

---

# 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  $\Gamma$  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  $\Gamma$  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 Evaluations  
13 & Datasets track: we provide the protocol, the calibration logic for its thresh-  
14 olds, a reference implementation, a five-method audit, and validation through tem-  
15 poral replay, cross-architecture checks, intervention-based disambiguation, and a  
16 documented catalog of pipeline pitfalls, in the spirit of critical evaluation analyses  
17 such as Jordan et al. [3], O’Bray et al. [2], Paleka et al. [1].

## 18 1 Introduction

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

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

Table 1: Main audit table for the 4-block  $d=256$  pre-LayerNorm ResMLP on CIFAR-10. The row and column structure is fixed here; fill from the three-seed audit output.

Method	Test acc.	Headline $\Gamma$	Status-quo verdict	Protocol verdict
BP	$0.615 \pm 0.003$	$\approx 1.0$	trustworthy	trustworthy
EP	$0.316 \pm 0.030$	0.008	trustworthy	trustworthy
DFA	$0.306 \pm 0.006$	0.10	trustworthy	walked back
State Bridge	$0.205 \pm 0.032$	0.005	trustworthy	walked back
Credit Bridge	$0.289 \pm 0.026$	0.07	trustworthy	walked back

36 collapses to  $\|g_L\| \sim 5 \times 10^{-10}$ , so the deep-layer cosine is measured against a clamp-dominated  
 37 floor rather than a meaningful target direction. The same figure also shows the second failure mode,  
 38 *Mode 2: low intrinsic credit-direction quality*, because even after comparing against the stronger  
 39 frozen-blocks baseline ( $0.349 \pm 0.002$ ) and looking layer-by-layer, DFA’s deep blocks remain essen-  
 40 tially null while only layer 0 is visibly positive. To test whether this is only a measurement problem,  
 41 the intervention results show a dissociation: with a residual penalty  $\lambda \|f_l(h_l)\|^2$ , the deepest state  
 42 scale falls toward  $4 \times 10^4$ , the reference gradient rises toward  $10^{-6}$ , and deep cosine can improve  
 43 to about  $+0.16$ , yet at  $\lambda=10^{-4}$  Mode 1 is alleviated while deep cosine still stays near zero, and at  
 44 vanilla DFA epoch 1 the reference is already meaningful at about  $6 \times 10^{-7}$  but the deep cosine is still  
 45  $-0.008 \pm 0.013$  across three seeds. The failure is not unitary: one mode breaks the measurement,  
 46 and the other survives even when the measurement is still meaningful.

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

## 58 2 Audit: Standard Reporting Walks Back Nothing

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

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

63 DFA’s headline  $\Gamma$  walks back from “small but positive alignment” to “a cosine measured against a  
 64 numerical-floor reference vector and driven by layer 0.”

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

## 67 3 Failure Mode 1: Measurement Degeneracy

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

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

72 Residual-stream growth provides the practical detector because it is the upstream quantity that makes  
 73 the terminal normalization step increasingly cancellation-prone in the settings audited here, espe-  
 74 cially on the pre-LayerNorm residual architectures where the final hidden-state scale is free to drift  
 75 [12, 7].

5-method audit on 4-block  $d=256$  ResMLP CIFAR-10 (3-seed mean  $\pm$  std)

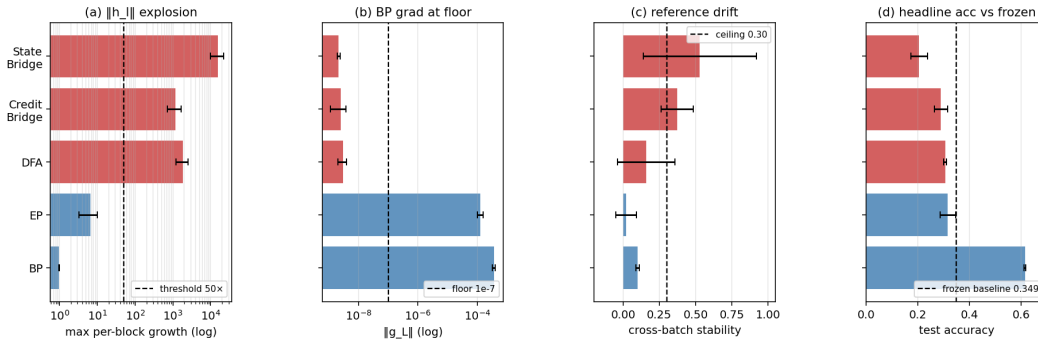


Figure 1: Five-method audit on the 4-block  $d=256$  pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

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

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

#### 79 4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

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

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

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

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

#### 91 5 Intervention and Cross-Architecture Evidence

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

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

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

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

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

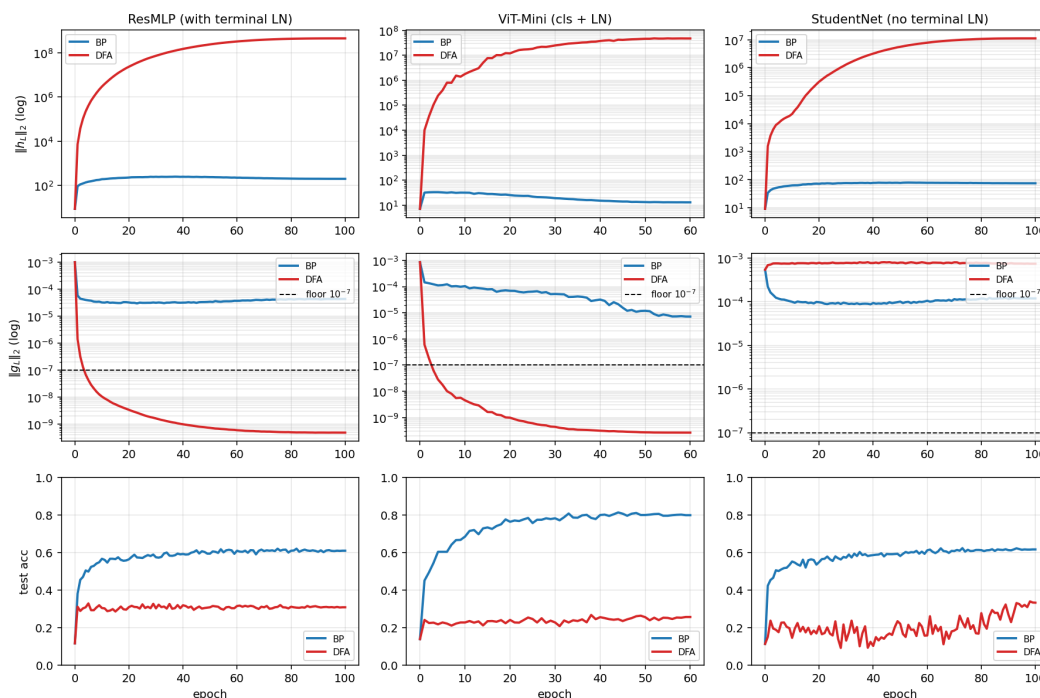


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

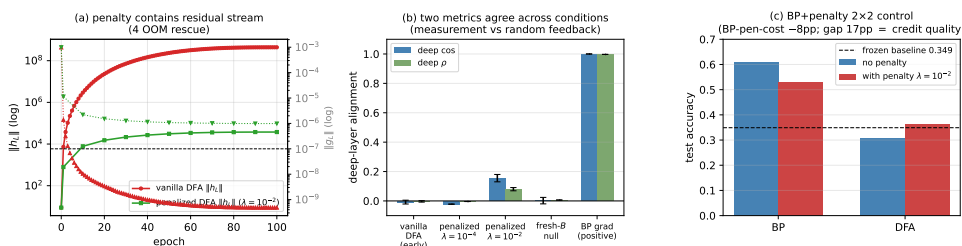


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

## 100 6 Recommended FA Evaluation Protocol

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

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

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

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



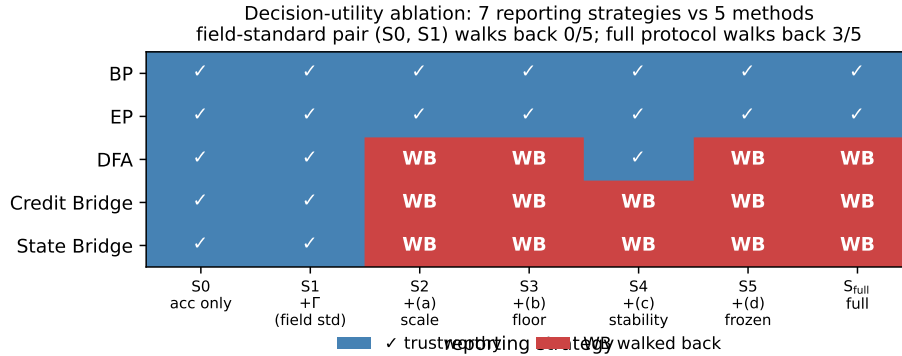


Figure 5: Decision-utility ablation comparing the field-standard reporting pair against progressively richer diagnostic strategies: accuracy only and accuracy+ $\Gamma$  walk back no audited failures, while the full protocol walks back the three silent failures.

- 139 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and  
140 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-  
141 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 142 [9] Ted H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep con-  
143 volutional networks. In *Advances in Neural Information Processing Systems*, 2018.
- 144 [10] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Florent Krzakala. Aligning residual  
145 pathways: normalization, scale, and feedback in deep networks. In *International Conference*  
146 *on Machine Learning*, 2023.
- 147 [11] Brian Crafton, Abhinav Parihar, Eric Gebhardt, and Arijit Raychowdhury. Backpropagation  
148 through feedback alignment for deep learning in analog hardware. In *International Conference*  
149 *on Acoustics, Speech, and Signal Processing*, 2019.
- 150 [12] Ruibin Xiong, Yunchang Yu, and others. On layer normalization in the transformer architecture.  
151 In *International Conference on Machine Learning*, 2020.

## 152 A Reference Implementation

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

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

## 167 B Pipeline Pitfalls Catalog

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

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

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

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

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

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

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

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

## 197 C Walk-Back Chain Methodology

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

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

## 211 D All Seven Validations

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

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

Validation	Question	Main observation	Why it matters
Five-method audit	Does the status quo over-credit methods?	Accuracy+ $\Gamma$ 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

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

## 222 E Threshold Sensitivity Full Sweep

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

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

## 234 F Per-Architecture Detailed Audits

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

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

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

## 246 **G Reproducibility**

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

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