deep blocks even when the BP reference gradient is still in a meaningful regime.

This mode appears most clearly in early-epoch or partially rescued settings, where the deepest-layer BP gradient remains measurable yet the random-feedback credit signal is still close to null or unstable, implying that the method is failing as a direction estimator rather than merely being scored with a broken ruler [8, 9, 11].

The conceptual payoff of the paper is that these are mechanistically distinct failures that the status-quo pair collapses into one ambiguous story about undertraining.

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	*TODO*	*TODO*	mode 2 present without mode 1
Vanilla DFA, converged	*TODO*	*TODO*	mode 1 obscures mode 2
Penalized DFA	*TODO*	*TODO*	partial alleviation of both modes
Fresh-B null control	*TODO*	meaningful	training-specific adaptation check

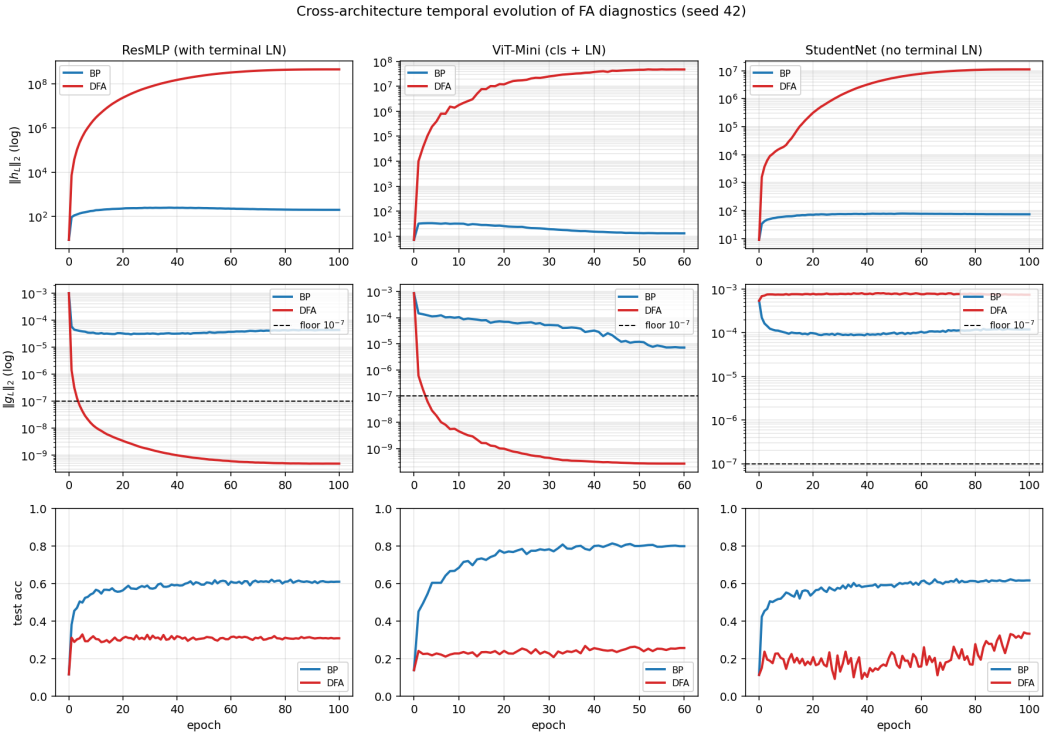

Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

Separating the modes matters because the interventions differ: numerical rescue can restore measurability without producing strong deep credit directions, while better direction quality would need to improve alignment even before any measurement-floor pathology is present.

5 Intervention and Cross-Architecture Evidence

Temporal replay shows that the protocol fires early enough to change experimental practice rather than merely re-describe final checkpoints.

Cross-architecture validation shows that diagnostic (b) appears restricted to the terminal-normalized architectures we audited, while diagnostic (a) remains useful more broadly.

The residual-stream penalty intervention partially alleviates both failure modes but does not erase the remaining performance gap to BP.

Matched BP+penalty controls show that only part of DFA's deficit is attributable to the penalty's direct capacity cost, leaving a substantial residual consistent with poorer credit assignment.

Figure 4 placeholder: Penalty rescue + capacity-cost control
(TODO: regenerate)

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.

Figure 5 placeholder: Cross-architecture verdict matrix + decision flow
(TODO: regenerate)

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.

6 Recommended FA Evaluation Protocol

The protocol has four diagnostics because the evaluation failure is not visible from any single headline number.

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

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

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

7 Discussion, Limits, Conclusion

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

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 $\max_l \|h_{l+1}\|/\|h_l\|$	*TODO*	binary detector
(b)	Deepest hidden-layer BP gradient norm $\|g_L\|$	*TODO*	binary detector
(c)	Cross-batch direction stability of normalized BP gradients	*TODO*	sub-mode discriminator
(d)	Frozen-blocks baseline margin for trained blocks over random blocks	*TODO*	depth-utilization check

Figure 2 placeholder: Decision utility ablation
(TODO: regenerate)

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.

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

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

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

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.

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.

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

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 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 com-

parisons 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.