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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

18 1 Introduction

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

21 Both numbers can silently mislead on the same trained network.

22 This paper argues that standard FA evaluation conflates two distinct failure modes and that the right
23 scientific object for this track is the evaluation protocol itself rather than a new benchmark or dataset
24 [3, 2, 1].

25 2 Audit: Standard Reporting Walks Back Nothing

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

28 DFA’s headline accuracy walks back from “the deep blocks are training” to “the trainable-blocks
29 model is below the architecture-matched random-blocks baseline.”

30 DFA’s headline Γ walks back from “small but positive alignment” to “a cosine measured against a
31 numerical-floor reference vector and driven by layer 0.”

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

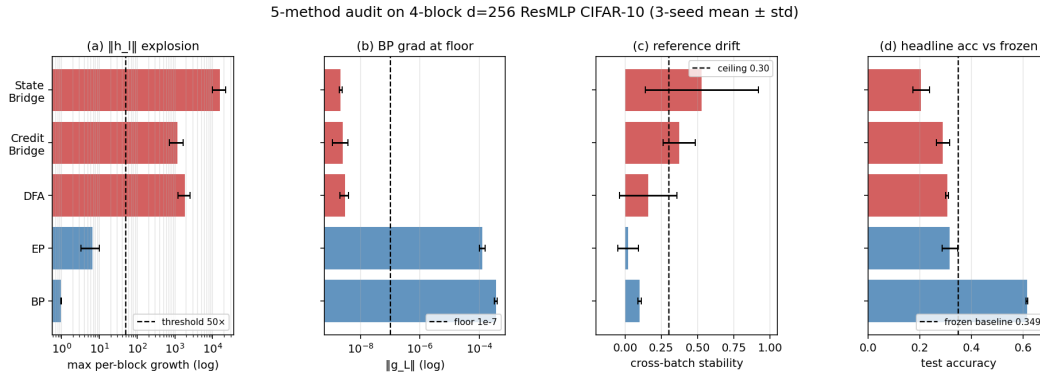


Figure 1: Five-method audit on the 4-block $d=256$ pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

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

34 **3 Failure Mode 1: Measurement Degeneracy**

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

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

39 Residual-stream growth provides the practical detector because it is the upstream quantity that makes
 40 the terminal normalization step increasingly cancellation-prone in the settings audited here, espe-
 41 cially on the pre-LayerNorm residual architectures where the final hidden-state scale is free to drift
 42 [12, 7].

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

46 **4 Failure Mode 2: Low Intrinsic Credit-Direction Quality**

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

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

53 The conceptual payoff of the paper is that these are mechanistically distinct failures that the status-
 54 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	$\overline{\text{cos}}_{deep} = -0.008 \pm 0.013, \overline{\rho}_{deep} = -0.003 \pm 0.005$	meaningful ($\ g\ \sim 10^{-6}$)	mode 2 present without m
Vanilla DFA, converged	$\overline{\text{cos}}_{deep} = -0.022, \overline{\rho}_{deep} = +0.002$	degenerate ($\ g\ \sim 10^{-9}$)	mode 1 obscures mod
Penalized DFA, $\lambda = 10^{-2}$	$\overline{\text{cos}}_{deep} = +0.155 \pm 0.025, \overline{\rho}_{deep} = +0.080 \pm 0.011$	meaningful ($\ g\ \sim 10^{-6}$)	partial alleviation of both
Fresh- B null control	$\overline{\text{cos}}_{deep} = +0.002 \pm 0.022$ ($n=20$ draws)	meaningful	training-specific adaptation

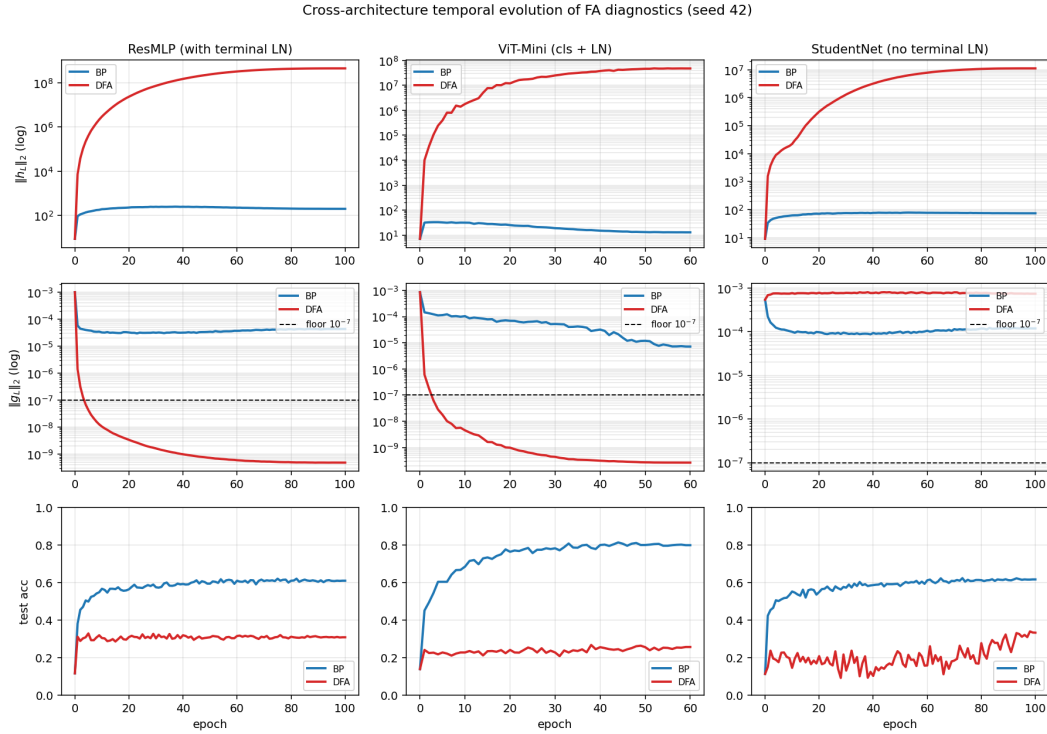


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

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

58 5 Intervention and Cross-Architecture Evidence

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

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

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

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

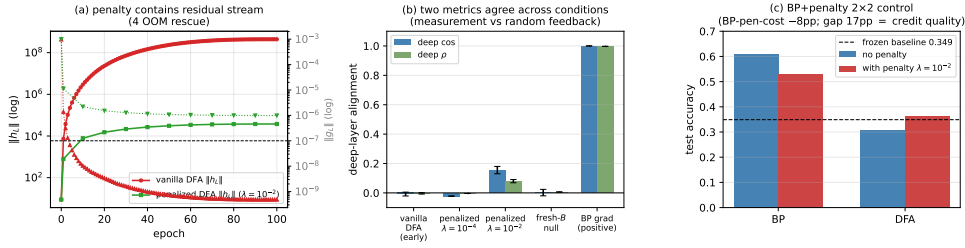


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

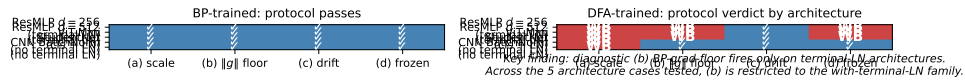


Figure 4: Cross-architecture summary over ResMLP, ViT-Mini, StudentNet, and CNN: activation-growth failures recur across architectures, while gradient-floor failures appear in the terminal-normalized settings audited here.

67 6 Recommended FA Evaluation Protocol

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

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

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

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

76 7 Discussion, Limits, Conclusion

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

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

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

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

89 References

- 90 [1] Daniel Paleka et al. Pitfalls in evaluating model behavior: measurement, reporting, and inter-
91 pretability failures. In *International Conference on Learning Representations*, 2026.
- 92 [2] Leslie O’Bray et al. Evaluation beyond leaderboard metrics: methodology matters. In *Interna-
93 tional Conference on Learning Representations*, 2022.

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

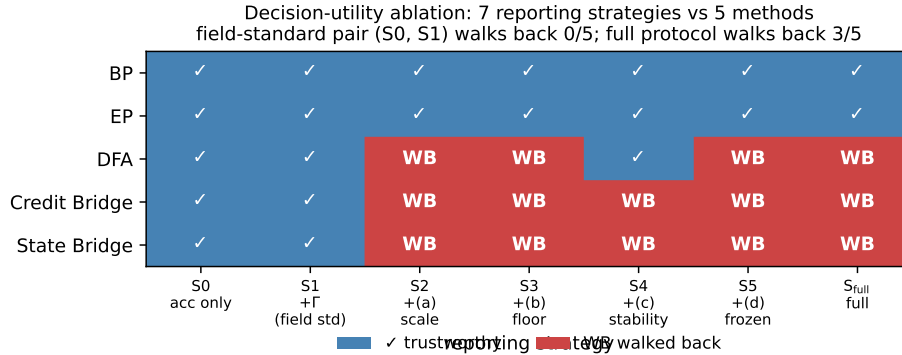


Figure 5: Decision-utility ablation comparing the field-standard reporting pair against progressively richer diagnostic strategies: accuracy only and accuracy+ Γ walk back no audited failures, while the full protocol walks back the three silent failures.

- 94 [3] Matt Jordan et al. Evaluating machine learning: tests, cases, and expectations. In *International*
95 *Conference on Machine Learning*, 2020.
- 96 [4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random
97 synaptic feedback weights support error backpropagation for deep learning. *Nature Communi-*
98 *cations*, 7:13276, 2016.
- 99 [5] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In
100 *Advances in Neural Information Processing Systems*, 2016.
- 101 [6] Mohamad Akrouf, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B.
102 Tweed. Deep feedback control. In *Advances in Neural Information Processing Systems*, 2019.
- 103 [7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback align-
104 ment scales to modern deep learning tasks and architectures. In *Advances in Neural Informa-*
105 *tion Processing Systems*, 2020.
- 106 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and
107 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-
108 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 109 [9] Ted H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep con-
110 volutional networks. In *Advances in Neural Information Processing Systems*, 2018.
- 111 [10] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Florent Krzakala. Aligning residual
112 pathways: normalization, scale, and feedback in deep networks. In *International Conference*
113 *on Machine Learning*, 2023.
- 114 [11] Brian Crafton, Abhinav Parihar, Eric Gebhardt, and Arijit Raychowdhury. Backpropagation
115 through feedback alignment for deep learning in analog hardware. In *International Conference*
116 *on Acoustics, Speech, and Signal Processing*, 2019.

117 [12] Ruibin Xiong, Yunchang Yu, and others. On layer normalization in the transformer architecture.
118 In *International Conference on Machine Learning*, 2020.

119 A Reference Implementation

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

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

134 B Pipeline Pitfalls Catalog

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

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

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

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

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

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

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

Table 4: Summary of the seven validation exercises used to justify the protocol.

Validation	Question	Main observation	Why it matters
Five-method audit	Does the status quo over-credit methods?	Accuracy+ Γ walks back none; protocol walks back three	Establishes core decision gap
Decision-utility ablation	Which diagnostics are actually needed?	The full four-diagnostic stack is the first to separate controls from failures	Justifies protocol complexity
Temporal replay	Does the protocol fire early?	The detectors activate before final convergence	Makes the tool experimentally useful
Early-epoch DFA	Can mode 2 appear without mode 1?	Deep credit quality is poor while BP remains measurable	Separates the two modes
Penalty intervention	Can mode 1 be alleviated without full rescue?	Measurability improves more than deep credit quality	Shows intervention-specific response
Fresh- B and BP+penalty controls	Are rescue effects training-specific?	Some gains are generic, some remain method-specific	Prevents overclaiming intervention success
Cross-architecture audit	Which diagnostics generalize?	Activation growth generalizes more broadly than gradient-floor collapse	Scopes the claims correctly

164 C Walk-Back Chain Methodology

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

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

178 D All Seven Validations

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

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

189 **E Threshold Sensitivity Full Sweep**

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

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

201 **F Per-Architecture Detailed Audits**

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

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

213 **G Reproducibility**

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

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